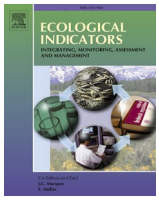
Ecological Indicators 171 (2025) 113132

Contents lists available at ScienceDirect

Ecological Indicators

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



Original Articles

Alleviating small sample problem in continuous forest monitoring with remote sensing-assisted Copulas

Xinjie Cheng [a,b], Zhengyang Hou [a,b,*], Annika Kangas [c], Jean-Pierre Renaud [d,e], Hao Tang [f], Weisheng Zeng [g], Qing Xu [h]

[a] The Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing 100083, China
[b] Ecological Observation and Research Station of Heilongjiang Sanjiang Plain Wetlands, National Forestry and Grassland Administration, Shuangyashan 518000, China
[c] Natural Resources Institute Finland (Luke), Bioeconomy and Environment Unit, Joensuu, Finland
[d] Office National des Forêts, Département Recherche Développement Innovation, 5 rue Girardet, 54052 Nancy, France
[e] Laboratoire d'inventaire forestier, ENSG, IGN, 14 rue Girardet, 54000 Nancy, France
[f] Department of Geography, National University of Singapore, Singapore
[g] Academy of Forest and Grassland Inventory and Planning, National Forest and Grassland Administration, Beijing 100714, China
[h] Key Laboratory of National Forestry and Grassland Administration/Beijing for Bamboo & Rattan Science and Technology, International Center for Bamboo and Rattan, Beijing 100102, China

ARTICLE INFO

Keywords:
Survey sampling
Machine learning
Model-assisted estimators
Copulas
Sample size optimization

ABSTRACT

With model-assisted (MA) estimation, remote sensing (RS) has provided auxiliary modeling data to enhance precision in estimators of forest parameters for continuous forest monitoring as mandated by various official reporting instruments. However, model-assisted estimation is largely reliant on a sample resulting from costly field surveys to meet the precision standard mandated by these instruments. While a large sample is more likely to represent the population in question and ensure meeting the prescribed precision, it is crucial to reduce costs by finding a balance between precision and sample size. Consequently, this study aims to (1) develop and demonstrate estimation using Copulas modeling; (2) propose a sample size optimization procedure for MA estimators in the context of continuous forest monitoring; and (3) compare survey precisions of the estimators using Copulas and Weighted Least Squares regression (WLS) as a function of sample sizes. Four main conclusions are relevant: for both Burkina Faso (BF) and Genhe (GH) study area, (1) Copulas outperforms WLS in modeling and prediction, both in terms of mean values and maximum/minimum values; (2) Copulas consistently demonstrates superior performance and precision across varying sample sizes compared to the WLS with MA estimators; (3) a straightforward sample size optimization approach reveals that variance estimates of Copulas remain lower than those of WLS as the sample size decreases in monitoring surveys; (4) Copulas requires about 20% smaller sample size than WLS does when achieving a specified precision, suggesting enhanced efficiency. Overall, Copulas appears promising to satisfy the precision, cost-efficiency, and flexibility requirements of monitoring surveys, particularly in situations involving small sample sizes.

1. Introduction

Annual estimates for forestry and ecological indicators are essential for monitoring changes in forest resources, the sustainability of forest management, and forest carbon emissions and sinks as mandated by intergovernmental organizations and processes, including the United Nations Framework Convention on Climate Change (UNFCCC).

Traditional large-scale inventory programs usually utilize field surveys employing probabilistic sampling designs to facilitate estimators with adequate precision, categorized as design-based inference (Gregoire, 1998; Tomppo et al., 2010). With the development of remote sense (RS) technology, National Forest Inventory (NFI) in many countries have combined expensive field plot data with remotely sensed information to enhance precision in estimators of forest parameters, categorized as model-based inference (Ståhl et al., 2011). The core issue in forest monitoring is thus determining and optimizing sample size to reduce costs while meeting the specified precision requirements.

Model-assisted estimators can serve as an adequate option for small

* Corresponding author.
E-mail address: houzhengyang@bjfu.edu.cn (Z. Hou).

https://doi.org/10.1016/j.ecolind.2025.113132
Received 20 June 2024; Received in revised form 4 January 2025; Accepted 19 January 2025
Available online 24 January 2025

sample size scenarios. The main difference between design-based and model-based inferences is the reliance on the probability sample and model (Ståhl et al., 2016). Design-based inference relies on large sample sizes for adequate precision (Cochran, 1977; Särndal et al., 1992). Compared to design-based inference, model-based inference can offer a higher level of precision with the same sample size. Considering the above two inferences, model-assisted approach integrates models to enhance the efficiency of estimation, also maintaining approximate unbiasedness (Breidt and Opsomer, 2017). While model-assisted estimation holds the potential to mitigate small sample problems (Faber and Fonseca, 2014; McRoberts et al., 2013), the selection of independent variables and the validity of the construction model can become obstacles. Subject to the selection of independent variables from the remote sensing auxiliary, better statistical or modeling approaches need to be attempted to improve precision on a sample basis. Data assimilation is also a viable option for improving prediction or estimation, including a broad category of mathematical procedures that refine existing parameter estimates to more accurately gauge the condition of a system by integrating data from various sources (Ehlers et al., 2013; Hou et al., 2023, 2021, 2019; Kangas et al., 2020; Mohamedou et al., 2022; Xu et al., 2023). However, meeting its high-quality requirements for numerical models and observing data may pose challenges.

Optimizing sample size is an essential procedure for both baseline and monitoring surveys. Considering model-assisted estimators (Särndal et al., 1992), it is preferable to provide fewer sample plots to establish models at a specified precision of NFI, thereby lowering cost. However, reduced sample sizes might pose a risk of failing to meet precision criteria because a large sample mitigates the adverse effects of underfitting with regression the extreme values or outliers which may have a greater ecological significance. Simple random sampling (SRS) is the simplest and most straightforward method used as a benchmark when assessing the effectiveness of estimation with auxiliary variables (Räty et al., 2020). Given that model-based and model-assisted estimators lack a standardized method for sample size calculation or utilize excessively complex methods (Charles et al., 2009; Nedyalkova and Tille, 2008), it is essential to develop an empirical formula for determining sample size based on the simple random sampling (SRS) estimation from previous surveys. This approach facilitates the optimization process (e.g., Frazer et al., 2011; McRoberts et al., 2014), which should aim to strike a balance between reducing costs and ensuring that precision requirements are met to yield reliable results.

Copula is a family of modeling and prediction procedures with promising properties, such as mitigating small sample size problems, outperforming regression in extrapolating maxima/minima, and so on. Copula captures the correlation/joint distribution variables to constitute the flexible modeling of complex dependencies even in larger dimensions (Bhatti and Do, 2019). First, copula provides a way to generate a large set of data from a small one considering the same multivariate distribution of the study variables (Houssou et al., 2022; Lin and Chaganty, 2021). Many recent studies of forest inventory have generated simulated populations through copula or Vine copula techniques (Nelsen, 2006) to conduct the applicability of certain estimators (e.g., Ene et al., 2012; Kangas et al., 2016; McRoberts et al., 2022) under design-based and model-based inference. Second, the capability of the copula method to effectively handle two or multiple random variables and its reduced susceptibility to collinearity issues (Kraus and Czado, 2017) make it a valuable choice for model prediction. For instance, Xu et al. (2019) suggested that the copula model outperformed LOESS model on prediction of ALS-based tree diameter at breast height (DBH). For modeling purposes, copula enables both linear and nonlinear dependence and accommodates extreme values (Kumar and Shoukri, 2007). Hence, utilizing copula theory to model will offer more effective and accurate predictions.

Consequently, the objectives of this study are threefold: (1) to propose and demonstrate estimation with Copulas; (2) to propose a handy procedure for sample size optimization in the context of continuous forest monitoring using model-assisted estimation; and (3) to compare survey precisions resulting from using Copulas and regression modeling as a function of sample sizes.

2. Materials

2.1. Field data

The study area of approximately 10,836 ha is located in Kou, Burkina Faso with a fragmented landscape of dry savanna due to agricultural land uses (11°41' ~ 11°47'N, 1°53' ~ 1°59'W) (Fig. 1). The target population with 120,756 elements was applied in this study. This study assumes that the sample of 160 plots with 0.1 ha in size was selected by simple random sampling, considered a viable and rational alternative to two-stage sampling as detailed in Appendix A. The plot centers were geo-referenced using Global Navigation Satellite System (GNSS) receivers that have a real-time accuracy of 60 cm from free corrections of Satellite-Based Augmentation Systems supported by European Geostationary Navigation Overlay Service. Selection and field measurements were conducted during the dry season between late November 2013 and early February 2014. The relevant statistics for the sample plots are shown in Table 1. The plot-level variable of interest (VOI) is firewood volume (m^3/ha), aggregated from within-plot woody material that was usable as fuelwood.

The another study area of approximately 632,100 ha is located in Genhe Forest Area (50°25' ~ 51°17'N, 120°41' ~ 122°42'E) in Greater Khingan Mountains (Fig. 1), characterized by temperate deciduous broad-leaved forest and coniferous and broad-leaved mixed forest. The target population, size with 10,049,221 was applied in this study. The sample of 97 plots of 0.06 ha was selected by systematic sampling, and field measurements were conducted in June 2018. Plot centers were geo-referenced using Real-Time Kinematic (RTK) high-precision positioning technology based on GNSS receivers, with a precision level of 5–10 m. The relevant statistics for the sample plots are shown in Table 1. The plot-level variable of interest (VOI) is standing volume (m^3/ha).

2.2. Remotely sensed independent variables

The remotely sensed data for the Burkina Faso (BF) and Genhe (GH) study area were provided by the Landsat 8 Operational Land Imager (OLI) georeferenced to WGS84/UTM Zone 30 N and Sentinel-2 L2A georeferenced to WGS84/UTM Zone 51 N, respectively, which were collected during or close to the time of the field campaign. A single scene for both datasets covered the entire study area. Landsat 8 OLI data with a spatial resolution (or pixel size) of 30 m were Provisional Surface Reflectance product obtained from the U.S. Geological Survey at no charge. The Sentinel-2 L2A data were also acquired free of charge from the European Space Agency, primarily consisting of atmospherically corrected bottom-of-the-atmosphere reflectance data. The Sentinel-2 L2A image pixel size was resampled to 25 m using SNAP software to match the information of sample plots, followed by pre-processing such as band fusion, mosaicing, cropping, etc.

For the Landsat 8 and Sentinel-2 L2A dataset, remotely sensed independent variables were calculated from original spectral bands, such as the first principal component (PCA), Enhanced Vegetation Index (EVI), Soil Adjusted Vegetation Index (SAVI) and Haralick textures. The R-package "rgdal" was used in data processing (Bivand et al., 2010).

3. Methods

3.1. Overview

The primary challenge in conducting monitoring surveys is optimizing sample size for specific precision or cost requirements, which this study sought to address by leveraging data from previous surveys,

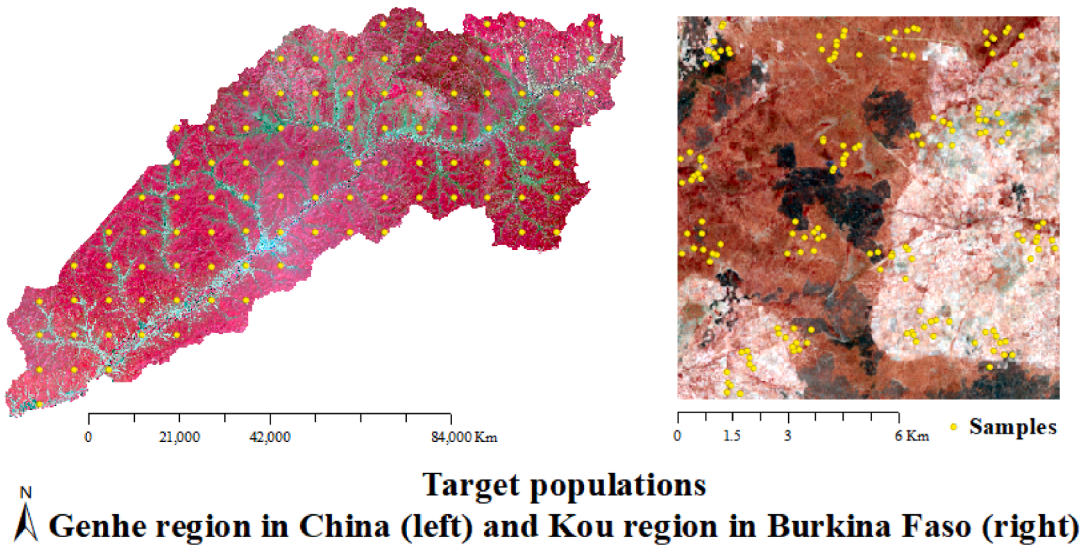


Fig. 1. The target populations and the sample plots.

Table 1
Sample statistics.

Study area	Attributes	Min	Max	Mean	SD
Burkina Faso (BF)	Tree density(stems/ha)	10	1935	494	401
	Mean diameters(cm)	6.4	40	15.2	8.5
	Basal area(m^2/ha)	0.2	16.1	5.6	3.6
	Firewood volume(m^3/ha)	0	29.1	6.6	6.2
Genhe	Mean DBH(cm)	0	29.1	12.2	7.4
(GH)	Mean height(m)	0	22.4	11.7	6.0
	Mean age(a)	0	185	75	45
	Standing volume(m^3/ha)	0	304.4	89.2	68.1

particularly the baseline survey. Assuming that the sample S of 160 or 97 plots is from either the first survey, i.e. the baseline survey, or the (n)th survey in the monitoring survey, then the study context is the relationship between precision and sample size trade-offs for monitoring the (n + 1)th survey over different estimators, where the sample size depends on the results of the (n)th survey. This study considered the model-assisted (MA) estimators supported by two models, WLS and Copulas, as well as the expansion estimator (Section 3.2.1), which is the baseline of the sample size optimization. A flowchart for the study is provided in Fig. 2, and the estimators considered above are listed in Table 2.

3.1.1. *Optimizing sample size with Monte Carlo simulation*

During the optimization procedure of the sample size using expansion estimators, the precision results of other different estimates and sample sizes, calculated based on the baseline survey, are used to develop linear/nonlinear models that could aid in determining sample sizes of the subsequent monitoring surveys.

Prior to sampling, the design variance is usually considered to be the one to use in calculating the sample size. After a particular sample has been selected and data collected, the variance computed under a reasonable model may be more appropriate for inference from that particular sample (Valliant et al., 2018, p. 53).

Table 2
The two cases for expansion (SRS), WLS and Copulas estimators.

Case	Estimator
Expansion estimator for simple random sampling (SRS)	$\hat{\mu}_1, \widehat{Var}(\hat{\mu}_1)$
Model-assisted estimator for WLS or Copulas model	$\hat{\mu}_2, \widehat{Var}(\hat{\mu}_2)$

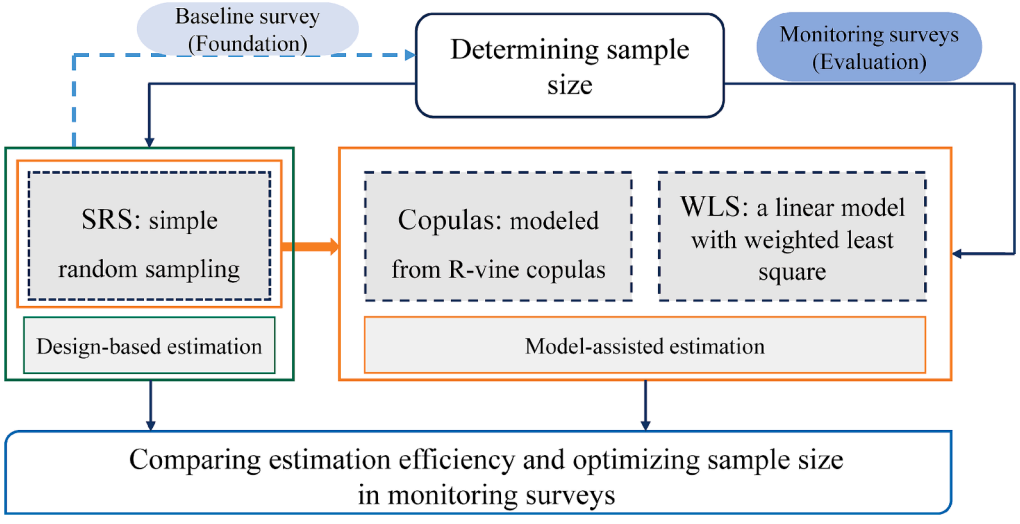


Fig. 2. The flowchart for the study.

Determining sample size gives a confidence interval of a specified width, equivalently twice the maximum error of estimation (Ryan, 2013). Supposing the margin of error is d and the goal is to be within d of the population mean μ with probability $1-\alpha$, this is,

$$\Pr(|\widehat{\mu}-\mu| \le d) = 1-\alpha \tag{1}$$

this is equivalent to setting the half-width of a $100(1-\alpha)\%$ two-sided confidence interval (CI) to $d = z_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{\mu})} = z_{1-\alpha/2}\sqrt{\sigma^2/n}$, assuming that $\widehat{\mu}$ can be treated as being normally distributed, which is shown in Fig. 3. The term $z_{1-\alpha/2}$ is the $100(1-\alpha)$ percentile of the standard normal distribution.

Since the population size N is large relative to the sample size n, the finite population correction factor can be ignored. Thus, a specified maximum error of estimation, considering the margin of error $d = d_s$, Eq.(1) can be manipulated to give the required sample size, n_s, as Thompson (2012),

$$n_s = \frac{z^2\sigma^2}{d_s^2} \tag{2}$$

where z indicates $z_{1-\alpha/2}$, and given the confidence level $1-\alpha = 95\%$, $z = 1.96$; σ^2 is the population variance, which generally is unknown, and was estimated using the designated sample variance, $\widehat{\sigma}^2$, introduced in Section 3.2.1 of this study. From the formula of determining sample size (Eq.(2)), subject to the specified maximum error of estimation, i.e., absolute error, the larger the margin of error d_s, the fewer plots are selected in the sample.

In addition, the Monte Carlo simulation with 1000 iterations was performed in this study to evaluate and compare expansion and model-assisted estimators under different sample sizes n_s, excluding $n_s \in \{160, 97\}$. It worked including four steps as Kalos and Whitlock (2008): (i) define the sample S with $n \in \{160, 97\}$ size as the population of possible inputs; (ii) generate the possible inputs with n_s size through simple random sampling across the population; (iii) calculate the population mean and variance estimates, $\widehat{\mu}_t$ and $\widehat{Var}(\widehat{\mu}_t)$, $t = 1, 2$ as described in Section 3.2, of the inputs obtained from the former step; (iv) aggregate the mean of the mean and variance estimates over respective iterations.

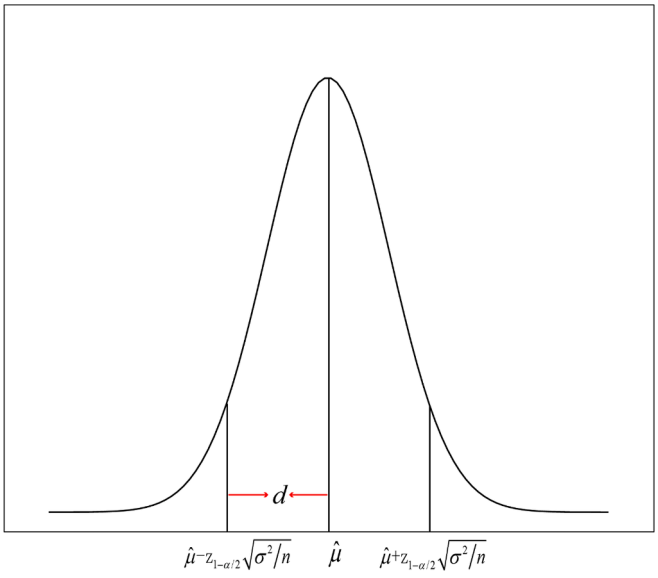


Fig. 3. The value of d, defined as the half-width of a $100(1-\alpha)\%$ two-sided confidence interval (CI).

3.2. Inference and estimators

3.2.1. Expansion estimator

With the probability-based and design-unbiased, the expansion estimator was the standard estimator defined from simple random sampling (SRS) design. Then, $\widehat{\mu}_1$ and its variance, $\widehat{Var}(\widehat{\mu}_1)$, were calculated as,

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3}$$

and

$$\widehat{Var}(\widehat{\mu}_1) = \frac{N-n}{N}\bullet\frac{\widehat{\sigma}^2}{n} \tag{4}$$

where i indexes the n sample observations and y_i is the observation for the i^{th} population element selected for the sample; and $\widehat{\sigma}^2 = \sum_{i=1}^{n}(y_i-\widehat{\mu}_1)^2/(n-1)$ is the sample variance. The finite-population correction factor, $(N-n)/N$, has effect on reducing $\widehat{Var}(\widehat{\mu}_1)$, but can be omitted for populations that are large relative to the sample size.

3.2.2. Model-assisted estimators

Model-assisted estimators based on SRS design were incorporated into the inference for comparison. With model-assisted estimators, an initial estimator of the population parameter of interest, obtained by predicting all population elements with $f(\boldsymbol{X}_i;\boldsymbol{\beta})$, is adjusted for systematic deviations between the predicted and observed values in the sample. A superpopulation model, $f(\boldsymbol{X}_i;\boldsymbol{\beta})$, one based on auxiliary data obtained from remote sensing (Ståhl et al., 2016). This model presumed to represent the underlying relationship between the independent and dependent variables takes the form,

$$y_i = f(\boldsymbol{X}_i;\boldsymbol{\beta}) + \varepsilon_i \tag{5}$$

where $\boldsymbol{\beta}$ indicates the model parameters; i indexes plots or population elements; y_i is the plot-level dependent variable; $\boldsymbol{X}_i$ is a vector of plot-level independent variables with realized values calculated using Landsat 8 auxiliary data that are available wall-to-wall for the population; and ε_i is a random residual term.

The model $f(\boldsymbol{X}_i;\boldsymbol{\beta})$ in model-assisted estimation casts a role for variance reduction but the inference itself is still probability-based for ensuring approximate unbiasedness. Therefore, model-assisted estimation is design-based with a distinguishing feature that approximately design-unbiasedness is hold (Särndal et al., 1992).

By simple random sample for the selection of sample S comprising of n elements, the estimator of μ takes the form,

$$\widehat{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N}\widehat{y}_i + \frac{1}{n}\sum_{i\in S} e_i \tag{6}$$

where $e_i = y_i - \widehat{y}_i$, and $\widehat{y}_i$ predicts y_i. The first term in Eq. (6), $\frac{1}{N}\sum_{i=1}^{N}\widehat{y}_i$, is the mean of the model predictions for all population elements, and the second term, $\frac{1}{n}\sum_{i\in S} e_i$, is an estimate of bias calculated over the sample elements to compensates for systematic model prediction errors. As seen in Särndal et al. (1992) and Gregoire et al. (2016), an asymptotically design-unbiased estimator of the variance of $\widehat{\mu}_2$ is,

$$\widehat{Var}(\widehat{\mu}_2) = \frac{N-n}{N}\bullet\frac{\widehat{\sigma}_e^2}{n} \tag{7}$$

where $\widehat{\sigma}_e^2 = \sum_{i\in S}\left(e_i - \frac{1}{n}\sum_{i\in S} e_i\right)^2/(n-1)$ is the residual variance. The primary benefit of the model-assisted estimator is that it takes advantage of the relationship between the sample observations and their model predictions to reduce the variance of the population mean estimate. It is

obvious from the above formula that the smaller the residual $e_i = y_i - \hat{y}_i$, the smaller the variance of the model-assisted estimator. This suggests that the way to reduce the uncertainty of the model-assisted estimator is to improve the model, thereby reducing the residuals (Hou et al., 2018). Therefore, two modeling techniques, Copulas and WLS, are presented in this study, as detailed in Section 3.3.

3.3. Modeling

As opposed to the conventional linear models, Copulas can capture a nonlinear dependence structure, developed as a more effective form of modeling using the R-vine copula. In this study, the chosen independent variables of Copulas would match those of WLS in order to examine the effectiveness and prediction accuracy of these two models.

For WLS modeling purpose, the selection of independent variables was conducted using the "bootstrap stepAIC" procedure parsimoniously from the extracted large set of Landsat 8 auxiliary variables, which integrates bootstrapping to assess the variability of stepwise model selections based on the Akaike information criterion (AIC), available in the R-package "bootStepAIC" (Rizopoulos, 2022).

3.3.1. Copulas modeling

Due to the flexibility and numerical applicability, vine copula is a choice for modeling Copulas, which uses bivariate copulas as the basic building blocks with vine structures to specify the dependence structure.

Copula is the function that joins or "copula" multivariate distribution functions to their one-dimensional marginal distribution functions (Nelsen, 2006). According the Sklar's theorem, for a m random variables $\boldsymbol{X} = \{X_1, X_2, \cdots, X_m\}$ following a joint distribution F with the j^{th} univariate margin F_j, the copula associated with F is a distribution function C : $[0,1]^m \rightarrow [0,1]$ with $U(0,1)$ margins, that satisfies,

$$F(\boldsymbol{x}) = C(F_1(x_1), \cdots, F_m(x_m)) = C(u_1, \cdots, u_m), \boldsymbol{x} \in \mathbb{R}^m \tag{8}$$

with associated density function $f(\boldsymbol{x}) = c(F_1(x_1), \cdots, F_m(x_m)) \bullet f_1(x_1) \cdots f_m(x_m)$, where c, f, f_j separately denotes the density function of $C, F, F_j, j = 1, \cdots, m$. Then, the function C is a bivariate copula when the dimension $m = 2$, is also unique if $F(\boldsymbol{x})$ is a continuous m-variate cumulative distribution function (CDF).

Vine copulas, originally proposed by Bedford and Cooke (2001), are hierarchical graphical models that decompose a high dimensional copula function into multiple bivariate copula functions, known as pair copulas (Aas et al., 2009), in the form of vines. Compared to the C-vine and D-vine copulas, the R-vine copula offers more versatile dependence structures without imposing constraints on edges in advance, allowing for a more realistic depiction of dependence structures among multiple variables (Yan et al., 2023; Zhang et al., 2014).

A formal definition of Regular vine (R-vine) is provided by Bedford and Cooke (2002), with edges in the first tree representing pairwise dependence and edges in the following trees representing conditional dependence. Here is an example of R-vine structure for $m = 5$ variables shown in Fig. 4. Generally, a regular vine (R-vine) of m variables is a nested ensemble of $m-1$ trees, where the edges of the first tree are nodes of the second tree, the edges of the second tree are nodes of the third tree, etc. (Joe, 2014).

To obtain a vine copula construction, for each edge $[j, k|G] \in E(\mathscr{V})$ in the vine, where $E(\mathscr{V}) = \bigcup_{l=1}^{m-1} E(T_l)$ denoting the set of edges of R-vine structure $\mathscr{V}$, there is a bivariate copula $C_{j,k|G}$ (e.g., three bivariate copula functions as listed in Table 3) associated with it. Then the m-dimensional joint probability density function of m dimensional distribution F in the R-vine copula model can be expressed as

$$f_{1,\cdots,m}(x_1, \cdots, x_m) = \prod_{i=1}^{m} f_i(x_i) \prod_{[j,k|G] \in E(\mathscr{V})} c_{j,k|G}\left(F_{j|G}\left(x_j|\boldsymbol{x}_G\right), F_{k|G}\left(x_k|\boldsymbol{x}_G\right)\right) \tag{9}$$

Such that for each $[j, k|G] \in E(\mathscr{V})$, we have for the distribution function of X_j and X_k given $\boldsymbol{X}_G = \boldsymbol{x}_G$, $F_{j,k|G}(x_j, x_k|\boldsymbol{x}_G) = C_{j,k|G}(F_{j|G}(x_j|\boldsymbol{x}_G), F_{k|G}(x_k|\boldsymbol{x}_G))$. Further the one dimensional margins of F are given by $F_i(x_i), i = 1, \cdots, m$. To get predictions of the corresponding variable with conditioning on the other variables, its inverse form of the distribution function F , i.e., F^{-1} is considered. When $m = 3$, the τ th copula-based conditional quantile function of the variable x_1 can be computed as follows,

$$Q_{x_1}(\tau|x_2, x_3) = F^{-1}(u_1) = F^{-1}\{h^{-1}[h^{-1}(\tau|h(u_3|u_2))u_2]\}$$

where F^{-1} is the inverse of u_1, h^{-1} is the inverse of the copula function, $h(u_3|u_2) = C_{3|2}(u_3|u_2)$. Therefore, the mean value of 500 realizations of $Q_{x_1}(\tau|x_2, x_3)$ with probabilities τ from the uniform distribution in the interval $[0,1]$ was considered to be the best prediction.

Additionally, the technique of selecting appropriate R-vine copula could utilize the R-package "VineCopula" (Nagler et al., 2023) which is based on the *simplifying assumption* that conditional copula functions in the second or higher trees are independent of conditional variables (Acar et al., 2012; Dißmann et al., 2013; Kraus and Czado, 2017).

Considering the multiple regression with dependent variable Y (denoted by X_1),and p independent variables $\{Z_1, \ldots, Z_p\}$ (denoted by $\boldsymbol{X} \backslash \{X_1\}$), individually selected from the field data and the corresponding remotely sensed auxiliary data, the R-vine copula was modeled to find the conditional distribution of the dependent variable, given the independent variables and to get its predictions based on the auxiliary data, which was specified by the R-package "copulareg" (Brant and Haff, 2021). In this study, the optimal copula for each pair was identified from the potential bivariate copulas presented in Table 3, which encompass a range of typical dependence characteristics, such as central, lower-, and upper-tail dependences (Wang et al., 2020; Xu et al., 2016). The selection criterion for the optimal copula is the smallest value of the Akaike

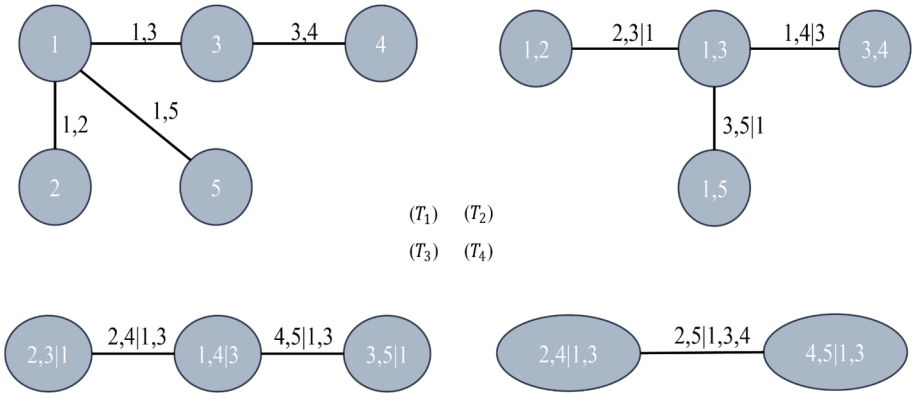


Fig. 4. The example of Regular vine structure of five-dimensional variables containing four trees $T_l, l = 1, 2, 3, 4$.

Table 3
Summary of the three candidate bivariate copula functions for Copulas modeling.

Copula	Copula function $C(u_1, u_2)$	Density function $c(u_1, u_2)$	Parameter θ
Gaussian[a]	$\Phi_\Sigma(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$	$\frac{1}{\sqrt{\lvert\Sigma\rvert}}\exp\left(-\frac{1}{2}\omega^T(\Sigma^{-1}-I_2)\omega\right)$	$[-1,1]$
Gumbel[b]	$\exp\left(-\left[(-\ln u_1)^\theta+(-\ln u_2)^\theta\right]^{1/\theta}\right)$	$\theta^2\exp\{-t_\theta(u_1,u_2)^\alpha\}\frac{(-\ln u_1)^{\theta-1}(-\ln u_2)^{\theta-1}}{t_\theta(u_1,u_2)^2 u_1 u_2}P^G_{2,\alpha}(t_\theta(u_1,u_2)^\alpha)$	$[1,\infty)$
Clayton	$\left(u_1{}^{-\theta}+u_2{}^{-\theta}-1\right)^{-1/\theta}$	$(1+\theta)u_1{}^{-(\theta+1)}u_2{}^{-(\theta+1)}\left(u_1{}^{-\theta}+u_2{}^{-\theta}-1\right)^{-1/\theta-2}$	$(0,\infty)$

[a] $\Sigma=\begin{pmatrix}1 & \theta\\ \theta & 1\end{pmatrix}$ is the correlation matrix, Φ_Σ is a standard bivariate normal distribution and Φ is a standard normal distribution, $\omega^T=(\Phi^{-1}(u_1),\Phi^{-1}(u_2))$, I_2 is the 2×2 identity matrix.

[b] $\alpha=1/\theta$, $P^G_{2,\alpha}(x)=\sum_{k=1}^{2}\varepsilon^G_{2k}(\alpha)x^k$, and $\varepsilon^G_{2k}(\alpha)=\frac{2}{k!}\sum_{j=1}^{k}\binom{k}{j}\binom{\alpha j}{2}(-1)^{2-j}$.

information criterion (AIC), which is expressed as follows: $AIC=-2loglik+2n_k$, where *loglik* is the maximized likelihood for the model, n_k is the number of the copula parameters with $n_k=1$ for all three candidates.

3.3.2. Using WLS model for comparison

The traditional parametric model, WLS, was applied here in the comparison because it is commonly used to extrapolate predictions and takes the heteroscedasticity of model residuals into account. For the model from the Eq. (5), a linear form $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ was constructed where $\boldsymbol{y}$ is the dependent variable; X is the design matrix of independent variables; $\boldsymbol{\beta}$ is the vector of model parameters; and ε is the vector of random errors assumed to be normally distributed with $E(\boldsymbol{\varepsilon})=0$ and $E(\boldsymbol{\varepsilon\varepsilon^T})=\boldsymbol{\phi}$, a positive definite matrix. To accommodate heteroscedasticity, the model parameters were estimated using weighted least squares (WLS) for which $\widehat{\boldsymbol{\beta}}=(\boldsymbol{X^TWX})^{-1}\boldsymbol{X^TWy}$, and $\boldsymbol{W}=\boldsymbol{\phi}^{-1}$ is a diagonal matrix with $w_{ii}=1/\sigma_i^2$. A four-step procedure (McRoberts et al., 2016) was used to estimate the diagonal elements of $\boldsymbol{W}$: (i) The pairs $(y_i,\widehat{y}_i)$ were ranked with respect to $\widehat{y}_i$; (ii) The pairs $(y_i,\widehat{y}_i)$ were assembled into groups of size 20 each; (iii) For each group, h, the mean of the predictions, $\overline{\widehat{y}_h}$, and the variance, $\widehat{\sigma}_h^2$, of the residuals, $\varepsilon_i=y_i-\widehat{y}_i$, were calculated; and (iv) The relationship between $\widehat{\sigma}_h^2$ and $\overline{\widehat{y}_h}$ was modeled as $\widehat{\sigma}_h^2=\lambda\cdot\overline{\widehat{y}_h}+\varepsilon_h$, where λ is the parameter to be estimated.

We were interested in comparing the Copulas with WLS model for the overall prediction accuracy and precision, in terms of population parameters, $\widehat{\mu}$, $\widehat{Var}(\widehat{\mu})$ and sampling precision. The prediction accuracy was also assessed with root mean square error (RMSE) where $RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}$, $RMSE_{\%}=\frac{RMSE}{\overline{y}}\times 100$, and $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$, n is the number of plots of sample S; y_i is the field measured firewood volume; $\widehat{y}_i$ is the predicted firewood volume.

3.4. Sampling precision

Sampling precision was used to evaluate the performance and accuracy of predictions with sample sizes changing. As a relative statistic that measures the degree of variability of data, the coefficient of variation (CV) is the ratio of the standard deviation to its mean, and is used to compare the degree of variability of two or more samples of data with different means. For this study, the coefficient of variation (CV) is chosen to define the sampling precision. Thus, the sampling precision $P(\%)$ is calculated as,

$$P(\%)=(1-cv)\times 100=\left(1-\frac{\sqrt{\widehat{Var}(\widehat{\mu})}}{\widehat{\mu}}\right)\times 100 \tag{10}$$

where cv is the coefficient of variation; $\widehat{\mu}$ and $\widehat{Var}(\widehat{\mu})$ indicate the corresponding estimators, i.e., $\widehat{\mu}_t$ and $\widehat{Var}(\widehat{\mu}_t), t=1,2$.

Typically, sampling precision decreases as the sample size decreases. With the sampling precision as an indication, we compared the predictions of models, WLS, and Copulas, with different sample sizes, to find the optimal model given a smaller sample size in respective inferential frameworks.

4. Results and discussion

4.1. The estimated models: Copulas outperforms WLS

The constructed models, WLS and Copulas, are summarized in Table 4. The graphs of observations versus predictions are presented in Fig. 5. For the both BF and GH study area, the RMSE and $RMSE_{\%}$ of both models displayed promising results, however, the Copulas model showcased a slight advantage in prediction accuracy over the WLS model. Additionally, Copulas exhibited a closer proximity to the diagonal compared to WLS, and showed closer alignment at the mean as well as at the maximum/minimum points. In contrast, WLS only closely aligned at the mean and was slightly out of calibration for the values at both ends (Fig. 5). These suggest that Copulas has better modeling and prediction performance compared to WLS, showing lower levels of saturation. Furthermore, the parsimony criterion was strictly followed during model construction, ensuring simplicity and facilitating meaningful comparisons between the two approaches.

4.2. Copulas outperforms WLS with model-assisted estimators

Copulas outperformed WLS when considering model-assisted estimators in two study areas, BF and GH. Under the condition of the sample S of 160 and 97 plots derived from the previously obtained measurements, the estimates resulting from model-assisted estimators are summarized in Table 5. Because the expansion estimator is unbiased, expansion estimates using simple random sample (SRS) were used as a standard for comparison for estimates obtained with other estimators.

Five findings are relevant. First, compared with the expansion estimator, the model-assisted estimators of the both study areas performed better for the precision, which is consistent with the conclusion in (e.g., Chen et al., 2023; McRoberts et al., 2016, 2013; Ståhl et al., 2016). Second, Copulas with smaller variance estimates demonstrated a higher level of accuracy in estimating the population mean for model-assisted estimators compared to WLS (Table 5). Moreover, the model-assisted estimators with WLS and Copulas model were close to the expansion estimator of the population mean, indicating that model-assisted estimators can mitigate the impact of zero values in the independent/dependent variables and are not prone to underestimation (Breidt and Opsomer, 2017; McRoberts et al., 2013).

Third, model-assisted estimators yielded smaller and more concentrated variance estimates than expansion estimators, making it more consistent and efficient. The model-assisted estimators resulted in slightly lower variance estimates, particularly developed with Copulas (Table 5), which performed the same for the BF and GH study area.

Table 4
Summary of the models.

Study area	Model	RMSE (m^3/ha)	RMSE%	Independent variable	Estimate	Std.Error	t value	Pr($>|t|$)
BF	WLS	4.515	66.660	(Intercept)	−6.22	1.16	−5.36	0.00***
				EVI	11.36	1.03	11.03	0.00***
	Copulas	4.386	64.755	EVI	NA	NA	NA	NA
				NA	NA	NA	NA	NA
GH	WLS	45.002	50.448	(Intercept)	336.10	70.37	4.78	0.00***
				SR.contrast	−2.47	0.49	−5.02	0.00***
				DVI.variance	1.05	0.21	4.92	0.00***
				SAVI.mean	−574.79	167.87	−3.42	0.00***
				band11	−895.21	128.19	−6.98	0.00***
	Copulas	44.551	49.942	SR.contrast	NA	NA	NA	NA
				DVI.variance	NA	NA	NA	NA
				SAVI.mean	NA	NA	NA	NA
				band11	NA	NA	NA	NA

*** : All parameter estimates were statistically significant at$\alpha = 0.001$.

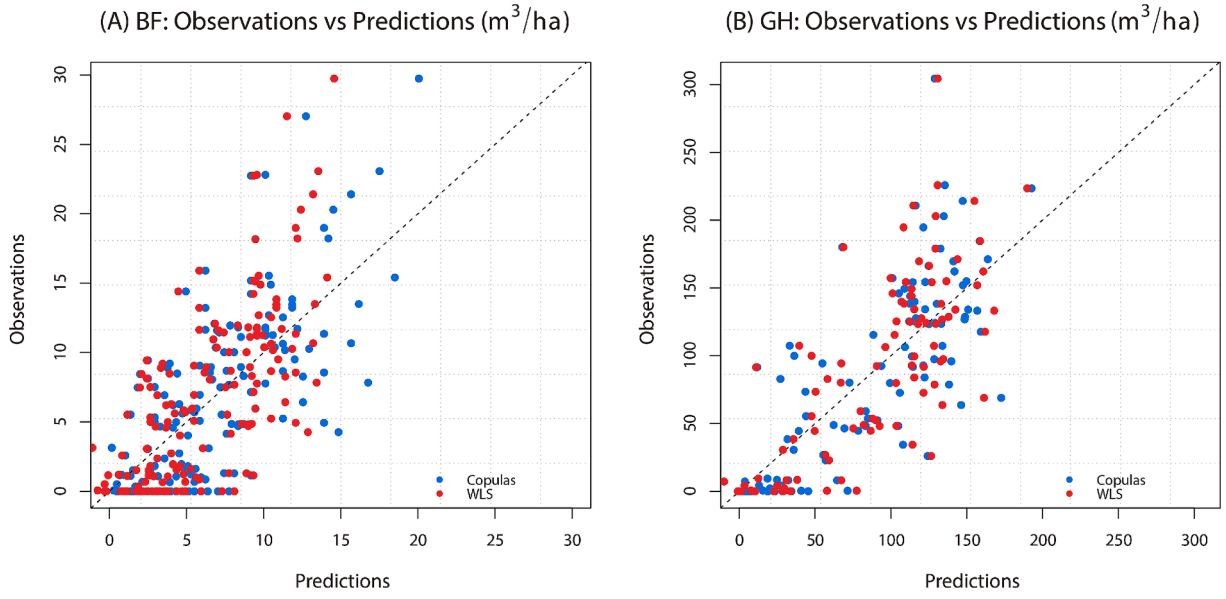


Fig. 5. The graph of observations versus predictions.

Table 5
The summary of estimators using WLS and Copulas model. A sample of 160 and 97 plots for BF and GH study area were used, respectively.

Estimation	Model	$\hat{\mu}$	$\widehat{Var}(\hat{\mu})$	P (%)
BF study area				
Expansion estimation	–	6.774	0.244	92.71
Model-assisted estimation (MA)	WLS	6.598	0.128	94.58
	Copulas	6.387	0.121	94.56
GH study area				
Expansion estimation	–	89.206	47.814	92.25
Model-assisted estimation (MA)	WLS	92.203	21.096	95.02
	Copulas	92.199	21.017	95.03

Fourth, comparing the estimate of the variance estimator $\widehat{Var}(\hat{\mu})$, Copulas exhibited a similar extent of reduction to WLS for both study areas, with the model-assisted estimators (Table 5), suggesting that Copulas could provide a better prediction accuracy, as well as offering unbiased correction from model-assisted. Fifth, using the same sample size in each study area, Copulas consistently displayed larger discrepancies in variance estimates between the expansion and model-assisted estimators compared to WLS (Table 5), indicating that Copulas reduces the uncertainty efficiently. Interestingly, the precision gaps of both study areas between two inferential estimates were nearly identical. This phenomenon continues to unfold in Section 4.4.

4.3. Copulas outperforms WLS with the conditions of different sample sizes

It is crucial to carefully consider the required sample size for monitoring surveys. This determination must be made prior to conducting any field measurements. Opting for a model method that demands the smallest sample size while maintaining the desired inferential precision can result in superior cost-effectiveness. Conversely, when faced with budget limitations, it becomes feasible to identify a modeling approach that fulfills survey precision needs within the constraints of available resources.

With the sample size n_s decreasing, the results of the estimates over 1000 iterations at varying numbers of sample plots for the BF study area are summarized in Table 6, and in Table 7 for the GH study area. The empirical formulas for models of different variance estimates (V) and sample sizes (n_s) for both study areas are depicted in Fig. 6. Four findings are relevant. First, in each sample size scenario for each study areas, the difference between the model-assisted estimators of the population mean for two models, Copulas and WLS, and the corresponding

Table 6
Summary of estimators, including expansion estimator using SRS and model-assisted estimator with WLS and Copulas modeling, at different sample sizes n_s (except for 160) obtained from the corresponding absolution error d_s, with 1000 iterations, for BF study area.

d_s	n_s	Expansion estimation			Model-assisted estimation					
		$\widehat{\mu}$	$\widehat{Var}(\widehat{\mu})$	P (%)	$\widehat{\mu}$		$\widehat{Var}(\widehat{\mu})$		P (%)	
					WLS	Copulas	WLS	Copulas	WLS	Copulas
0.97	160	6.774	0.244	92.71	6.598	6.387	0.128	0.121	94.58	94.56
1.03	140	6.816	0.299	91.97	6.571	6.387	0.151	0.138	94.10	94.18
1.12	120	7.057	0.341	91.72	6.568	6.383	0.182	0.161	93.50	93.71
1.17	110	6.311	0.324	90.98	6.582	6.383	0.201	0.175	93.19	93.45
1.22	100	6.946	0.395	90.95	6.583	6.375	0.223	0.193	92.83	93.11
1.29	90	6.802	0.448	90.16	6.573	6.384	0.253	0.214	92.35	92.75
1.37	80	6.811	0.477	89.86	6.585	6.373	0.284	0.242	91.91	92.28
1.46	70	6.486	0.471	89.42	6.587	6.370	0.332	0.274	91.25	91.78
1.58	60	6.344	0.725	86.57	6.579	6.349	0.379	0.323	90.64	91.06
1.73	50	6.739	0.991	85.22	6.566	6.321	0.463	0.381	89.64	90.24
1.94	40	7.084	1.117	85.08	6.586	6.304	0.570	0.469	88.54	89.14

Table 7
Summary of estimators, including expansion estimator using SRS and model-assisted estimator with WLS and Copulas modeling, at different sample sizes n_s (except for 97) obtained from the corresponding absolution error d_s, with 1000 iterations, for GH study area.

d_s	n_s	Expansion estimation			Model-assisted estimation					
		$\widehat{\mu}$	$\widehat{Var}(\widehat{\mu})$	P (%)	$\widehat{\mu}$		$\widehat{Var}(\widehat{\mu})$		P (%)	
					WLS	Copulas	WLS	Copulas	WLS	Copulas
13.55	97	89.206	47.814	92.25	92.203	92.199	21.096	21.017	95.02	95.03
14.07	90	89.114	52.091	91.90	92.130	92.525	22.681	22.101	94.83	94.92
14.92	80	89.154	58.346	91.43	92.021	93.042	25.461	24.657	94.52	94.66
15.95	70	89.322	66.992	90.84	91.998	93.758	29.344	27.982	94.11	94.36
17.23	60	89.334	77.868	90.12	92.116	94.960	34.193	32.547	93.65	93.99
18.88	50	88.722	93.550	89.10	92.172	94.862	40.552	38.027	93.09	93.50
21.11	40	89.093	117.436	87.84	92.321	95.144	50.742	46.168	92.28	92.86
24.37	30	88.566	156.765	85.86	92.225	96.156	65.436	58.588	91.23	92.04
26.70	25	88.832	189.203	84.52	92.867	96.944	78.178	65.677	90.48	91.64
29.85	20	89.041	232.038	82.89	92.359	96.585	90.610	79.974	89.69	90.74

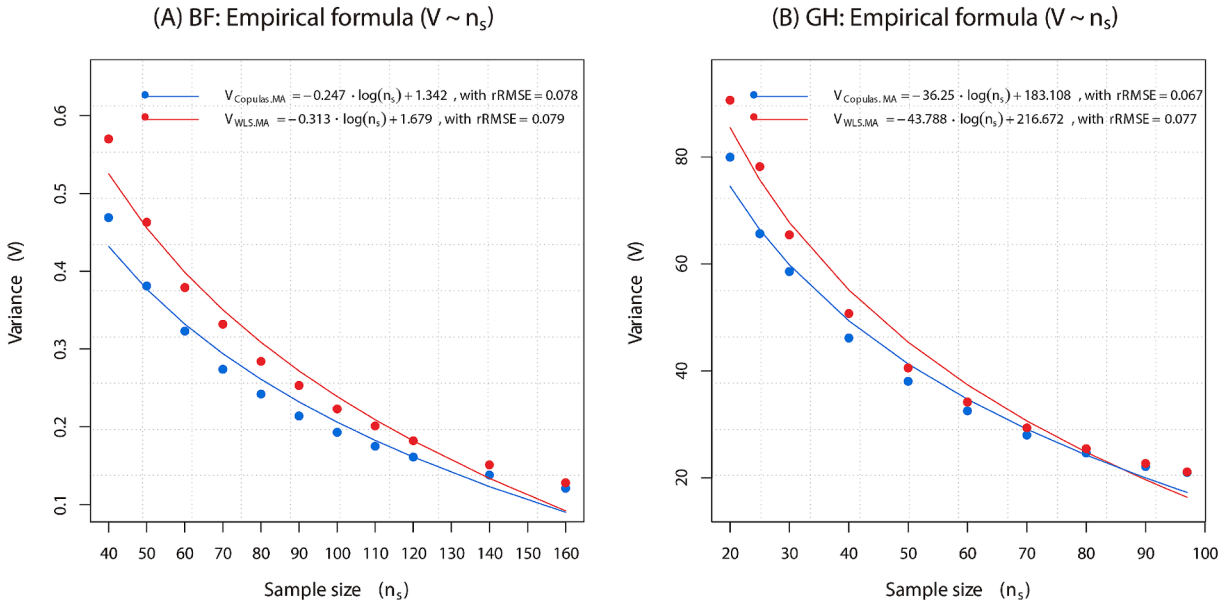


Fig. 6. The empirical formulas of different variance estimates (V) and sample sizes (n_s) for the BF and GH study area.

expansion estimators was significantly smaller than the absolute error d_s (Table 6, 7). This suggests that (1) with the given variance estimator, the formula of determining sample size (Eq. (2)) could offer the appropriate relationship between the absolute error d_s and the sample size n_s for this study; (2) both models, Copulas and WLS, have the accuracy and usability of prediction; (3) model-assisted estimators for both models are relatively stable (i.e. have lower rRMSE) considering the estimates of the population mean. Second, the log-transformed model of variance estimates and sample sizes was developed in this study, which exhibited varying convergence rates for different variance estimates with the

sample size increasing (Fig. 6). As the sample size increased, the smooth fitted curve of Copulas consistently remained below the WLS, in line with the original data point conditions, regardless of BF or GH aspect. This indicates that (1) Copulas has smaller variance estimates at the same sample size than WLS in this study; (2) these modeling formulas could contribute to enhancing the precision and reliability of further survey analyses, serving as a reference for optimizing sample sizes in monitoring surveys.

Third, with model-assisted estimators, when the sample size was 160, Copulas had the lowest variance estimate of 0.121 for BF (Table 6), and the lowest variance estimate of 21.017 for GH at a sample size of 97 (Table 7). In particular, the uncertainty of all Copulas and WLS estimation obtained over 1000 iterations in the BF study area was elevated to varying degrees when the sample size varied from 60 to 50, with Copulas at 18 % for model-assisted, roughly, and WLS at approximately 22 %, respectively. Similarly, in the GH study area, Copulas and WLS elevated approximately 22 % and 16 % when the sample size varied from 25 to 20. It suggests that there would be a significant turnaround in inferential precision as the sample size decreased in this study. Fourth, as the sample size decreases, the variance estimates of Copulas remained smaller than WLS, regardless of BF or GH study area, indicating that (1) with the conditions of different sample size, Copulas outperforms WLS in overall prediction; (2) Copulas may offer more stable predictions with lower uncertainty, making it a favorable choice for meeting the demands of inferential precision, especially in small sample scenarios.

4.4. The comparison of sampling precision between Copulas and WLS

In contrast to the design-based inference without auxiliary information, which relies on a sufficient sample size, model-based inference enhances inferential precision by utilizing an accurate model (McRoberts, 2006; Saarela et al., 2015; Zheng et al., 2024). Furthermore, model-assisted estimation integrates the model and probability samples to guarantee both unbiasedness and precision (Opsomer et al., 2007; Ståhl et al., 2016). Hence, the design effect (Särndal et al., 1992, p. 54), defined as the variance ratio, $\frac{\widehat{Var}(\widehat{\mu}_2)}{\widehat{Var}(\widehat{\mu}_1)}$ was also employed to evaluate the sampling precision of Copulas and WLS with model-assisted estimators, in comparison to the reference that use the variance estimate $\widehat{Var}(\widehat{\mu}_1)$ of the expansion estimator.

Then, the results of precision with sample size changing for two study areas are summarized in Table 6 and Table 7, and the corresponding trends of the design effect and sampling precision for different estimators are shown in Fig. 7. Four findings are relevant. The general minimum standard for NFI is a sampling precision of 90 % (Wang et al., 2018; Zhou et al., 2018). First, Copulas and WLS yielded the better sampling precision than the expansion estimation. When the sample size was 60 in the GH study area, the precision of expansion estimation was

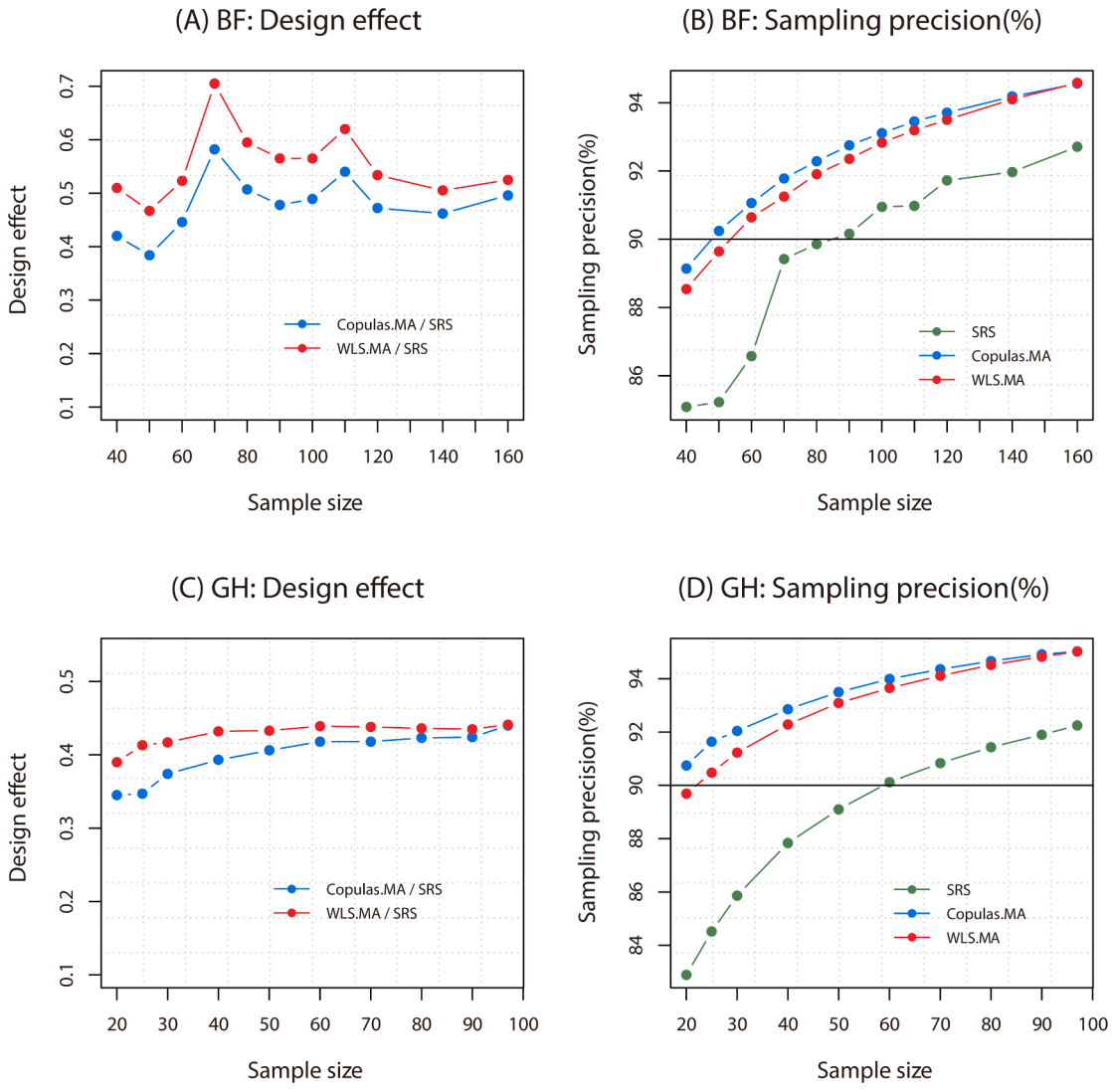


Fig. 7. Effects of sample size on the design effect (on the left) of estimators and the sampling precision (%) (on the right) for both BF and GH study area.

approximately 90 %, while Copulas and WLS were both about 94 % (Table 7). In fact, the precision of expansion estimation was approximately 90 % when the sample size was decreased to 90, whereas the precision of Copulas and WLS ranged from around 92 % to 93 %, representing the precision of expansion estimation at the maximum sample size of 160 for the BF study area (Table 6). It suggests that developing suitable model forms could aid to evaluate the efficiency and robustness of sampling strategy (Saarela et al., 2015) and optimize sample sizes.

Second, the design effects were all less than the unit value (Fig. 7A, C), implying that both Copulas and WLS gained precision over the expansion estimation. Additionally, the variance ratios of Copulas were closer to the zero value, thus suggesting the better accuracy. Third, Copulas was more efficient at small sample size, rather than WLS, similarly at the large sample size. The precision versus sample size (Fig. 7B, D) revealed that (1) the precision of Copulas remained at the highest level at different sample sizes; (2) the trend in precision is slightly smoother for model-assisted estimation, either Copulas or WLS; (3) for BF study area, at a sample size of 60, a distinct turning point is evident. Below 60, the sample size was too small to highlight the difference of precision between the different estimators, and once this threshold was surpassed, Copulas for the model-assisted estimators was notably superior to that of the other residual estimators, which was consistent with the GH study area at the threshold of 25. This phenomenon is more pronounced for the design effect in Fig. 7A, C. That is, the copula method demonstrates strong performance of prediction in small sample sizes, achieving high levels of accuracy (Grover et al., 2020). Fourth, Copulas consistently outperformed WLS with the same sample size. Conversely, employing the Copulas model reduces the sample size compared to WLS while maintaining the same level of precision. As the sample size decreased from 160 to 40 and 97 to 20, the precision of Copulas varied from 94.56 % to 89.14 %, and the precision of WLS varied from 94.58 % to 88.54 % in BF; and 95.03 % to 90.74 % for Copulas, 95.02 % to 89.69 % for WLS in GH (Table 6, 7). For BF study area, WLS achieved the precision threshold of 90 % with a sample size of 60, whereas Copulas achieved with a sample size of 50. The sample size of Copulas and WLS for GH study area were 20 and 25, respectively, when achieving the precision threshold of 90 %. These findings suggest that (1) Copulas could reduce the sample size by about 20 % than WLS at the minimum threshold requirement of 90 % precision; compared to WLS, (2) Copulas is a better choice for solving small sample problems, which could balance tradeoffs between cost, accuracy and flexibility.

5. Conclusions

This study proposed and demonstrated eight relevant conclusions: (1) Copulas is highly superior to WLS in modeling and prediction, with its prediction points more closely aligned with the observations at the mean, minimum, and maximum values; (2) Copulas with model-assisted estimator yielded the smallest variance estimate at the sample size of 160 or 97; as sample size reduces, (3) the Copulas still performs more effectively and accurately in monitoring surveys, compared to the WLS under the model-assisted estimation, as well as the expansion estimation; (4) a simple method for optimizing sample size, provided a log-transformed model exhibiting that Copulas has smaller variance estimates than WLS at any same sample size; (5) the difference of the population mean between the Copulas and WLS estimators and the expansion estimators was smaller than the corresponding absolute error, signifying that Copulas and WLS possess the accuracy and usability for prediction; (6) Copulas consistently outperformed WLS in terms of sampling precision, that is, as the sample size decreased in the GH study area, the precision of Copulas decreased from 95.03 % to 90.74 %, whereas it varied from 95.02 % to 89.69 % for WLS, which was similarly found in the BF study area; (7) beyond a distinct turning point at a certain sample size for two study areas, the precision of the model-assisted estimators with Copulas was notably superior to that of the other estimators; (8) Copulas reduces by 1/6 or 1/5 the sample sizes as compared to WLS to reach a precision threshold of 90 %, indicating that Copulas remains a highly effective tool for small sample sizes. Overall, Copulas could adequately fulfill the requirements for accuracy, cost-effectiveness and flexibility of monitoring surveys, especially for situations with small sample sizes.

CRediT authorship contribution statement

Xinjie Cheng: Writing – review & editing, Writing – original draft, Software, Formal analysis, Data curation. **Zhengyang Hou:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Annika Kangas:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis. **Jean-Pierre Renaud:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis. **Hao Tang:** Writing – original draft, Validation, Investigation, Formal analysis. **Weisheng Zeng:** Writing – original draft, Validation, Investigation, Formal analysis. **Qing Xu:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2023YFF1304002-05); and the National Social Science Fund of China (Grant No. 22BTJ005). Dr. Qing Xu was also supported by the National Natural Science Foundation of China (Grant No. 32001252); and the International Center for Bamboo and Rattan (Grant No. 1632020029, Grant No. 1632021024 and Grant No.1632022024). We are grateful to the international project, Building Biocarbon and Rural Development in West Africa (BIODEV), and our Finnish colleague, Dr. Janne Heiskanen, for organizing the field campaign in Burkina Faso.

Appendix A

Horvitz-Thompson (HT) estimator has a general form of

$$\widehat{\mu}_{\pi} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} \tag{A.1}$$

with an unbiased estimator for its variance as,

$$\widehat{Var}(\hat{\mu}_{\pi}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\left(\frac{\Delta_{kl}}{\pi_{kl}} \bullet \frac{y_k}{\pi_k} \bullet \frac{y_l}{\pi_l}\right) \tag{A.2}$$

where $\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k\pi_l, k \neq l \\ \pi_k(1-\pi_k), k = l \end{cases}$, π_k is the inclusion probability for element k, $\pi_k = \frac{n}{N}$ for simple random sampling (SRS); $\pi_k = \frac{n}{N} = \frac{c \bullet s}{C \bullet S}$, i.e., $N = \sum_i^C S_i$, $n = \sum_i^c s_i$ for two-stage sampling.

Two-stage (TS) estimator takes the form

$$\hat{\mu}_{TS} = \frac{1}{N} \bullet \frac{C}{c} \bullet \sum_i^c \hat{\tau}_i \tag{A.3}$$

where $\hat{\tau}_i = \frac{S_i}{s_i}\sum_k^{s_i} y_{ik}$ estimates the total y-value of the i-th PSU. Its variance estimator is

$$\widehat{Var}(\hat{\mu}_{TS}) = \frac{1}{N^2}\left[C(C-c)\frac{\sigma_I^2}{c}\right] + \frac{1}{N^2}\left[\frac{C}{c} \bullet \sum_i^c S_i(S_i - s_i)\frac{\sigma_i^2}{s_i}\right] \tag{A.4}$$

where $\sigma_I^2 = \sum_i^c (\hat{\tau}_i - \bar{\hat{\tau}}_i)^2/(c-1)$, with $\bar{\hat{\tau}}_i = \sum_i^c \hat{\tau}_i/c$; and $\sigma_i^2 = \sum_k^{s_i}(y_{ik} - \bar{y}_i)^2/(s_i - 1)$, with $\bar{y}_i = \sum_k^{s_i} y_{ik}/s_i$.

Remark:

(I) The variance estimator always has $\widehat{Var}(\hat{\mu}_{TS}) > \widehat{Var}(\hat{\mu}_{\pi})$*forSRS*, making the sample size determination using SRS estimators is stricter than using TS;

(II) The mean estimators for SRS, TS in Eq.(A.1) and (A.3) are the same;

(III) Especially, let $S = 1$, there is $S = s = 1$, i.e., $N = C, c = n$, such that the second term of Eq.(A.4) equals zero, and the first term now expresses as

$$\frac{1}{N^2}\left[C(C-c)\frac{\sigma_I^2}{c}\right] = \frac{1}{N^2}\left[N(N-n)\frac{\sigma_I^2}{n}\right] = \frac{N-n}{N} \bullet \frac{\sigma_I^2}{n}$$

where $\sigma_I^2 = \sum_i^n (y_i - \bar{y})^2/(n-1)$, with $\bar{y} = \sum_i^n y_i/n$. This equals to the SRS variance estimators.

Data availability

Data will be made available on request.

References

Aas, K., Czado, C., Frigessi, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. Insur. Math. Econ. 44, 182–198. https://doi.org/10.1016/j.insmatheco.2007.02.001.

Acar, E.F., Genest, C., Nešlehová, J., 2012. Beyond simplified pair-copula constructions. J. Multivar. Anal Special Issue on Copula Modeling and Dependence 110, 74–90. https://doi.org/10.1016/j.jmva.2012.02.001.

Bedford, T., Cooke, R.M., 2001. Probability density decomposition for conditionally dependent random variables modeled by vines. Ann. Math. Artif. Intell. 32, 245–268. https://doi.org/10.1023/A:1016725902970.

Bedford, T., Cooke, R.M., 2002. Vines–a new graphical model for dependent random variables. Ann. Stat. 30. https://doi.org/10.1214/aos/1031689016.

Bhatti, M.I., Do, H.Q., 2019. Recent development in copula and its applications to the energy, forestry and environmental sciences. Int. J. Hydrogen Energy 44, 19453–19473. https://doi.org/10.1016/j.ijhydene.2019.06.015.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., 2010. rgdal: Bindings for the Geospatial Data Abstraction Library.

Brant, S.B., Haff, I.H., 2021. copulareg: Copula Regression.

Breidt, F.J., Opsomer, J.D., 2017. Model-assisted survey estimation with modern prediction techniques. Stat. Sci. 32, 190–205. https://doi.org/10.1214/16-STS589.

Charles, P., Giraudeau, B., Dechartres, A., Baron, G., Ravaud, P., 2009. Reporting of sample size calculation in randomised controlled trials: review. BMJ 338, b1732–b. https://doi.org/10.1136/bmj.b1732.

Chen, F., Hou, Z., Saarela, S., McRoberts, R.E., Ståhl, G., Kangas, A., Packalen, P., Li, B., Xu, Q., 2023. Leveraging remotely sensed non-wall-to-wall data for wall-to-wall upscaling in forest inventory. Int. J. Appl. Earth Obs. Geoinformation 119, 103314. https://doi.org/10.1016/j.jag.2023.103314.

Cochran, W., 1977. Sampling Techniques, 3rd ed. Wiley, New York.

Dißmann, J., Brechmann, E.C., Czado, C., Kurowicka, D., 2013. Selecting and estimating regular vine copulae and application to financial returns. Comput. Stat. Data Anal. 59, 52–69. https://doi.org/10.1016/j.csda.2012.08.010.

Ehlers, S., Grafström, A., Nyström, K., Olsson, H., Ståhl, G., 2013. Data assimilation in stand-level forest inventories. Can. J. for. Res. 43, 1104–1113. https://doi.org/10.1139/cjfr-2013-0250.

Ene, L.T., Næsset, E., Gobakken, T., Gregoire, T.G., Ståhl, G., Nelson, R., 2012. Assessing the accuracy of regional LiDAR-based biomass estimation using a simulation approach. Remote Sens. Environ. 123, 579–592. https://doi.org/10.1016/j.rse.2012.04.017.

Faber, J., Fonseca, L., 2014. How sample size influences research outcomes. Dent. Press J. Orthod. 19, 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo.

Frazer, G.W., Magnussen, S., Wulder, M.A., Niemann, K.O., 2011. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. Remote Sens. Environ. 115, 636–649. https://doi.org/10.1016/j.rse.2010.10.008.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. for. Res. 28, 1429–1447. https://doi.org/10.1139/x98-166.

Gregoire, T.G., Næsset, E., McRoberts, R.E., Ståhl, G., Andersen, H.-E., Gobakken, T., Ene, L., Nelson, R., 2016. Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass. Remote Sens. Environ. 173, 98–108. https://doi.org/10.1016/j.rse.2015.11.012.

Grover, K., Acar, E.F., Torabi, M., 2020. Copula-based predictions in small area estimation. Can. J. Stat. 48, 685–711. https://doi.org/10.1002/cjs.11558.

Hou, Z., McRoberts, R.E., Ståhl, G., Packalen, P., Greenberg, J.A., Xu, Q., 2018. How much can natural resource inventory benefit from finer resolution auxiliary data? Remote Sens. Environ. 209, 31–40. https://doi.org/10.1016/j.rse.2018.02.039.

Hou, Z., Mehtätalo, L., McRoberts, R.E., Ståhl, G., Tokola, T., Rana, P., Siipilehto, J., Xu, Q., 2019. Remote sensing-assisted data assimilation and simultaneous inference for forest inventory. Remote Sens. Environ. 234, 111431. https://doi.org/10.1016/j.rse.2019.111431.

Hou, Z., Domke, G.M., Russell, M.B., Coulston, J.W., Nelson, M.D., Xu, Q., McRoberts, R. E., 2021. Updating annual state- and county-level forest inventory estimates with data assimilation and FIA data. For. Ecol. Manag. 483, 118777. https://doi.org/10.1016/j.foreco.2020.118777.

Hou, Z., Yuan, K., Ståhl, G., McRoberts, R.E., Kangas, A., Tang, H., Jiang, J., Meng, J., Xu, Q., Li, Z., 2023. Conjugating remotely sensed data assimilation and model-assisted estimation for efficient multivariate forest inventory. Remote Sens. Environ. 299, 113854. https://doi.org/10.1016/j.rse.2023.113854.

Houssou, R., Augustin, M.-C., Rappos, E., Bonvin, V., Robert-Nicoud, S., 2022. Generation and Simulation of Synthetic Datasets with Copulas. https://doi.org/10.48550/arXiv.2203.17250.

Joe, H., 2014. Dependence Modeling with Copulas. CRC Press.

Kalos, M.H., Whitlock, P.A., 2008. Monte Carlo Methods | Wiley Online Books. John Wiley & Sons Ltd.

Kangas, A., Myllymäki, M., Gobakken, T., Næsset, E., 2016. Model-assisted forest inventory with parametric, semiparametric, and nonparametric models. Can. J. for. Res. 46, 855–868. https://doi.org/10.1139/cjfr-2015-0504.

Kangas, A., Gobakken, T., Næsset, E., 2020. Benefits of past inventory data as prior information for the current inventory. For. Ecosyst. 7, 20. https://doi.org/10.1186/s40663-020-00231-6.

Kraus, D., Czado, C., 2017. D-vine copula based quantile regression. Comput. Stat. Data Anal. 110, 1–18. https://doi.org/10.1016/j.csda.2016.12.009.

Kumar, P., Shoukri, M.M., 2007. Copula based prediction models: an application to an aortic regurgitation study. BMC Med. Res. Methodol. 7, 21. https://doi.org/10.1186/1471-2288-7-21.
Lin, H., Chaganty, N.R., 2021. Multivariate distributions of correlated binary variables generated by pair-copulas. J Stat Distrib App 8, 4. https://doi.org/10.1186/s40488-021-00118-z.
McRoberts, R.E., 2006. A model-based approach to estimating forest area. Remote Sens. Environ. 103, 56–66. https://doi.org/10.1016/j.rse.2006.03.005.
McRoberts, R.E., Næsset, E., Gobakken, T., 2013. Inference for lidar-assisted estimation of forest growing stock volume. Remote Sens. Environ. 128, 268–275. https://doi.org/10.1016/j.rse.2012.10.007.
McRoberts, R.E., Næsset, E., Gobakken, T., 2014. Estimation for inaccessible and non-sampled forest areas using model-based inference and remotely sensed auxiliary information. Remote Sens. Environ. 154, 226–233. https://doi.org/10.1016/j.rse.2014.08.028.
McRoberts, R.E., Chen, Q., Domke, G.M., Ståhl, G., Saarela, S., Westfall, J.A., 2016. Hybrid estimators for mean aboveground carbon per unit area. For. Ecol. Manag. 378, 44–56. https://doi.org/10.1016/j.foreco.2016.07.007.
McRoberts, R.E., Næsset, E., Heikkinen, J., Chen, Q., Strimbu, V., Esteban, J., Hou, Z., Giannetti, F., Mohammadi, J., Chirici, G., 2022. On the model-assisted regression estimators using remotely sensed auxiliary data. Remote Sens. Environ. 281, 113168. https://doi.org/10.1016/j.rse.2022.113168.
Mohamedou, C., Kangas, A., Hamedianfar, A., Vauhkonen, J., 2022. Potential of Bayesian formalism for the fusion and assimilation of sequential forestry data in time and space. Can. J. for. Res. 52. https://doi.org/10.1139/cjfr-2021-0145.
Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E.C., Graeler, B., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., Killiches, M., Joe, H., Vatter, T., 2023. VineCopula: Statistical Inference of Vine Copulas.
Nedyalkova, D., Tille, Y., 2008. Optimal sampling and estimation strategies under the linear model. Biometrika 95, 521–537. https://doi.org/10.1093/biomet/asn027.
Nelsen, R.B., 2006. An introduction to copulas. Springer series in statistics, 2nd ed. Springer, New York.
Opsomer, J., Breidt, F., Moisen, G., Kauermann, G., 2007. Model-assisted estimation of forest resources with generalized additive models. J. Am. Stat. Assoc. 102, 400–409. https://doi.org/10.1198/016214506000001491.
Räty, M., Kuronen, M., Myllymäki, M., Kangas, A., Mäkisara, K., Heikkinen, J., 2020. Comparison of the local pivotal method and systematic sampling for national forest inventories. For. Ecosyst. 7, 54. https://doi.org/10.1186/s40663-020-00266-9.
Rizopoulos, D., 2022. bootStepAIC: Bootstrap stepAIC.
Ryan, T.P., 2013. Sample Size Determination and Power, 1st ed, Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9781118439241.
Saarela, S., Schnell, S., Grafström, A., Tuominen, S., Nordkvist, K., Hyyppä, J., Kangas, A., Ståhl, G., 2015. Effects of sample size and model form on the accuracy of model-based estimators of growing stock volume. Can. J. for. Res. 45, 1524–1534. https://doi.org/10.1139/cjfr-2015-0077.
Särndal, C.E., Swensson, B., Wretman, J.H., 1992. Model Assisted Survey Sampling. Springer, New York, p. 54.
Ståhl, G., Holm, S., Gregoire, T.G., Gobakken, T., Næsset, E., Nelson, R., 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County. Norway. Can. J. for. Res. 41, 96–107. https://doi.org/10.1139/X10-161.
Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E., Gregoire, T.G., 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. For. Ecosyst. 3, 5. https://doi.org/10.1186/s40663-016-0064-9.
Thompson, S.K., 2012. Sampling. John Wiley & Sons.
Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E., 2010. National Forest Inventories: Pathways for Common Reporting. Springer, Netherlands, Dordrecht, 10.1007/978-90-481-3233-1.
Valliant, R., Dever, J.A., Kreuter, F., 2018. Practical Tools for Designing and Weighting Survey Samples, Statistics for Social and Behavioral Sciences. Springer International Publishing, Cham, p. 53, 10.1007/978-3-319-93632-1.
Wang, M.-X., Huang, D., Wang, G., Du, W., Li, D.-Q., 2020. Vine copula-based dependence modeling of multivariate ground-motion intensity measures and the impact on probabilistic seismic slope displacement hazard analysis. Bull. Seismol. Soc. Am. 110, 2967–2990. https://doi.org/10.1785/0120190244.
Wang, Y., Liu, L., Shangguan, Z., 2018. Dynamics of forest biomass carbon stocks from 1949 to 2008 in Henan Province, east-central China. J. for. Res. 29, 439–448. https://doi.org/10.1007/s11676-017-0459-7.
Xu, Q., Li, B., Maltamo, M., Tokola, T., Hou, Z., 2019. Predicting tree diameter using allometry described by non-parametric locally-estimated copulas from tree dimensions derived from airborne laser scanning. For. Ecol. Manag. 434, 205–212. https://doi.org/10.1016/j.foreco.2018.12.020.
Xu, Q., Li, B., McRoberts, R.E., Li, Z., Hou, Z., 2023. Harnessing data assimilation and spatial autocorrelation for forest inventory. Remote Sens. Environ. 288, 113488. https://doi.org/10.1016/j.rse.2023.113488.
Xu, Y., Tang, X., Wang, J.P., Kuo-Chen, H., 2016. Copula-based joint probability function for PGA and CAV: a case study from Taiwan. Earthq. Eng. Struct. Dyn. 45, 2123–2136. https://doi.org/10.1002/eqe.2748.
Yan, D., Zhao, T., Xu, L., Zuo, L., Wen, H., Ren, J., 2023. Statistical modeling of multivariate loess properties in Taiyuan using regular vine copula with optimized tree structure. Transp. Geotech. 41, 101025. https://doi.org/10.1016/j.trgeo.2023.101025.
Zhang, B., Wei, Y., Yu, J., Lai, X., Peng, Z., 2014. Forecasting VaR and ES of stock index portfolio: A Vine copula method. Phys. Stat. Mech. Its Appl. 416, 112–124. https://doi.org/10.1016/j.physa.2014.08.043.
Zheng, Y., Hou, Z., Ståhl, G., McRoberts, R.E., Zeng, W., Næsset, E., Gobakken, T., Li, B., Xu, Q., 2024. Nexus of certain model-based estimators in remote sensing forest inventory. Forest Ecosyst. 11, 100245. https://doi.org/10.1016/j.fecs.2024.100245.
Zhou, R., Wu, D., Fang, L., Xu, A., Lou, X., 2018. A Levenberg–Marquardt backpropagation neural network for predicting forest growing stock based on the least-squares equation fitting parameters. Forests 9, 757. https://doi.org/10.3390/f9120757.