Modelling forest volume with small area estimation of forest inventory using GEDI footprints as auxiliary information

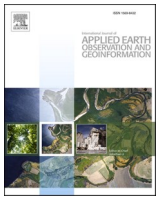
Shaohui Zhang, Cédric Vega, Christine Deleuze, Sylvie Durrieu, Pierre M Barbillon, Olivier Bouriaud, Jean-Pierre Renaud

International Journal of Applied Earth Observations and Geoinformation 114 (2022) 103072

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



Modelling forest volume with small area estimation of forest inventory using GEDI footprints as auxiliary information

Shaohui Zhang [a,b,*], Cédric Vega [c], Christine Deleuze [d], Sylvie Durrieu [e], Pierre Barbillon [f], Olivier Bouriaud [g,c], Jean-Pierre Renaud [b,c]

[a] University of Eastern Finland, Yliopistokatu 7, 80130 Joensuu, Finland
[b] Office National des Forêts, 8 Allée de Longchamp, 54600 Villers lès Nancy, France
[c] Institut National de L'information Géographique et Forestière (IGN), Laboratoire d'inventaire forestier, 14 Rue Girardet, 54000 Nancy, France
[d] Office National des Forêts, 21 rue du Muguet, 39100 Dole, France
[e] TETIS, Inrae, AgroParisTech, CIRAD, CNRS, Univ Montpellier, 500 Rue Jean-François Breton, 34196 Montpellier, France
[f] UMR MIA-Paris, Université Paris-Saclay, AgroParisTech, INRAE 16 Rue Claude Bernard, 75 231 Paris Cedex 05, France
[g] Ștefan cel Mare University of Suceava, 13 University Street, 720229 Suceava, Romania

ARTICLE INFO

Keywords:
Forest inventory
Small area estimation
LiDAR
GEDI
Semivariogram
Forest dominant height

ABSTRACT

The French National Forest Inventory provides detailed forest information up to large national and regional scales. Forest inventory for small areas of interest within a large population is equally important for decision making, such as for local forest planning and management purposes. However, sampling these small areas with sufficient ground plots is often not cost efficient. In response, small area estimation has gained increasing popularity in forest inventory. It consists of a set of techniques that enables predictions of forest attributes of subpopulation with the help of auxiliary information that compensates for the small field samples.

Common sources of auxiliary information usually come from remote sensing technology, such as airborne laser scanning and satellite imagery. The newly launched NASA's Global Ecosystem Dynamics Investigation (GEDI), a full waveform Lidar instrument, provides an unprecedented opportunity of collecting large-scale and dense forest sample plots given its sampling frequency and spatial coverage. However, the geolocation uncertainty associated with GEDI footprints create important challenges for their use for small area estimations.

In this study, we designed a process that provides NFI measurements at plot level with GEDI auxiliary information from nearby footprints. We demonstrated that GEDI RH_{98} is equivalent to NFI dominant height at plot level. We stressed the importance of pairing NFI plots with nearby GEDI footprints, based on not only the distance in between but also their similarities, i.e., forest heights and forest types. Subsequently, these NFI-GEDI pairs were used for small area estimations following a two-phase sampling scheme. We showcased that, with an adequate sample size, small area estimation with GEDI auxiliary data can improve the accuracy of forest volume estimates.

1. Introduction

National Forest Inventories (NFIs) play an important role in understanding the state of forests at the national and regional levels. NFIs are based on data collected in the field at the level of a set of forest plots that are spatially distributed according to a specific sampling design. Forest inventory for small territorial areas, such as municipalities, is also important for decision-makers; however, the information is often relatively limited at this level. As a result, developing small area estimation (SAE) approaches has gained increasing popularity in the field of forest inventory (Hill et al., 2021). It enables prediction of forest attributes, including forest volume, for small areas by using regression models based on auxiliary data commonly derived from remote sensing techniques over the Area of Interest (AOI). It has been reported that SAE can improve forest inventory precision without increasing costs (Mandallaz et al., 2013) and may produce reliable predictions of forest attributes locally, even when field plots are only available outside the small area under assessment (Molina & Marhuenda, 2015).

* Corresponding author at: University of Eastern Finland, Yliopistokatu 7, 80130 Joensuu, Finland.
E-mail addresses: shaohui.zhang@uef.fi (S. Zhang), cedric.vega@ign.fr (C. Vega), christine.deleuze@onf.fr (C. Deleuze), sylvie.durrieu@inrae.fr (S. Durrieu), pierre.barbillon@agroparistech.fr (P. Barbillon), olivier.bouriaud@ign.fr, obouriaud@usm.ro (O. Bouriaud), jean-pierre.renaud-02@onf.fr (J.-P. Renaud).

https://doi.org/10.1016/j.jag.2022.103072
Received 31 May 2022; Received in revised form 11 October 2022; Accepted 18 October 2022
Available online 22 October 2022

Tomppo (2006) is a pioneer in the use of auxiliary data for multi-source forest inventory. Previously, common sources of auxiliary data used in SAE often came from satellite-based imagery (McRoberts et al., 2007), digital aerial photogrammetry (Breidenbach et al., 2018), and airborne laser scanning (Magnussen et al., 2014). The newly launched NASA's Global Ecosystem Dynamics Investigation (GEDI) is a full waveform LiDAR instrument aboard the International Space Station (ISS), which will produce footprint measurements covering over 4 % of the global land surface at the end of the mission (Dubayah et al., 2020). This provides an unprecedented opportunity of systematically collecting samples of forest information that can be used in a SAE approach on a large scale. However, large geolocation uncertainties associated with GEDI footprints (Dubayah et al., 2020; Roy et al., 2021) may impact their usability in estimating forest attributes in fragmented or heterogeneous forests.

The objective of this study was therefore to explore the possibility of using GEDI auxiliary data in a design-based model-assisted approach to improve forest inventory precision and accuracy for a large natural area in France (Sologne) as well as for small areas defined by administrative boundaries (departments). The results were evaluated against estimations obtained using a simple random sampling (SRS) design to assess the efficiency brought by using GEDI as auxiliary data informing about the forest structure.

2. Material

2.1. Study area

Our study is in Sologne, Central France, which covers an area of approximately 6000 km^2 (Fig. 1a). The topography is mostly flat, with most elevations falling within the range of 70–180 m. The climate is temperate Atlantic, with mean annual temperature and precipitation of 11 °C and 725 mm. Forests cover approximately 48 % of the area and are dominated by broadleaved stands (75.3 %). Conifer and mixed stands account for 15.5 % and 9.2 % of the forest areas respectively.

2.2. NFI data

Surveyed between 2015 and 2019, 635 NFI plots were available over the study area. Such a 5-year time interval is routinely used for the official French NFI statistics, and it allows the shortest time gap of data acquisition between NFI surveys and GEDI auxiliary data. Details of inventory schemes and methods can be found in Hervé et al. (2014).

Each NFI plot has detailed inventory information, including density (trees/ha), quadratic mean diameter (cm), basal area (m^2/ha), dominant height (m), and volume (m^3/ha). For this study we focused on two variables of high importance in forestry, i.e., forest dominant height when modelling forest canopy and forest volume when performing SAE. Table 1 provides descriptive statistics for both of these variables obtained from the NFI data at both overall AOI and individual department levels.

2.3. Auxiliary data

The auxiliary data consist of 1) GEDI Level 2A products; 2) the forest mask from the French National Institute of Geographic and Forest Information (IGN) (BDForêt® V2, https://geoservices.ign.fr/bdforet); and 3) a digital terrain model (DTM, BD ALTI® 25 m, https://geoservices.ign.fr/bdalti) over the study area.

GEDI Level 2A product provides footprint information of multiple layers, including beam types, sensitivity, geo-located elevation and height metrics. Both full-power and coverage footprints were used in this study. Height metrics include standard relative height (RH) percentiles from 0 to 100 (e.g., percentiles 95 and 100 are RH_{95} and RH_{100}). According to the literature, multiple GEDI RHs can be used as an assessment of forest height (Duncanson et al., 2022). We retrieved and downloaded the data from the NASA/USGS Land Processes Distributed Active Archive Centre (https://lpdaac.usgs.gov). The data acquisition dates were between 2019 and 04-22 and 2020–04-14, the latest data available at the time. The bounding box used to capture spatial GEDI footprints has the following WGS84 coordinates: 1.417434 (xmin), 2.792492 (xmax), 47.11044 (ymin), 48.1175 (ymax).

The BDForêt® V2 forest is a vector product derived from the interpretation of near-infrared aerial images and it provides information of vegetation composition (Vega et al., 2021). For instance, it captures forest stands of at least 0.5 ha in size. The forest mask was mainly used to filter out irrelevant GEDI footprints, i.e., whether or not GEDI footprints are located in vegetation. In this study, the vegetation classification was divided into three categories: broadleaved forest, coniferous forest, and mixed forest.

The digital terrain model (BD ALTI®) used in this study results from a Lidar campaign conducted in 2014. It was resampled and made available by IGN at a 25 m resolution. The forest mask and DTM can be downloaded from the IGN web platform (https://geoservices.ign.fr).

3. Methods

The flowchart below (Fig. 2) illustrates the overall framework of this study. In brief, we adopted the framework of a two-phase sampling scheme described in Hill et al. (2021). We first had to select the suitable GEDI RH metric that can represent the NFI dominant height at plot level. Based on a semi-variogram examination, we paired GEDI auxiliary information with NFI measurements at plot level up to a critical distance obtained from the semi-variogram. Finally, small area estimation of

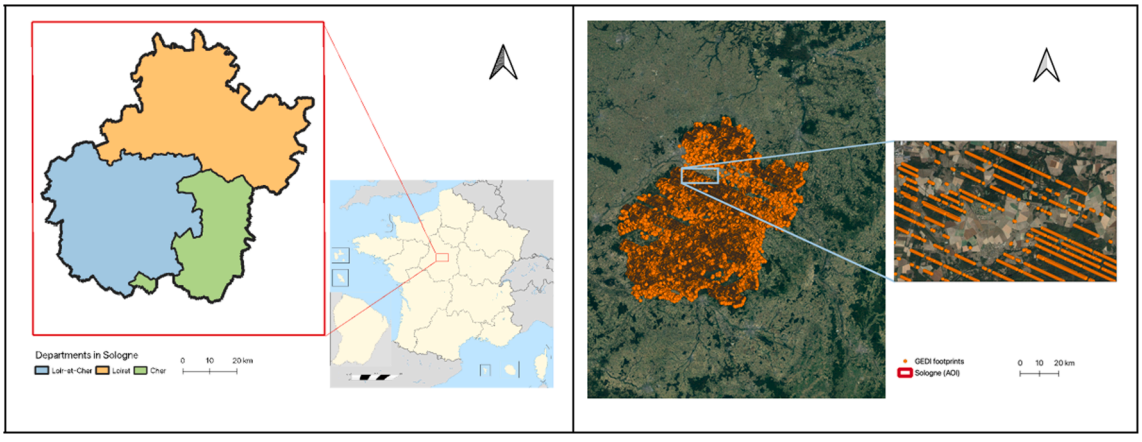


Fig. 1. A. Location of the study area (departments have different colours). B. Availability of filtered GEDI footprints over the AOI.

Table 1
Statistics of the forest volumes and dominant height from the 635 plots in Sologne, and its partitioning within the 3 administrative units (departments).

Area	Plot N	Stem Volume (m^3/ha)				Dominant height (m)			
		Min.	Mean	Max.	SD	Min.	Mean	Max.	SD
Overall AOI	635	1.45	176.5	926.4	128.5	4.9	18.8	38.2	5.2
Cher	115	1.80	172.8	462.6	108.6	7.7	18.9	30.8	5.0
Loiret	243	1.53	174.8	926.4	134.6	4.9	18.7	38.2	5.7
Loir-et-Cher	277	1.45	179.6	699.3	130.9	5.8	18.8	30.1	4.7

Note: dominant height (m) is defined as the average height of the 100 largest trees per hectare.

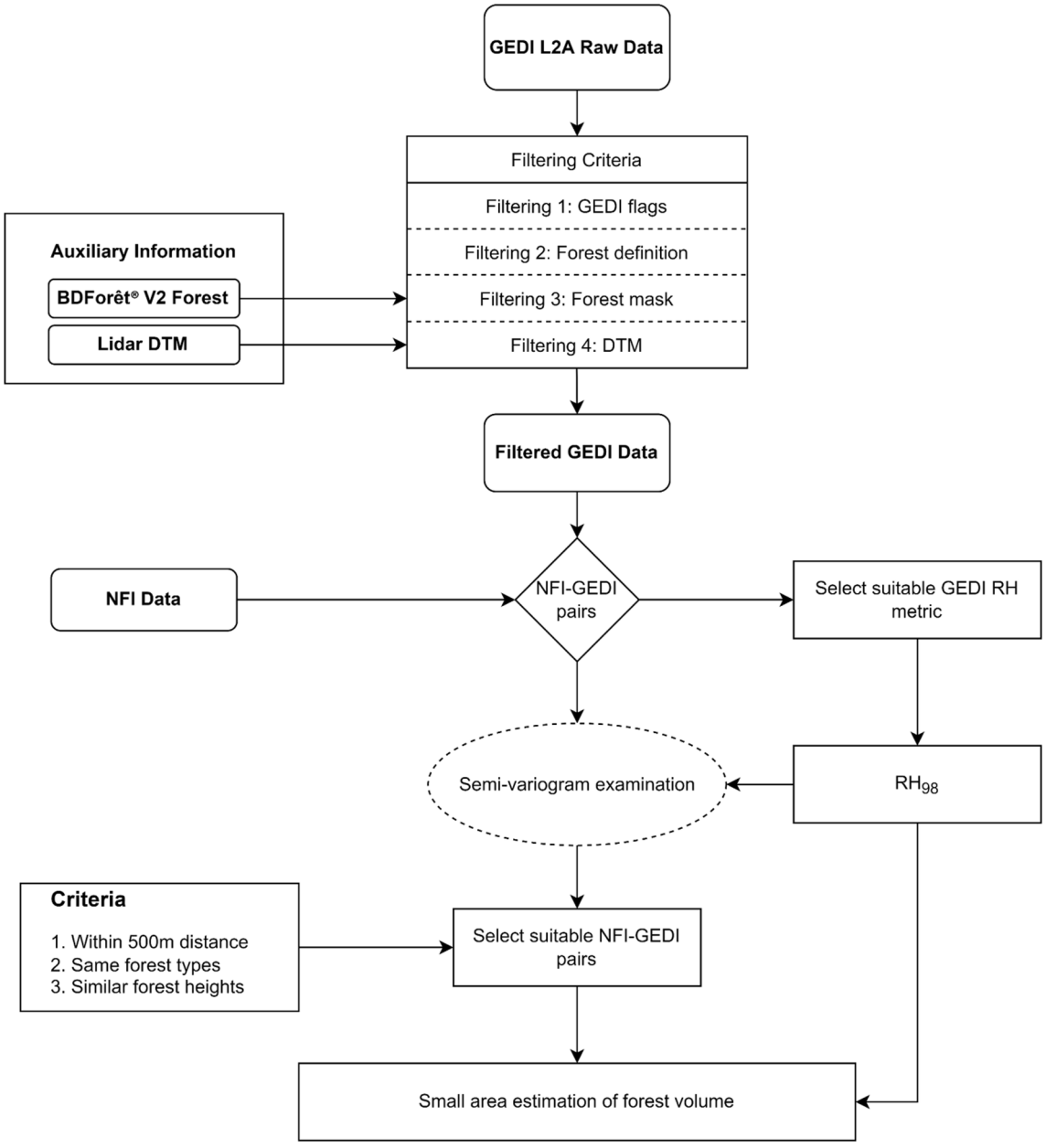


Fig. 2. Overall workflow of this study.

forest volume was performed with a simple linear regression using the R package "forestinventory".

3.1. GEDI footprints processing

Footprint processing aimed first at selecting valuable footprints and second at matching GEDI footprints to NFI plots with similar stand characteristics. Firstly, we only kept those footprints whose "quality flag" was 1 and "degrade flag" 0, which ensures the removal of error or low-quality footprints (Hofton & Blair, 2019). Next, we followed the UNFCCC definition, which states that "trees in a forest reach a minimum height of 2–5 m at maturity". Therefore, we used an average height of 3 m as a threshold and removed those footprints whose RH_{100} values were smaller than 3 m and thus could not be qualified as forest at the time of

the survey. In addition, the footprints were intersected with forest masks to ascertain their forest types (i.e., in which type of forest they are located, broadleaved, mixed or coniferous forests) and footprints located outside forest masks were excluded. Lastly, we removed footprints that have an elevation discrepancy of more than 50 m compared to Lidar DTM. As a result, a total number of 112,569 footprints were included for further analysis and the available footprints over AOI after filtering is illustrated in Fig. 1b. Height metrics and forest types were then extracted from the footprints, which formed the auxiliary data frame.

3.2. Harmonisation of the projection system and DTM correction

All the data but GEDI products are available in the French Lambert-93 projection system. Geographic coordinates of the GEDI footprints, in the WGS84 system, were transformed to map coordinates in the Lambert-93 projection system. During this transformation, particular attention must be paid to the vertical coordinate as GEDI elevations are provided taking an ellipsoid as reference when the values in the reference DTM are altitudes with the geoid as reference. It was then necessary to rectify the differences brought by ellipsoids before comparing elevation estimations.

In France, the IGN's Service of Geodesy and Metrology disseminates quasi-geoid grids of 1 km resolution (RAF 18) that are commonly used to convert ellipsoidal heights into altitudes. Its latest update was in 2020. According to the grids, a mean altitude difference of 48 m was detected and thus added to the DTM to ensure the accuracy of GEDI elevation.

3.3. Variogram examination of the GEDI values

To explore the spatial correlation of the GEDI footprints, the semi-variance of the selected GEDI RH was calculated following Eq. (1) and the corresponding empirical semi-variogram was established using the variogram function from the 'gstat' package (Pebesma, 2004). The selected RH was used as a surrogate of the field dominant height, with the aim of obtaining an idea of the spatial dependence of the canopy cover present in the AOI, which can help to determine the optimal distance threshold used to pair GEDI footprints with corresponding NFI plots.

$$Semivariance = \frac{1}{2n} \times \sum_{i \neq j}^{n} \left(RH_{x,i} - RH_{x,j}\right)^2 \quad (1)$$

where n is the number of GEDI footprints, $RH_{x,\,i}$ and $RH_{x,\,j}$ are the selected RH_x values of GEDI footprints recorded at 2 separate locations i and j.

Next, NFI plots were paired up with the closest GEDI footprints based on their Euclidean distances (NFI-GEDI pairs). The semi-variogram was used to identify the NFI-GEDI pairs whose distances were smaller than the distance at which the semi-variogram reached the horizontal asymptote (i.e., 500 m). We purposely considered that GEDI footprints and NFI plots are most likely not geographically co-located. Indeed, due to the innate geolocation errors of GEDI footprints, it is nearly impossible to obtain a perfect co-location, and a pairing step is therefore necessary. Thus, we deemed that footprints located within such distance can to some extent represent similar, if not identical, forest structures of their paired NFI plots. Pairs having a height difference (between the selected GEDI RH and NFI dominant height) of more than 10 m and located in different forest types were excluded from further analysis. We further divided the distance into three distance groups, 100, 300, 500 m, to assess their impacts on small area estimation of forest volume.

3.4. Small area estimation

Unit-level SAE was performed using the two-phase non-exhaustive estimation procedure provided in the R package "forestinventory" and described in Hill et al. (2021). The first phase is associated with the auxiliary GEDI information used to generate model predictions based on a linear regression using the method of ordinary least squares. The second phase contains the field NFI plot attributes (i.e., forest volume), that was used to generate model coefficients and correct bias. Using field data alone (all available plots within the AOI, not restricted to the paired plots), the mean and variance of the SRS were calculated as Eqs. (2) and (3):

$$\widehat{\mu}_{SRS} = \frac{1}{n}\sum\nolimits_{i=1}^{n} y_i \quad (2)$$

$$\widehat{Var}\left(\widehat{\mu}_{SRS,G}\right) = \frac{1}{n \times (n-1)}\sum\nolimits_{i=1}^{n} \left(y_i - \widehat{\mu}_{SRS}\right)^2 \quad (3)$$

where n is the total number of NFI sample plots in the AOI in this case, and y_i is the observed value of forest volume of plot *i*.

For small domain estimation, the procedure relied on the small area estimators provided in Hill et al. (2021, Equations 4b and 5b) for non-exhaustive availability of auxiliary information. The prediction model used here is internal, where the estimate accounted for fitting the inventory data at hand. The mean (Eq. (4)) and g-variance (Eq. (5)) estimators are defined as follow:

$$\widehat{Y}_{G,psmall} = \widehat{\bar{Z}}_G^{\,t}\widehat{\beta}_{s2} + \frac{1}{n_{2G}}\widehat{R}(x) \quad (4)$$

$$\widehat{Var}\left(\widehat{Y}_{G,psmall}\right) = \widehat{\bar{Z}}_G^{\,t}\widehat{\Sigma}_{\widehat{\beta}_{s2}}\widehat{\bar{Z}_G} + \widehat{\beta}_{s2}^{\,t}\widehat{\Sigma}_{\widehat{\bar{Z}_G}}\widehat{\beta}_{s2} + \frac{1}{n_{2G}}\widehat{Var}_{s2G}(\widehat{R}(x)) \quad (5)$$

where the predictions over the small area G (Eq. (4), first term) are corrected by the mean bias of the model (Eq. (4), second term) and for the variance of the predictions (5) the first term is related to the variance–covariance matrix $\widehat{\Sigma}_{\widehat{\beta}_{s2}}$ of the regression coefficient ($\widehat{\beta}_{s2}$, estimated on s2), the second term to the variance–covariance matrix $\widehat{\Sigma}_{\widehat{\bar{Z}_G}}$ of the auxiliary vector $\widehat{\bar{Z}}_G$, and the last term to the residual correction term ($\widehat{R}(x)$).

We do acknowledge that the co-location of units of s1 and s2 is not fulfilled here, since s2 units are not strictly speaking a subsample of s1 but rather assimilated. We however conjecture that this imperfect match of units, unavoidable given GEDI footprint's location errors, does not compromise the approach since it solely translates into larger model errors.

The performance of SAE using the GEDI auxiliary data was evaluated by measuring relative efficiency (RE), as shown in Eq. (6). RE estimates the gain in precision brought by using auxiliary information that is incorporated into the model estimators in comparison to using field estimates alone. In the latter case, all field plots were used as references (not only the paired data). RE values higher than 1 indicate increased precision.

$$RE = \widehat{Var}\left(\widehat{\mu}_{SRS,G}\right) / \widehat{Var}\left(\widehat{Y}_{G,psmall}\right) \quad (6)$$

4. Results

4.1. Select suitable GEDI RH metric

We had to decide which GEDI RH metric is comparable with NFI dominant height and thus can be used in modelling forest volume. We first inspected the distributions of the three GEDI RH metrics selected at the upper limit (RH_{100}, RH_{98} and RH_{95}) with that of the NFI dominant height over the AOI. At the AOI level, the mean of NFI dominant height (18.8 m) was found to be closer to the mean of GEDI RH_{98} (18.4 m) than those of RH_{100} (20.0 m) and RH_{95} (17.1 m). In addition, Fig. 3 illustrates that the distribution of NFI dominant height in 5 m intervals overall resembles all three GEDI height metrics. The distribution of RH_{98} appears closer to that of NFI dominant height, while those of RH_{100} and RH_{95} tend to over- and underestimate it respectively. Note that a small

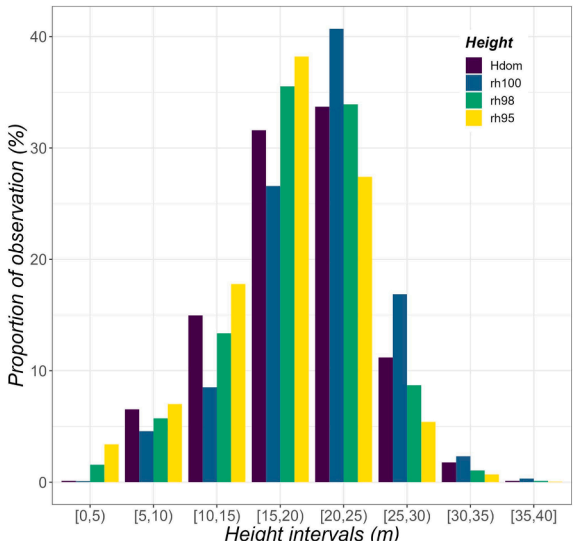


Fig. 3. Comparing NFI dominant height with GEDI RH_{100}, RH_{98} and RH_{95} in 5 m intervals.

peak of GEDI height metrics commonly exists on the left side of the distributions.

In addition, we compared the correlations between NFI dominant height with the three selected GEDI RH metrics. Statistics at the whole AOI level showed that the NFI dominant height is strongly correlated with GEDI height metrics, with RH_{100}, RH_{98} and RH_{95} having correlations of 0.90, 0.90 and 0.89 respectively. Fig. 4 illustrates the field measured NFI dominant heights and their counterparts predicted by GEDI RH_{100}, RH_{98} and RH_{95} using all available 635 plots. The RMSE obtained by RH_{100}, RH_{98} and RH_{95} were 2.3 m, 2.3 m and 2.5 m respectively. As a result, we used RH98 as a proxy of NFI forest dominant height. It was later used as an input in the semi-variance analysis and forest volume estimation.

4.2. *Select suitable NFI plots based on the semi-variogram result*

Using the selected GEDI metric RH_{98}, we calculated the semi-variance using all filtered GEDI footprints. Fig. 5 illustrates that the values of GEDI RH_{98} between two footprints were to some extent spatially correlated depending on the distance between them. The semi-variogram reaches a horizontal asymptote at approximately 500 m, and the process could then be considered as second-order stationary, which suggests that the spatial correlation disappears when two footprints are located farther than approximately 500 m away from each other. Based on this, every NFI plot was paired with the nearest GEDI footprint within

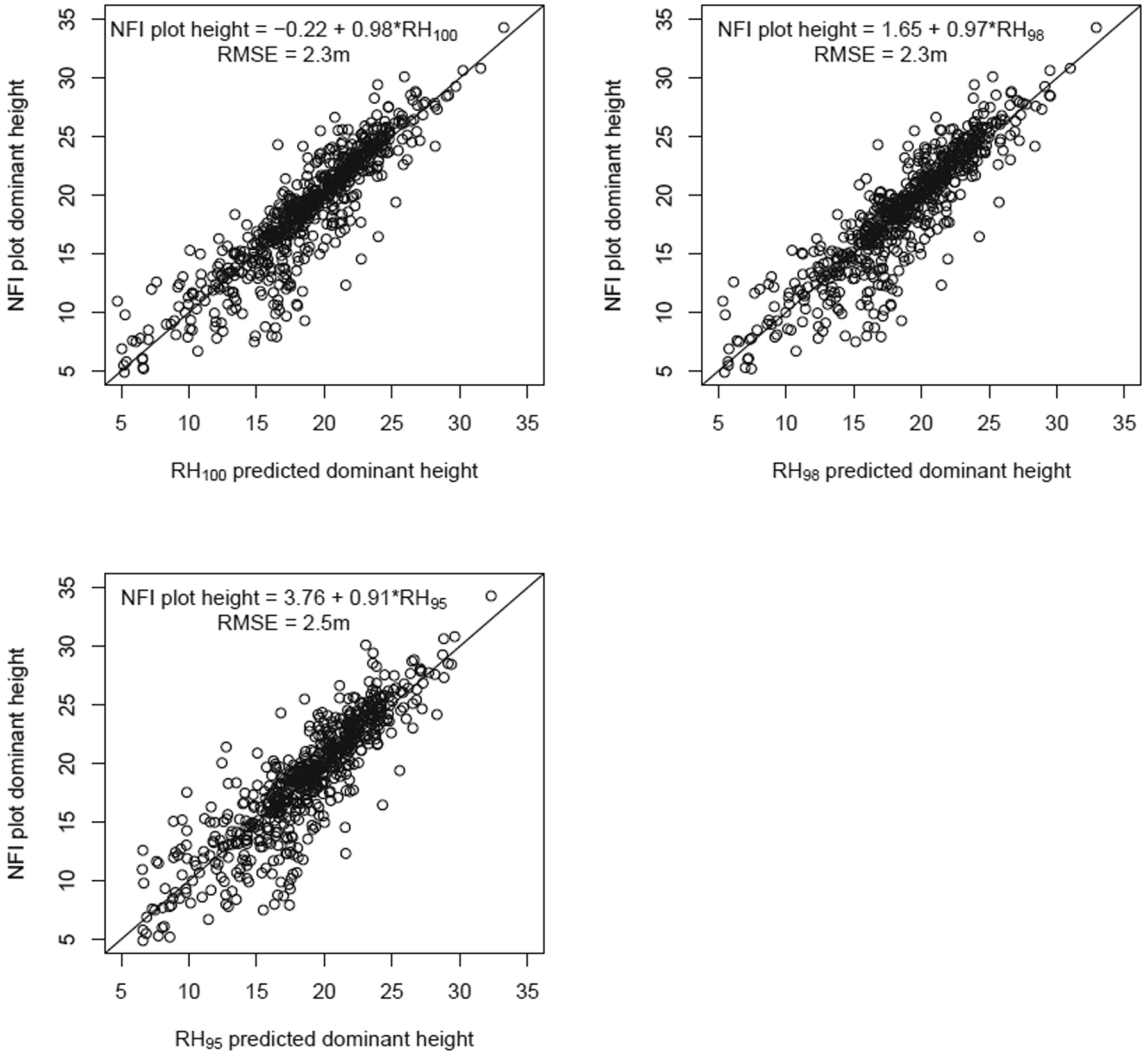


Fig. 4. Predicted forest dominant height by GEDI RH_{100}, RH_{98} and RH_{95} vs observed forest dominant height from NFI using all 635 available plots.

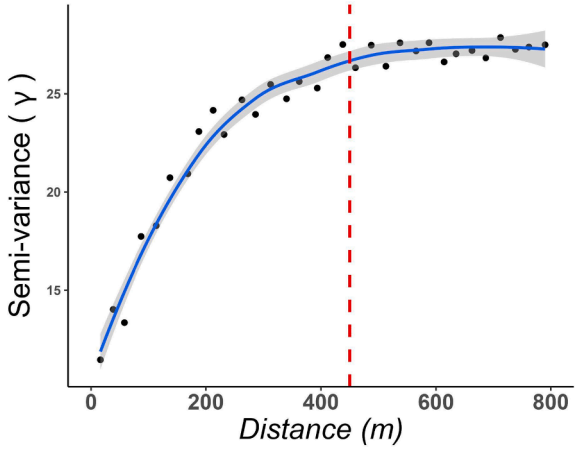


Fig. 5. Semi-variance (ϒ) of GEDI height (RH_{98}) for Sologne (AOI) according to the distance between footprints (in metres). (A vertical line is placed at 500 m).

the range of 500 m that have identical forest types (NFI forest type = GEDI forest type) and similar forest heights (|NFI dominant height – GEDI RH_{98}| < 10 m).

Next, we assessed the statistics of NFI-GEDI pairs in each distance group (Table 2). Originally, there were 635 NFI plots available across our study area and only 366 plots were retained after the pairing process. The number of NFI-GEDI pairs was further reduced to 331 and 141 when the distance threshold was set to 300 m and 100 m respectively. It was expected that the number of available pairs would increase with increasing distance threshold. The mean height difference varied from 1.3 to 1.5 m depending on the distance groups.

4.3. Small area estimation of forest volume

With the NFI-GEDI pairs and filtered GEDI RH_{98}, small area estimation of forest volume was performed at both AOI and sub-area levels using a simple linear internal model (i.e., built from the paired plots of the small area). Results showed that GEDI auxiliary information slightly improved the estimation accuracy for the AOI level at the maximum distance threshold used for pairing (i.e., 500 m). A similar increase in relative efficiency was also witnessed in every sub-area with an increasing number of NFI-GEDI pairs used to build the models (Table 3).

At department level, when limiting the spatial distance between paired NFI plots and GEDI footprints to 100 m, there was no gain in relative efficiency observed in all three departments. This meant that GEDI auxiliary data were not able to reduce model variance under this condition as compared to the standard NFI sample. With increasing distance, the sample size of NFI-GEDI pairs used to calibrate SAE models in the second phase also enlarges. When the spatial distance was restricted at 300 m, relative efficiency fluctuated from 0.9 in Cher to 1.3 in Loiret. A gain in relative efficiency by a factor of 1.1 to 1.4 was observed in every department when using all matched NFI-GEDI pairs located within 500 m. In the department of Cher, the mean estimate of volume was slightly overestimated compared to the one of SRS (although within the confidence interval ± 2 sd), while in other departments and at the whole AOI level, the mean volume estimates remained relatively stable (Fig. 6).

Table 3
Volume estimations of SAE and SRS at both AOI and sub-area (departments) levels.

Distance	Area	NFI-GEDI pairs	SAE estimates	SRS estimates	Relative efficiency
100 *m*	AOI	141	176.5 (7.7)	176.5 (5.1)	0.4
	Cher	26	183.5 (16.6)	172.8 (10.1)	0.4
	Loir-et-Cher	51	169.3 (12.0)	179.6 (7.9)	0.4
	Loiret	64	179.5 (12.6)	174.8 (8.6)	0.5
300 *m*	AOI	331	179.2 (5.1)	176.5 (5.1)	1.0
	Cher	59	202.0 (10.4)	172.8 (10.1)	0.9
	Loir-et-Cher	124	177.1 (7.8)	179.6 (7.9)	1.0
	Loiret	148	171.6 (7.6)	174.8 (8.6)	1.3
500 *m*	AOI	366	180.1 (4.8)	176.5 (5.1)	1.1
	Cher	66	199.2 (9.8)	172.8 (10.1)	1.1
	Loir-et-Cher	137	178.8 (7.5)	179.6 (7.9)	1.1
	Loiret	163	173.4 (7.2)	174.8 (8.6)	1.4

Note: SAE and SRS estimates have mean volume predictions and standard deviation in parenthesis.

5. Discussion

5.1. Selecting suitable GEDI RH metric

Our results suggest that GEDI RH_{98} can be used to represent NFI dominant height and consequently be used to model forest volume. Although Fig. 3 illustrated that GEDI RH_{98} and NFI dominant height had certain inconsistency in terms of their distributions, RH_{98} was found overall as a good predictor of forest volumes. The 'bumps' commonly existing at the lower end of the GEDI distributions are most likely bad footprints resulting from noises. Given that the GEDI system pulse (FWHM: full width at half-minimum) is 15.6 ns (Hancock et al., 2019), the ground signal starts at approximately 13 ns above the ground peak for a flat terrain, which is equivalent to a precision of approximately 2 m. This means that a single ground pulse is likely to be found at 2 m above the ground and that a 5 m forest threshold could have been used to avoid slopes being misclassified as trees. This explains the seemingly high values at the lower end of the distributions.

In addition, estimating NFI dominant height with RH_{98} yielded a low RMSE of 2.3 m at forest plot level. In fact, choosing the suitable GEDI RH metric to represent NFI dominant height is difficult since this attribute may be affected by fragmented forest structure where canopy height can change drastically within a close distance. RH_{98} records the height at which 98 percentiles returned energy relative to the footprint centre. Higher percentiles, such as RH_{100}, have been found sensitive to noise and thus less precise (Silva et al., 2018). Previous research has also

Table 2
The number of retained NFI plots under each distance class. Distance statistics and absolute height difference between field dominant height and GEDI RH_{98} also presented.

Distance conditions (m)	NFI-GEDI pairs	Statistics of the distances (m)				|NFI dominant height – GEDI RH_{98}| (m)			
		Min.	Median	Mean	Max.	Min.	Mean	Max.	SD
100	141	6.2	50.0	52.2	100.0	~0	1.3	8.1	1.5
300	331	6.2	112.0	125.9	299.0	~0	1.5	9.8	1.7
500	366	6.2	124.4	150.0	496.6	~0	1.5	9.8	1.7

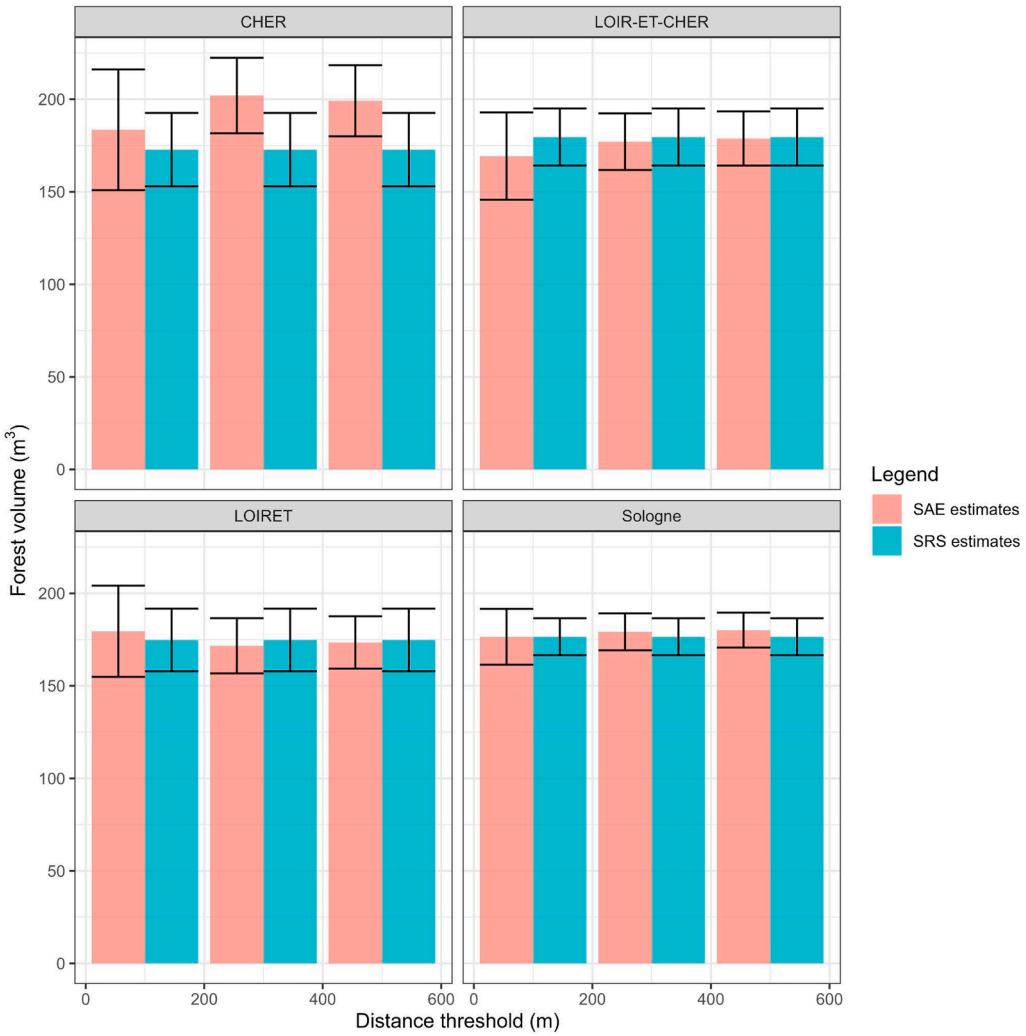


Fig. 6. Estimated mean forest volume obtained by SRS and SAE methods at AOI and sub-are level, with error bar indicating mean ± 2*sd.

demonstrated that RH_{98} is a more stable predictor used in modelling many forest attributes, including forest fuel load (Leite et al., 2022) and above-ground biomass (Duncanson et al., 2022). The fact that we included both full-power and coverage footprints with a relatively low sensitivity threshold (greater than0.9) made RH_{98} a much safer choice, which was later used as an input in the semi-variance examination and modelling forest volume.

5.2. *Pairing issues and estimation uncertainties*

We drew inspiration from the two-phase sampling scheme proposed by (Mandallaz et al., 2013), where partially-exhaustive auxiliary data were used in a model-assisted estimation process. As we used non-exhaustive auxiliary data in this study, the details of the final estimators as stated in the method section could be found in Mandallaz et al. (2016) and Hill et al. (2021). Since the NFI plots followed a similar design-based sampling scheme, it was tempting to use this approach to evaluate the benefits that GEDI could bring in terms of estimation efficiency. However, due to the geolocation uncertainties associated with GEDI (Dubayah et al., 2020), pairing field plots is difficult. This uncertainty compromises the co-location, and consequently, the NFI plots are not a subsample of the first phase sample. This however only exacerbates an issue present whenever using remote-sensed data along with ground plot, whereby a perfect location match is probably more hypothetical than real.

To overcome these limits imposed by GEDI itself, we tested a pairing method based on a neighbourhood consistency and a threshold distance between field plots and GEDI footprints. Our results showed that the uncertainty introduced by the pairing, when the pairing consistency is controlled, did not compromise the efficiency of SAEs.

As the whole AOI and sub-areas have relatively large spatial extents, an adequate amount of calibration samples within the domain appeared to be necessary to achieve relative efficiencies greater than 1 (Table 3). This is particularly obvious for the sub-area Loiret, which yields the highest RE (i.e., 1.4) at the maximum neighbouring distance tested. Globally, a sample size of at least 100 paired plots appeared to be required in our study to reduce the model's variability (RMSE), enabling a gain in RE compared to the SRS of the whole NFI plots. This is therefore the main limitation of the present approach, where the precision gain appears to depend on a sufficient amount of model's calibration plots. Note that we only used GEDI data collected over the time span of approximately-one year (2019–04-22 and 2020–04-14) at the time when this study was undergoing. We expect that both the number of quality GEDI footprints and their spatial distribution over our study area will increase at least twofold by the end of the two-year GEDI mission. In other words, the number of both NFI-GEDI pairs used for model calibration and footprints as auxiliary data will increase, which is likely to further improve volume estimation accuracy.

6. Conclusion

This study showcased how to match NFI measurements with GEDI

footprints at plot level and subsequently perform unit-level small area estimation using GEDI Level 2A products as auxiliary data. We emphasised the importance of having the same forest types and similar forest heights when pairing NFI plots with nearby GEDI footprints. In our case, we found that GEDI RH_{98} was a suitable candidate to represent NFI dominant height at plot level. The fact that in Sologne, GEDI RH_{98} appeared to become spatially independent after approximately 500 m provided valuable insights into the NFI sampling design in order to achieve optimal efficiency. With these NFI-GEDI pairs and filtered GEDI data, the results showed that GEDI auxiliary information can improve forest volume estimation compared to SRS. This is particularly affected by the number and similarity of NFI-GEDI pairs used to calibrate models. The fact that GEDI data are open-access and cover the entire country of France makes it particularly attractive to improve forest inventory precision at regional and local levels.

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.

Data availability

Data will be made available on request.

Acknowledgements

We are grateful to the two anonymous reviewers for their valuable comments on improving the quality of this manuscript. This research received the financial support of the TOSCA Continental Surface program of the Centre National d'Etudes Spatiales (CNES) through the project SLIM "Space Lidar for Improved Multisource Forest Inventory" (grant number 4800001129). IGN and the ONF Département RDI are supported by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (ANR-11-LABX-0002-01, Lab of Excellence ARBRE). SZ also received financial support from the Academy of Finland Flagship Programme (Forest-Human-Machine Interplay - Building Resilience, Redefining Value Networks and Enabling Meaningful Experiences (UNITE); grant number 337127). OB acknowledged funding from project Pro-USV-Biom, contract no. 10PFE/2021 financed by the Ministry of Research, Innovation and Digitalization within Program 1 - Development of national research and development system.

References

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. https://doi.org/10.1016/j.rse.2018.04.028.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. Science of Remote Sensing 1, 100002. https://doi.org/10.1016/j.srs.2020.100002.

Duncanson, L., Kellner, J. R., Armston, J., Dubayah, R., Minor, D. M., Hancock, S., Healey, S. P., Patterson, P. L., Saarela, S., Marselis, S., Silva, C. E., Bruening, J., Goetz, S. J., Tang, H., Hofton, M., Blair, B., Luthcke, S., Fatoyinbo, L., Abernethy, K., et al. 2022. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. Remote Sensing of Environment, 270, 112845. https://doi.org/10.1016/j.rse.2021.112845.

Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L.I., Kellner, J.R., Dubayah, R., 2019. The GEDI Simulator: A Large-Footprint Waveform Lidar Simulator for Calibration and Validation of Spaceborne Missions. Earth Space Sci. 6 (2), 294–310. https://doi.org/10.1029/2018EA000506.

Hervé, J.-C., Wurpillot, S., Vidal, C., & Roman-amat, B. 2014. L'inventaire des ressources forestières en France: un nouveau regard sur de nouvelles forêts. Revue Forestière Française, 3, Fr.], ISSN 0035. https://doi.org/10.4267/2042/56055.

Hill, A., Massey, A., Mandallaz, D., 2021. The R Package forestinventory: Design-Based Global and Small Area Estimations for Multiphase Forest Inventories. J. Stat. Softw. 97 (4). https://doi.org/10.18637/jss.v097.i04.

Hofton, M., & Blair, B. 2019. Algorithm Theoretical Basis Document (ATBD). Report. Goddard Space Flight Center, Maryland. https://lpdaac.usgs.gov/documents/581/GEDI_WF_ATBD_v1.0.pdf.

Leite, R. V., Silva, C. A., Broadbent, E. N., Amaral, C. H. do, Liesenberg, V., Almeida, D. R. A. de, Mohan, M., Godinho, S., Cardil, A., Hamamura, C., Faria, B. L. de, Brancalion, P. H. S., Hirsch, A., Marcatti, G. E., Dalla Corte, A. P., Zambrano, A. M. A., Costa, M. B. T. da, Matricardi, E. A. T., Silva, A. L. da, et al. 2022. Large scale multi-layer fuel load characterization in tropical savanna using GEDI spaceborne lidar data. Remote Sensing of Environment, 268, 112764. https://doi.org/10.1016/j.rse.2021.112764.

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. https://doi.org/10.1139/cjfr-2013-0448.

Mandallaz, D., Breschan, J., Hill, A., 2013. New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. Can. J. For. Res. 43 (11), 1023–1031. https://doi.org/10.1139/cjfr-2013-0181.

Mandallaz, D., Hill, A., & Massey, A. F. 2016. Design-based properties of some small-area estimators in forest inventory with two-phase sampling. Report. ETH Zurich. https://doi.org/10.3929/ethz-a-010579388.

McRoberts, R.E., Tomppo, E.O., Finley, A.O., Heikkinen, J., 2007. Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. Remote Sens. Environ. 111 (4), 466–480. https://doi.org/10.1016/j.rse.2007.04.002.

Molina, I., & Marhuenda, Y. 2015. sae: An R Package for Small Area Estimation. The R Journal, 7(1), 81. https://doi.org/10.32614/RJ-2015-007.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.

Roy, D.P., Kashongwe, H.B., Armston, J., 2021. The impact of geolocation uncertainty on GEDI tropical forest canopy height estimation and change monitoring. Science of Remote Sensing 4, 100024. https://doi.org/10.1016/j.srs.2021.100024.

Silva, C.A., Saatchi, S., Garcia, M., Labriere, N., Klauberg, C., Ferraz, A., Meyer, V., Jeffery, K.J., Abernethy, K., White, L., Zhao, K., Lewis, S.L., Hudak, A.T., 2018. Comparison of Small- and Large-Footprint Lidar Characterization of Tropical Forest Aboveground Structure and Biomass: A Case Study From Central Gabon. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (10), 3512–3526. https://doi.org/10.1109/JSTARS.2018.2816962.

Tomppo, E. 2006. The Finnish Multi-source National Forest Inventory - Small Area Estimationand Map Production. In A. Kangas & M. (Eds.) Maltamo (Eds.), Forest Inventory (Vol. 10, pp. 195–224). Springer. https://doi.org/10.1007/1-4020-4381-3_12.

Vega, C., Renaud, J.-P., Sagar, A., Bouriaud, O., 2021. A new small area estimation algorithm to balance between statistical precision and scale. Int. J. Appl. Earth Obs. Geoinf. 97, 102303 https://doi.org/10.1016/j.jag.2021.102303.