

Sentinel-1 and Sentinel-2 data for predicting forest stand variables in GEE: A case study of timber plantations in northern Spain

Iyán Tejjido-Murias¹, Marcos Barrio-Anta¹ and Carlos A. López-Sánchez^{1*}

¹*SMartForest Research Group, AsRaM Asturias Raw Material Institute, 33600 University of Oviedo. Mieres, Spain.*

* Correspondence should be addressed to Carlos A. López-Sánchez: lopezscarlos@uniovi.es

Abstract

Aim of study: In this study we developed machine learning models for estimating five forest stand variables, including total over bark volume, number of trees per hectare, dominant height, basal area and aboveground biomass, from remote sensing data and ancillary variables in forest plantations of *Eucalyptus globulus* Labill., *Pinus pinaster* Aiton and *Pinus radiata* D. Don.

Area of study: The study was conducted in northern Spain (Autonomous Communities of Asturias, Cantabria, Basque Country and Galicia) where plantations of the species of interest are mainly concentrated due to the high productivity of forest stands in these areas.

Material and methods: We used ground-truth data from Spanish National Forest Inventory plots, two sources of remote sensing data and images from the Sentinel-1 constellation (radar) and from the Sentinel-2 constellation (optical) combined with terrain and climate data.

We used the Google Earth Engine platform using Random Forest algorithm with 10-fold cross-validation to obtain forest stand variables, including density, size and yield variables.

Main results: Evaluation of the model accuracy and variable importance in estimating forest variables showed that the total volume models performed best, yielding R^2 values between 0.39 and 0.45 for the different tree species. Optical bands were very important in all cases, while radar bands were less important.

Research highlights: The use of synthetic aperture radar, although promising, is limited in platforms like GEE, especially for complex terrains where shadowing and angles affect the image quality. The research contributions include the automation of preprocessing steps, and the findings highlight the need to develop more robust, adaptable models using AI and new radar sensors.

Keywords: aboveground biomass; Google Earth Engine; machine learning; modelling; total volume.

Datos de Sentinel-1 y Sentinel-2 para predecir variables de rodal en GEE: Un caso de estudio en las plantaciones forestales del norte de España

Resumen

Objetivo del estudio: En este trabajo se desarrollaron modelos de aprendizaje automático para estimar cinco variables de rodal, incluyendo volumen total con corteza, número de árboles por hectárea, altura dominante, área basimétrica y biomasa forestal aérea a partir de datos de teledetección y variables auxiliares en plantaciones forestales de *Eucalyptus globulus* Labill., *Pinus pinaster* Aiton y *Pinus radiata* D. Don.

Área de estudio: El estudio se llevó a cabo en el norte de España (Comunidades Autónomas de Asturias, Cantabria, País Vasco y Galicia), donde, debido a su alta productividad forestal, se concentran las principales plantaciones de las especies de interés.

Materiales y métodos: Se utilizaron datos de campo provenientes de parcelas del Inventario Forestal Nacional de España, junto con dos fuentes de datos de teledetección: imágenes de la constelación Sentinel-1 (radar) y de la constelación Sentinel-2 (óptico), combinadas con datos de terreno y clima. Se empleó la plataforma Google Earth Engine (GEE) utilizando el algoritmo Random Forest con validación cruzada de 10 carpetas para obtener variables de rodal.

Resultados principales: La evaluación de la precisión de los modelos y la importancia de las variables mostró que los modelos de volumen total fueron los que mejor estadísticos presentaron, con valores de R^2 entre 0,39 y 0,45 según la especie. Las bandas ópticas fueron muy importantes en todos los casos, mientras que las bandas de radar tuvieron menor relevancia.

Aspectos destacados de la investigación: El uso de radar de apertura sintética (SAR), aunque prometedor, está limitado en plataformas como GEE, especialmente en terrenos complejos donde el sombreado y los ángulos afectan la calidad de la imagen. Las contribuciones del estudio incluyen la automatización de los pasos de preprocesamiento y destacan la necesidad de desarrollar modelos más robustos utilizando inteligencia artificial y nuevos sensores radar.

Palabras clave: biomasa forestal aérea; Google Earth Engine; machine learning; modelización; volumen total.

Citation: Tejjido-Murias, I; Barrio-Anta, M; López-Sánchez, CA (2025). Sentinel-1 and Sentinel-2 data for predicting forest stand variables in GEE: A case study of timber plantations in northern Spain. *Forest Systems*, Volume 34, Issue 3, 20986. <https://doi.org/10.5424/fs/2025343-20986>

Received: 04/12/2024. **Accepted:** 14/05/2025. **Published:** 27/11/2025.

Copyright © 2025 CSIC. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License.

This article belongs to the Special Issue “Advances in national forest monitoring networks: integrating data for ecosystem assessment and management”: <https://fs.revistas.csic.es/index.php/fs/announcement/view/29>

Introduction

The area encompassing the autonomous communities of Galicia, Asturias, Cantabria and Basque Country is the most productive forest region in Spain (Aguirre et al., 2022). Forest plantations of *Eucalyptus globulus* Labill, *Pinus pinaster* Aiton and *Pinus radiata* D. Don occupy 34% of the wooded surface area while generating 88% of the wood harvested in the region, which accounted for 62% of the total produced in Spain in the period 2005-2021 (MITECO, 2023).

Owing to the importance of forestry, reliable sources of information about forest resources are very important for the different stakeholders involved, such as forest officers, forest owners and timber companies. In the past, the only way of obtaining this information was by traditional forest inventory. Recent research has demonstrated that remote sensing can provide high spatial resolution images and spatially continuous data on forest variables. The ability to obtain up-to-date, accurate information about forest ecosystems at a low cost and with high temporal resolution has led to a new way of managing forest resources.

The most accurate spatially continuous predictions for large areas of different forest stand variables are obtained using Light Detection and Ranging (LiDAR) data. However, the data provided by national governments, such as data from the National Aerial Photography LiDAR Plan (PNOA-LiDAR), are characterized by low temporal resolution and high acquisition costs (Novo-Fernández et al., 2019). In the study area, the time between two LiDAR flights is often more than 6 years, which is unsuitable for short-rotation species, such as *E. globulus*. There are currently no high-resolution, temporally and spatially continuous LiDAR data available that would enable estimation of these forest parameters.

Given these limitations, and the need for forest stakeholders to have up-to-date information, we propose using satellite-derived variables, as the high temporal resolution of the data will allow this information gap to be filled. Although Novo-Fernández et al. (2024) recently reported advances in this field, they focused on the use of Sentinel-2 constellation (optical) in the Copernicus program. Using other satellite constellations, such as Sentinel-1 (radar), could provide valuable additional information about forests.

The Sentinel-1 (S1) satellite constellation consists of the Sentinel 1A and 1B satellites (launched into orbit in 2014 and 2016 respectively), which are equipped with synthetic aperture radar (SAR) systems. These

satellites operate at a radar frequency in the C-band (5.405 GHz). They provide global revisit coverage every 12 days with one satellite and every 6 days with both satellites (Sentinel-1B is now out of service) with a spatial resolution in Interferometric Wide Swath (IW) of 5 x 20 m. The system supports both single and dual polarization (HH or VV and HH+HV or VV+VH respectively).

In 2015, the European Space Agency (ESA) initiated the Sentinel-2 (S2) mission by launching Sentinel-2A (S2A), which was later complemented by the launch of Sentinel-2B (S2B), in 2017, and more recently Sentinel-2C (S2C), in 2024. Sentinel-2 satellites use 13 spectral bands across the electromagnetic spectrum. The spatial resolution varies among bands, ranging from 10 to 60 m per pixel, and the temporal resolution is of 5 days for the study area.

Despite the advantages in terms of temporal resolution, some well-known limitations of these sensors must be also mentioned. The presence of clouds greatly hampers the generation of optical images, while the speckle effect is a limiting factor in obtaining radar images. Additionally, in monitoring mountainous areas both types of imaging are hampered by difficulties due to slopes and the shadows they generate (Tejjido-Murias et al., 2024).

The Spanish National Forest Inventory (SNFI) provides up-to-date forest information every 5 years for the most productive forest species in northern Spain and it is a valuable source of data to determine the ground truth for modelling with remote sensing data (MAPAMA, 2019). Thanks to the availability of free satellite data and of SNFI data, models can be constructed to estimate the different forest variables. In addition, cloud computing tools, such as Google Earth Engine (GEE), facilitate access to remote sensing data, accelerate data processing and complement standard modelling methods.

The combined use of optical satellites and field databases enabled the development of predictive models in the previously mentioned recent study (Novo-Fernández et al., 2024). However, this previous study did not include radar data, which could be used to gain additional insights into the structure of the forest canopy. The present study aimed to improve two key aspects of the products developed by Novo-Fernández et al. (2024): i) including variables derived from radar imagery, which represents active remote sensing data, and ii) implementing the models in a cloud computing platform, such as GEE. This will enable integration of data from new sources and will also improve the efficiency and accessibility of data processing required for the development of these predictive models.

The overall aim of this research was to construct models to estimate aboveground biomass (*AGB*), basal area (*G*), dominant height (H_0), number of trees per hectare (*N*) and total over bark volume (*TV*) and for *E. globulus*, *P. pinaster* and *P. radiata* in northern Spain using S1, S2 data available in the GEE platform and other ancillary variables that we uploaded in GEE.

Materials and methods

Study area and data collection

The study area covers a total area of 52,821 km². The area is characterized by a temperate climate and complex orography due to the high variability in slope and elevation (ranging from sea level to 2,650 m.a.s.l.).

The development of forest models utilized four different datasets: i) field measurements, ii) remote sensing data, iii) terrain information and iv) climatic variables.

The ground truth data used in the study were obtained from the SNFI for commercially productive species in northern Spain (SNFI-4.5) conducted in mid-2018 (MAPAMA, 2019) by the Ministry for Ecological Transition and the Demographic Challenge, Spain. Sample plots of the three target species were established in pure stands, where at least 80% of the total basal area within each plot corresponded to the species of interest. Following this criterion, a total of 1,471 plots within the study area were available for analysis. Among these plots, 589 were dominated by *E. globulus*, 474 by *P. pinaster* and 408 by *P. radiata* (Table 1).

We used Sentinel-1 and Sentinel-2 images, which are available in the GEE repository (Sentinel-1 SAR GRD: C-band Synthetic Aperture Radar Ground Range Detected, COPERNICUS/S1_GRD and Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level 1C, COPERNICUS/S2_HARMONIZED).

The ancillary data included terrain variables obtained from a Digital Elevation Model (DEM) with a spatial resolution of 5 m, provided by the Spanish National Center for Geographic Information (CNIG), along with five climatic variables sourced from the Digital Climatic Atlas of the Iberian Peninsula, featuring a spatial

resolution of 200 m (Ninyerola et al., 2005). All variables in the dataset were resampled to a 10 m·pixel⁻¹ resolution using the cubic convolution method for terrain variables and the nearest-neighbour method for climatic variables. The names and abbreviations of the optical variables used in this study are provided in Supplementary Table S1. The equations and parameters used for the calculation of the indices and other variables can be consulted in Novo-Fernández et al. (2024). The new radar variables incorporated into this study, obtained through the GEE platform, are listed in Table 1.

Table 1. Radar characteristics considered candidate independent variables for predicting forest stand variables.

Group	Variable name
Radar bands	Vertical-Vertical polarization (<i>VV</i>), Vertical-Horizontal polarization (<i>VH</i>)
Radar index	Vegetation Radar Index (<i>RVVv</i>), Vertical Dual Depolarization Index (<i>VDDPI</i>), <i>VVVH</i> difference, sum, product and ratio (<i>VVVH_div</i> , <i>VVVH_sum</i> , <i>VVVH_prod</i> and <i>VVVH_ratio</i>)

Data preprocessing

Ground truth data from SNFI-4.5 plots were processed to produce the following forest state variables: two density variables (number of trees per hectare and basal area), one size-related variable (dominant height), and two yield variables (total over-bark volume and total aboveground biomass). The summary statistics of the forest state variables considered for the three forest species in the study area are included in Table 2.

Table 2. Descriptive statistics of the dependent forest variables analyzed in the study for each tree species. The statistics include the mean, minimum (Min.), maximum (Max.), and standard deviation (SD) of five structural attributes: number of trees per hectare (*N*), basal area (*G*), dominant height (*H₀*), total volume (*TV*), and aboveground biomass (*AGB*). The statistics are presented separately for each species: *Eucalyptus globulus*, *Pinus pinaster*, and *Pinus radiata*.

Species	N° plots	Forest variable	Descriptive statistic			
			Mean	Min.	Max.	SD.
<i>E. globulus</i>	589	<i>N</i> (trees ha ⁻¹)	833.83	10.19	2,695.02	499.93
		<i>G</i> (m ² ha ⁻¹)	18.30	0.44	52.25	0.44
		<i>H₀</i> (m)	21.43	6.70	43.55	7.26
		<i>TV</i> (m ³ ha ⁻¹)	148.42	0.68	522.67	118.14
		<i>AGB</i> (Mg ha ⁻¹)	99.44	0.98	371.55	81.68
<i>P. pinaster</i>	474	<i>N</i> (trees ha ⁻¹)	574.60	10.19	3,176.03	439.15
		<i>G</i> (m ² ha ⁻¹)	22.60	0.42	55.73	13.70
		<i>H₀</i> (m)	16.67	3.40	31.78	6.34
		<i>TV</i> (m ³ ha ⁻¹)	164.05	0.88	460.72	119.37
		<i>AGB</i> (Mg ha ⁻¹)	92.26	0.80	298.64	68.25
<i>P. radiata</i>	408	<i>N</i> (trees ha ⁻¹)	453.66	25.46	1,773.48	294.07
		<i>G</i> (m ² ha ⁻¹)	27.82	0.67	66.62	13.54
		<i>H₀</i> (m)	22.55	5.70	39.55	6.28
		<i>TV</i> (m ³ ha ⁻¹)	246.23	2.25	699.31	147.64
		<i>AGB</i> (Mg ha ⁻¹)	127.43	1.59	356.93	75.38

Minimal preprocessing is applied to the datasets available in GEE, to preserve the maximum amount of information. Depending on the specific application needs, the user can perform different processing steps for both image collections. In this case, we used two different methods of optimizing the remote sensing data by using the cloud computing solutions provided by GEE: one for the Sentinel-1 data (Mullisa et al., 2021) and another for the Sentinel-2 data (Tejido-Murias et al., 2024).

Regarding the Sentinel-1 data, we selected the best images for 2018 based on cover, direction and polarization. The process used to correct Sentinel-1 data is described in Mullissa et al. (2021) and includes additional border noise correction, speckle filtering and radiometric terrain normalization. After applying the corrections, we derived two SAR indices, the Radar Vegetation Index (Trudel et al., 2012) and Vertical Dual Depolarization Index (Periasamy, 2018) and four basic operations between bands.

Regarding Sentinel-2 data, we selected the best images for 2018 based on the solar angle and the percentage of clouds. We corrected the images using the process that we previously developed, which includes atmospheric, topographic and Bidirectional Reflectance Distribution Function (BRDF) corrections (Tejido-Murias et al., 2024), in the GEE platform. After applying the corrections, we derived 21 spectral indices and 18 texture variables based on the normalized difference vegetation index (NDVI), as previously proposed (Novo-Fernández et al., 2024).

Finally, we uploaded forest stand variables and terrain and climate variables from the GEE assets to calculate the zonal statistics in this platform. We combined these variables with remotely sensed data and calculated zonal statistics for all variables in each field plot.

Data analysis

The models were fitted using Random Forest (Breiman, 2001), and repeated 10-fold cross-validation was applied in the R 4.4.2 package 'caret' (Kuhn, 2024) for model evaluation. The overall goodness-of-fit statistics were computed as the average test values obtained from 100 model iterations, based on a 10-fold cross-validation repeated 10 times using the training dataset.

Model performance was assessed using various goodness-of-fit metrics, including the coefficient of determination for nonlinear regression (R^2) (Ryan, 1997), bias (Bias), Root Mean Squared Error (RMSE), and their corresponding relative measures (Bias% and RMSE%).

In this study, since we are dealing with non-parametric and non-linear models, R^2 has been used as a general measure of model fit, as suggested by Ryan (1997), despite the limitations it presents in non-linear regressions. Therefore, we used the coefficient of determination for nonlinear regression (R^2) as suggested by Ryan (1997):

$$R^2 = r_{y_i \hat{y}_i}^2$$

where $r_{y_i \hat{y}_i}$, is the correlation coefficient for a linear regression between the observed and predicted values of the dependent variable (Ryan, 1997).

We applied variable dimensionality reduction, to minimize the collinearity of the data in Weka 3.8 software, by using Wrapper methods (Zhiwei and Xinghua, 2010). We then optimized the Random Forest associated parameters (*mtry* and *N var*) using the 'caret' package (Kuhn, 2024) in R 4.4.2. Finally, we computed the models in the GEE platform using the variables obtained in the dimensionality reduction process and the random forest parameters obtained in the local model optimization process.

To guide the selection of predictors for the final models, the variable importance measure (VIM), computed in R 4.4.2, using the 'caret' package (Kuhn, 2024) was used as a measure to rank the variables in accordance with their predictive power. This measure uses the minimization of the mean square error (MSE) in each of the nodes of the regression tree. To ensure that values of variable importance were expressed on comparable scales, the VIM scores were expressed in relative values (relative importance, $VIM_R = (VIM/ΣVIM) \times 100$). The values per species and the totals have been obtained by averaging the individual values.

A comprehensive description of the data processing workflow is shown in Figure 1.

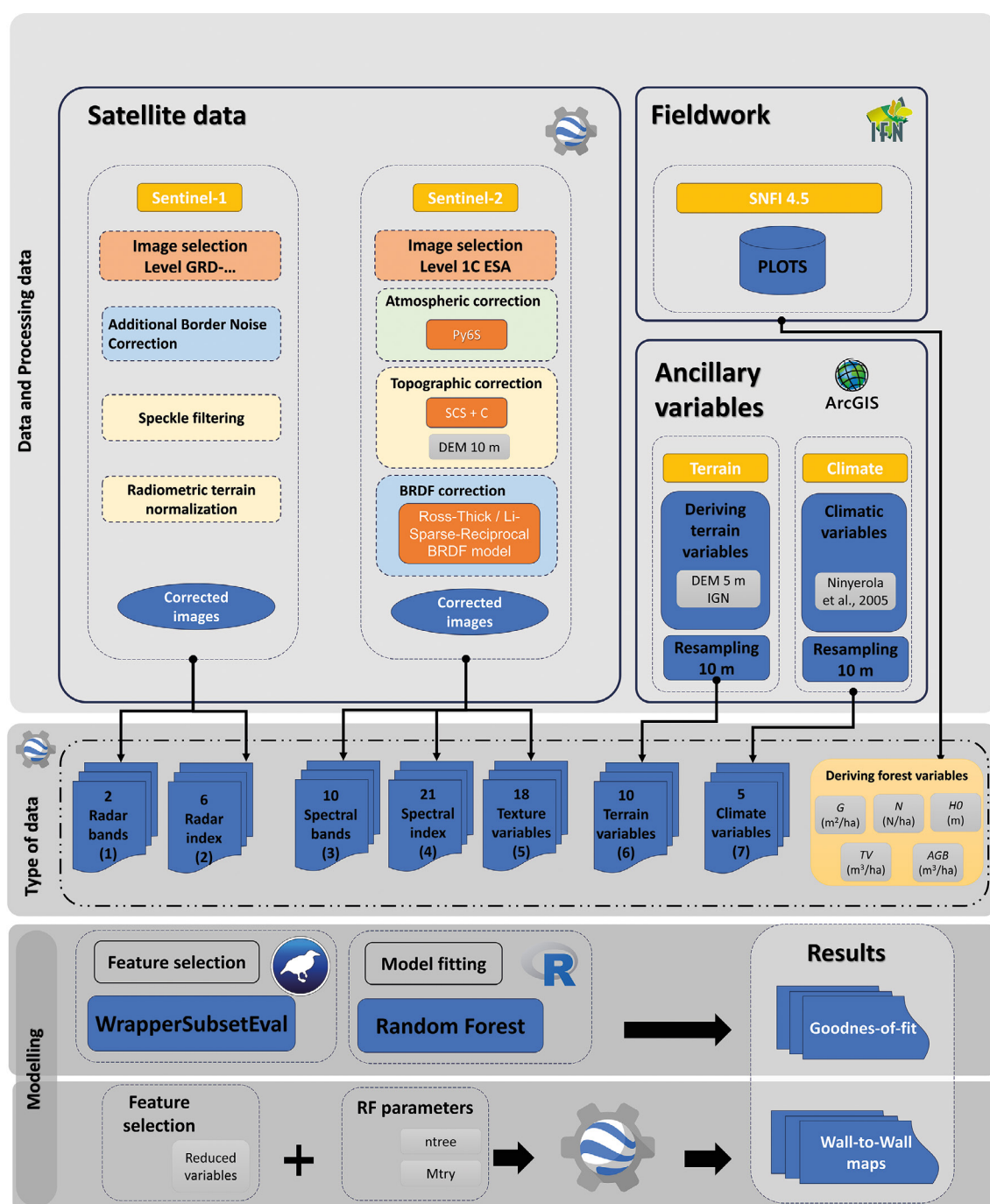


Figure 1. Workflow adopted in the study to model five forest stand variables —basal area (G , $m^2 ha^{-1}$), number of trees per hectare (N , trees ha^{-1}), dominant height (H_0 , m), total volume (TV , $m^3 ha^{-1}$), and aboveground biomass (AGB , $Mg ha^{-1}$)— using multisource data. The input data included Sentinel-1 and Sentinel-2 imagery processed through a series of corrections, National Forest Inventory field data (SNFI 4.5), and ancillary variables. All variables were standardized to a spatial resolution of 10 m. Input features included radar bands and indices, spectral bands and indices, texture metrics, and environmental variables. Feature selection was carried out and Random Forest models were implemented in R to estimate the target forest variables. Model outputs included goodness-of-fit statistics in R and wall-to-wall predictive maps in GEE.

For each species and forest stand variable, we developed a full model including all available explanatory variables listed in Table 1 and Suppl. Table S1. We conducted an initial dimensionality reduction using the Wrapper methods mentioned earlier, and we then developed the prediction models for each species and forest stand variable.

Results

The poorest model fits were obtained for the variables number of trees per hectare (N), with R^2 values ranging from 0.12 to 0.27 and RMSE% between 51.5% and 68.6%, followed by dominant height (H_0), with R^2 values between 0.33 and 0.38 and RMSE% ranging from 22.9% to 30.4%.

By contrast, the best fitting results were obtained for the yield variables, total over-bark volume (TV), aboveground biomass (AGB) and stand density, measured as basal area (G). For these variables the R^2 values were greater than 0.39, except for G in *P. radiata*, which had an R^2 of 0.37, with maximum R^2 values reaching 0.45 (Table 3). The RMSE% for these variables ranged from 38.1% to 62.0%.

Table 3. Goodness-of-fit statistics, number of plots (N° plots), and Random Forest parameters ($mtry$ and N var) for each tree species and dependent variable analyzed in the study. The evaluated forest variables include number of trees per hectare (N), basal area (G), dominant height (H_0), total volume (TV), and aboveground biomass (AGB). Reported statistics include the coefficient of determination (R^2), bias, relative bias (Bias %), and root mean square error expressed as a percentage (RMSE %). The statistics are presented separately for each species: *Eucalyptus globulus*, *Pinus pinaster*, and *Pinus radiata*

Species	N° plots	Forest variable	Goodness-of-fit statistics					
			mtry	N var	R ²	Bias	Bias%	RMSE(%)
<i>E. globulus</i>	544	<i>N</i> (trees ha ⁻¹)	2	16	0.273	-4.390	-0.005	51.48%
		<i>G</i> (m ² ha ⁻¹)	7	16	0.451	0.049	0.003	49.84%
		<i>H</i> ₀ (m)	2	9	0.339	0.006	0.000	27.80%
		<i>TV</i> (m ³ ha ⁻¹)	3	13	0.445	-0.376	-0.003	59.76%
		<i>AGB</i> (Mg ha ⁻¹)	5	13	0.434	-0.515	-0.005	62.03%
<i>P. pinaster</i>	415	<i>N</i> (trees ha ⁻¹)	2	13	0.192	-8.992	-0.016	68.57%
		<i>G</i> (m ² ha ⁻¹)	4	13	0.436	-0.124	-0.005	46.00%
		<i>H</i> ₀ (m)	4	18	0.378	-0.018	-0.001	30.40%
		<i>TV</i> (m ³ ha ⁻¹)	3	12	0.415	-1.746	-0.011	56.32%
		<i>AGB</i> (Mg ha ⁻¹)	6	17	0.394	-0.945	-0.010	58.15%
<i>P. radiata</i>	381	<i>N</i> (trees ha ⁻¹)	13	21	0.120	-7.791	-0.017	61.11%
		<i>G</i> (m ² ha ⁻¹)	8	11	0.393	0.051	0.002	38.13%
		<i>H</i> ₀ (m)	6	11	0.330	-0.005	0.000	22.92%
		<i>TV</i> (m ³ ha ⁻¹)	9	9	0.398	-0.209	-0.001	46.74%
		<i>AGB</i> (Mg ha ⁻¹)	2	2	0.37	0.092	0.001	47.24%

Despite the variability, all models demonstrated very low bias, with higher values observed in the N models. The Bias% ranged between -0.017% and 0.001% (Table 3). Overall, the models can be considered unbiased, as the average Bias% across all models was -0.005%, consistently remaining below 1%.

Following the VIM_R scores (shown in Suppl. Table S2), spectral bands made the largest contribution to the model success (39.11% of the VIM_R), followed by spectral indices (contributing 25.69% of the VIM_R), texture variables (11.69% to the VIM_R), terrain variables (10.63% to the VIM_R), climate variables (7.62% of the VIM_R), radar indices (3.55% of the VIM_R) and radar bands (1.71% of the VIM_R).

Considering the results by species, spectral bands made the greatest contribution to the *E. globulus* and *P. radiata* models (VIM_R of 46.19% and 43.45%, respectively). For *P. pinaster*, the most influential set was the

spectral indices bands, followed by spectral bands (VIM_R of 34.04% and 27.70%, respectively). The second most important variable for *E. globulus* was the spectral index (VIM_R of 27.50%), while for *P. radiata* it was the set of texture variables (VIM_R of 16.97%). However, for all three species, the variables derived from Sentinel-2 images yielded a cumulative importance of between 58.96% and 73.70% (bands + indices).

Other variables contributed less than 16% of accumulative VIM_R , with texture and terrain variables contributing 11.21% and 9.17% for *P. pinaster*, 17.20% and 10.54% for *P. radiata* and 6.88% and 12.18% for *E. globulus*. Climatic variables contributed only 10.81% of the VIM_R for *P. pinaster*, 8.99% for *P. radiata* and 3.08% for *E. globulus*. Finally, the radar bands and derived indices contributed only 5.26% (cumulative VIM_R) with radar bands and radar indices contributing 0.95% and 6.12% for *P. pinaster*; 0% and 4.54% for *P. radiata* and 4.17% and 0% for *E. globulus*.

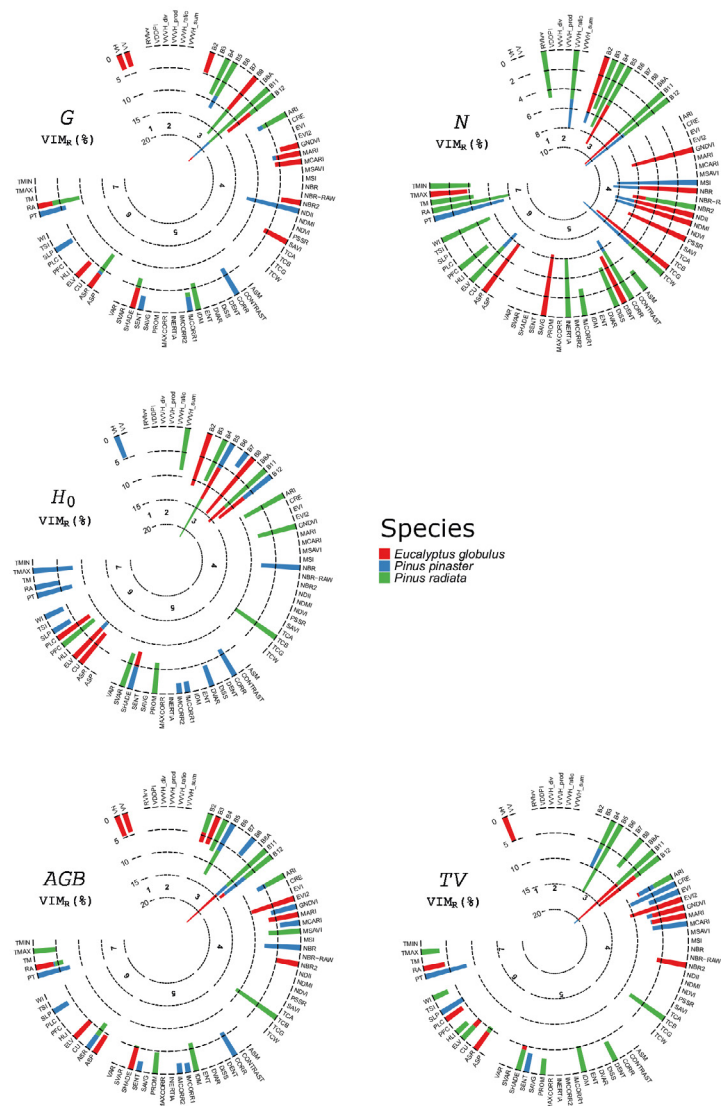


Figure 2. Bar plots showing the VIM_R for the five forest state variables and the three species. The numbers within each plot correspond to the groups of predictor variables: 1) radar bands; 2) radar indices; 3) spectral bands; 4) spectral indices; 5) texture variables; 6) terrain variables; and 7) climate variables. The discontinuous lines, centered in the middle of each plot, indicate the percentage of VIM_R for each variable. All plots have a range from 0 to 20, except for the plot corresponding to the variable “n,” which has a range from 0 to 10. In all cases, zero point is located at the outer ring of the plot. The colors in the plots represent the different species.

The contribution of the radar variables was more important in the *P. pinaster* models than in the *E. globulus* model. For *P. radiata*, the radar variables only contributed information for predicting the total number of trees per hectare and the dominant height. For *E. globulus*, the radar variables were only important in the yield models (*AGB* and *TV*) and density models (*G* and *N*) but not the size model (*H₀*). The greatest contribution was made by the *AGB* model for *E. globulus*, in which the radar band contributed 9% of the VIM_r (see Suppl. Table S2 for more information).

Considering the contribution of each independent variable in the different datasets, we can see that the spectral bands made the greatest contribution in all cases, except for the number of trees per hectare (Suppl. Table S3). The VIM_r for each independent variable, models and species is represented in Figure 2.

Discussion

Model accuracy and variable importance

The model evaluation results are typical for these types of models (Dos Reis et al., 2018; Astola et al., 2019). The evaluation results for the models developed in this study are very similar to those obtained for the models developed by Novo-Fernández et al. (2024). Optical bands yielded the highest VIM_r in all cases, and the contribution of radar bands was less important.

Contribution of radar data

Synthetic Aperture Radar (SAR) is a promising tool for studying forest biomass due to its capability to penetrate the forest canopy to varying depths depending on the wavelength of the sensor, its sensitivity to water content in vegetation, and its independence from weather conditions (Le Toan et al., 1992; Kasischke et al., 2011).

Several studies have provided information regarding the prediction of forest stand variables using radar satellite data (Roy et al., 2021; Fang et al., 2023). However, our study differs from many of these studies, which were conducted at a local scale, using a small number of ground truth plots or focused on a single species or type of forest. As a result, studies covering large areas, with different species and large numbers of plots, are scarce (Kellndorfer et al., 2014).

Our findings indicate that the inclusion of radar bands does not improve the models relative to previously developed models (Novo-Fernández et al., 2024b). These results are similar to those reported by (Fang et al., 2023) but contrast to those reported by other authors, e.g. (Roy et al., 2021). Such discrepancies may be related to the differences in the study conditions and the difference in the size of the area covered (local vs regional) and the database used (small vs large database). In addition, the topography, which varies widely in the study area, can affect the radar information (layover and foreshortening) thus hampering the use of this data in more rugged areas.

Backscattering of land cover surfaces is greatly influenced by variables that depend on the sensor, such as the wavelength (e.g. X, C, L, P) and polarization, DUAL-POL or QUAD-POL (e.g. HH, VV, HV, VH), and by others that depend on the surface, such as land cover and terrain properties (e.g. roughness, dielectric constant or wetness) and yet others affected by both factors, such as incidence angle. The variations within and between images lead to significant differences in backscatter values available from optical and radar-derived variables for a plot, even when the plot remains constant (e.g. the same volume of wood in a month), which hampers use of this type of information for large-scale analysis.

Sentinel-1 provide two types of data: i) Single Look Complex (SLC) images, which preserve the phase information and ii) Ground Range Detection (GRD), in which phase information is lost. These are the only types of data available in the GEE repository. Moreover, Sentinel-1, which operates in the C-band, which is not generally used in forestry structure predictions (Le Toan et al., 1992; Lu, 2006). While the L-band enables biomass information to be obtained and is the basis for the upcoming NISAR mission (NASA) and ROSE-L mission (ESA), the X-band enables the retrieval of interferometric heights that can be correlated with average or dominant height, for example TanDEM-X constellation. However, the C-band, which lies between the two previous bands, does not enable either of these types of analysis to be performed adequately, although it has been used in the analysis of other forestry variables such as canopy water content and wetness (Rao et al., 2020).

The type of study conducted with radar images is also a factor to consider in the comparison, as Single Look Complex (SLC) and QUAD-POL images are required in the complete study.

Due to the incompatibility of the GEE platform to use Single Look Complex (SLC) images, interferometric, polarimetric or tomography analyses cannot be performed in GEE. These types of analysis are commonly used to estimate forest variables such as forest height, which is correlated with other important forest variables (Papathanassiou et al., 2021, Ramachandran et al., 2023, Xing et al., 2023 and; Solberg et al., 2024).

In special cases, radar information has been used to complete the optical information when high cloud cover occurred in the study area (Aranguren-Erice, 2021), and Sentinel-1 images can be used to complement the information provided by Sentinel-2 images, as previously reported (Fang et al., 2023). On the other hand, although the interferometry synthetic aperture radar (InSAR) can provide information about forest horizontal structure, it cannot be developed in the GEE platform as the phase information is lost during storage.

Automatization, limitations and future developments

One of the strong points of this study is the automatized image pre-processing, which enables useful models and tools to be developed that will provide replicable information in the coming years. By contrast, the main limitation is the difficulty in obtaining images without clouds.

Future developments will include attempting to adjust the models using temporal series or to use deep learning algorithms to minimize the dependence on images without clouds. The construction of models for certain forest areas and species using other techniques such as InSAR, PolSAR, PolInSAR and TomoSAR is also a possible solution for accurate estimation of forest production in these areas. Other possible developments include the inclusion of other radar sensors, such as NISAR.

Conclusions

GEE processing and local processing produce the same results regarding estimation of forest parameters. In the case of GEE, images of the study area can be updated with a simple script, do not require downloading and preprocessing can be conducted with the server side. Sentinel-1 GRD radar variables do not provide further information about forest state variables in forest plantations in northern Spain. The search for the best use of radar information should involve trying different approaches to preprocessing and/or using different types of data (bands, satellite, etc.) not included in the GEE platform.

Supplementary material (Tables S1, S2 and S3): accompanies the paper on *Forest System's* website.

Competing of Interest: The authors have declared that no competing interests exist.

Author Contributions: **Iyán Tejjido-Murias:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Marcos Barrio-Anta:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Carlos Antonio López-Sánchez:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **All authors have read and agreed to the published version of the manuscript.**

Funding:

Funding agencies/institutions	Project / Grand
MCIN/AEI/10.13039/ 501100011033	PID2020-112839RB-I00
Severo Ochoa Fellowship from the Asturias Government	BP21-125

References

- Aguirre A, Moreno-Fernández D, Alberdi I, Hernández L, Adame P, Cañellas I, Montes F, 2022. Mapping forest site quality at national level. *Ecol Manag* 508, 120043. <https://doi.org/10.1016/j.foreco.2022.120043>
- Aranguren-Erice I, 2021. Relleno de huecos en series temporales de índices Sentinel-2 en cubiertas agrícolas con datos procedentes de Sentinel-1. Trabajo Fin de Master. Universidad Pública de Navarra. <https://academica-e.unavarra.es/entities/publication/flb611ff-156e-42a5-ac46-47f8c2629f22>.
- Astola H, Häme T, Sirro L, Molinier M, Kilpi J, 2019. Comparison of Sentinel-2 and Landsat 8 imagery for forest variable prediction in boreal region. *Remote Sens Environ* 223:257–73. <https://doi.org/10.1016/j.rse.2019.01.019>
- Breiman L, 2001. Random forests. *Mach Learn*. 45: 5–32. <https://doi.org/10.1023/A:1010933404324>
- Fang G, Xu H, Yang SI, Lou X, Fang L. 2023. Synergistic use of Sentinel-1, Sentinel-2, and Landsat 8 in predicting forest variables. *Ecol Indic* 151:110296. <https://doi.org/10.1016/j.ecolind.2023.110296>
- Kasischke ES, Tanase MA, Bourgeau-Chavez LL, Borr M, 2011. Soil moisture limitations on monitoring boreal forest regrowth using spaceborne L-band SAR data. *Remote Sens Environ* 115:227–32. <https://doi.org/10.1016/j.rse.2010.08.022>
- Kellndorfer J, Cartus O, Bishop J, Holecz WW and F, Kellndorfer J, Cartus O, 2014. Large Scale Mapping of Forests and Land Cover with Synthetic Aperture Radar Data. *Land Applications of Radar Remote Sensing*. <https://doi.org/10.5772/58220>
- Kuhn, M. 2024. *caret: Classification and Regression Training* (Version 7.0-1) [R package]. Comprehensive R Archive Network (CRAN). <https://github.com/topepo/caret/>
- Le Toan T, Beaudoin A, Riom J, Guyon D, 1992. Relating Forest Biomass to SAR Data. *IEEE Transactions on Geoscience and Remote Sensing*; 30 (2), 403–11. <https://doi.org/10.1109/36.134089>
- Lu D, 2006. The potential and challenge of remote sensing-based biomass estimation. *Int J Remote Sens* 27: 1297–328. <https://doi.org/10.1080/01431160500486732>
- MAPAMA, 2019. Spanish National Fourth Inventory updating. Ministerio de Agricultura, Pesca y Alimentación Gobierno de España. https://www.miteco.gob.es/es/biodiversidad/temas/inventarios-nacionales/inventario-forestal-nacional/cuarto_inventario.html
- MITECO, 2023. Anuario de Estadística Forestal. Gobierno de España. Ministerio para la Transición Ecológica y el Reto Demográfico. https://www.miteco.gob.es/es/biodiversidad/estadisticas/forestal_anuarios_todos.html.
- Mullissa A, Vollrath A, Odongo-Braun C, Slagter B, Balling J, Gou Y, 2021. Sentinel-1 SAR Backscatter Analysis Ready Data Preparation in Google Earth Engine. *Remote Sensing* 13 (10) 1954. <https://doi.org/10.3390/rs13101954>
- Ninyerola M, Pons X y Roure JM. 2005. Atlas Climático Digital de la Península Ibérica. Metodología y aplicaciones en bioclimatología y geobotánica. ISBN 932860-8-7. Universidad Autónoma de Barcelona, Bellaterra.
- Novo-Fernández A, Barrio-Anta M, Recondo C, Cámara-Obregón A, López-Sánchez CA, 2019. Integration of National Forest Inventory and Nationwide Airborne Laser Scanning Data to Improve Forest Yield Predictions in North-Western Spain. *Remote Sensing* 11 (14). <https://doi.org/10.3390/rs11141693>
- Novo-Fernández A, López-Sánchez CA, Cámara-Obregón A, Barrio-Anta M, Tejjido-Murias I, 2024. Estimating Forest Variables for Major Commercial Timber Plantations in Northern Spain Using Sentinel-2 and Ancillary Data. *Forests* 15 (1), 99. <https://doi.org/10.3390/f15010099>
- Papathanassiou KP, Cloude SR, Pardini M, Quiñones MJ, Hoekman D, Ferro-Famil L, 2021. Forest Applications. *Remote Sensing and Digital Image Processing* 25:59–117. https://doi.org/10.1007/978-3-030-56504-6_2
- Periasamy S, 2018. Significance of dual polarimetric synthetic aperture radar in biomass retrieval: An attempt on Sentinel-1. *Remote Sens Environ* 217:537–49. <https://doi.org/10.1016/j.rse.2018.09.003>
- Ramachandran N, Saatchi S, Tebaldini S, d'Alessandro MM, Dikshit O, 2023. Mapping tropical forest aboveground biomass using airborne SAR tomography. *Sci Rep* 13, 6233. <https://doi.org/10.1038/s41598-023-33311-y>
- Rao K, Williams AP, Flefil JF, Konings AG, 2020. SAR-enhanced mapping of live fuel moisture content. *Remote Sens of Environ* 245, 111797. <https://doi.org/10.1016/j.rse.2020.111797>
- Dos-Reis AA, Carvalho MC, de Mello JM, Gomide LR, Ferraz Filho AC, Acerbi FW, 2018. Spatial prediction of basal area and volume in Eucalyptus stands using Landsat TM data: an assessment of prediction methods. *N Z J For Sci* 48, 1. <https://doi.org/10.1186/s40490-017-0108-0>
- Roy S, Mudi S, Das P, Ghosh S, Shit PK, Bhunia GS, 2021. Estimating Above Ground Biomass (AGB) and Tree Density using Sentinel-1 Data. *Environmental Science and Engineering* 259–80. https://doi.org/10.1007/978-3-030-56542-8_11

- Ryan TP, 1997. Modern regression methods. John Wiley & Sons, New York. <https://doi.org/10.1002/9780470382806>
- Solberg S, Bollandsås OM, Gobakken T, Næsset E, Basak P, Duncanson LI, 2024. Biomass Change Estimated by TanDEM-X Interferometry and GEDI in a Tanzanian Forest. *Remote Sensing*, 16 (5), 861. <https://doi.org/10.3390/rs16050861>
- Tejjido-Murias I, Barrio-Anta M, López-Sánchez CA, 2024. Evaluation of Correction Algorithms for Sentinel-2 Images Implemented in Google Earth Engine for Use in Land Cover Classification in Northern Spain. *Forests*, 15 (12), 2192. <https://doi.org/10.3390/f15122192>
- Trudel M, Charbonneau F, Leconte R, 2012. Using RADARSAT-2 polarimetric and ENVISAT-ASAR dual-polarization data for estimating soil moisture over agricultural fields. *Canadian Journal of Remote Sensing*, 38 (4), 514–27. <https://doi.org/10.5589/m12-043>
- Xing C, Wang H, Zhang Z, Yin J, Yang J, 2023. A Review of Forest Height Inversion by PolInSAR: Theory, Advances, and Perspectives. *Remote Sensing*, 15 (15), 3781. <https://doi.org/10.3390/rs15153781>
- Zhiwei X, Xinghua W, 2010. Research for Information Extraction Based on Wrapper Model Algorithm. 2010 Second International Conference on Computer Research and Development, p. 652–5. <https://doi.org/10.1109/ICCRD.2010.141>