# Bayesian Models to Predict Above Ground-Biomass

Ricardo Coelho[1(✉)], Isabel Natário[1,2], and Silvia Fraile[3]

[1] NOVA MATH – Center for Mathematics and Applications, Nova University of Lisbon, Caparica, Portugal
rpe.coelho@campus.fct.unl.pt
[2] Department of Mathematics, Nova School of Science and Technology, Nova University of Lisbon, Caparica, Portugal
icn@fct.unl.pt
[3] GEOSAT SATELITES SLU, Boecillo, Valladolid, Spain
silvia.fraile@geosat.space

**Abstract.** Accurate estimation of Above-Ground Biomass (AGB) plays a key role in monitoring forest carbon stocks and supporting climate mitigation policies. This study compares four Bayesian models for AGB prediction, a Geostatistical model (BGM), a Bayesian Additive Regression Trees (BART) model, the recently proposed BARTSIMP model, and a novel Hybrid BART (HBART) framework inspired by Random Forest–Ordinary Kriging. The HBART model combines the flexibility of BART in capturing non-linear relationships with a spatial model for residual dependence. These models are applied to predict AGB in forest areas of the *Páramos-Valles* region (Spain). To this end, remote sensing data from the GEOSAT-2 satellite are integrated with field inventory data from the Spanish National Forest Inventory, including satellite variables such as reflectance bands, vegetation indices, and texture. In addition, topographic covariates such as altitude, slope, and aspect are also considered. Predictive performance is assessed using spatial block cross-validation and a set of accuracy and uncertainty metrics. Results show that the BGM achieves the best point prediction accuracy, while HBART provides the most reliable uncertainty quantification, with substantially higher coverage probabilities.

**Keywords:** Above-Ground Biomass · Bayesian machine learning · Geostatistical model · Remote sensing data

## 1 Introduction

Monitoring carbon sequestration plays a crucial role in climate change mitigation and ecosystem management. Above-Ground Biomass (AGB) is an indicator that well describes the amount of carbon present in forests and is widely used [17]. AGB can be estimated using direct or indirect methods [18]. Direct methods involve the destruction of trees while indirect methods do not. Among indirect methods, the allometric equations are quite used. These equations use field

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16767, pp. 660–678, 2027.
https://doi.org/10.1007/978-3-032-30491-9_41

data about characteristics of trees (for example, diameter at breast height, tree height, among others) to estimate AGB. Although this method provides reliable estimates at the plot level, it is typically labor-intensive, costly, and spatially limited, thus making it difficult to apply over large areas. Recent advances in remote sensing technologies have provided a faster and more cost-effective alternative. Satellite imagery enables the extraction of variables, including spectral bands, vegetation indices, and texture measures, which are commonly used as covariates in AGB estimation. Integrating field measurements with variables derived from remote sensing data enables the use of statistical models and Machine Learning (ML) models to produce estimates of AGB.

Several ML models such as Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting, have been applied to AGB prediction using remote sensing data [10,17]. Although these models can capture complex non-linear relationships between AGB and covariates, many traditional ML approaches do not explicitly account for predictive uncertainty, which is particularly important in environmental applications. Bayesian ML models provide a useful alternative, as they incorporate uncertainty both in the model parameters and in the predictions. One of the most widely used Bayesian ML models is the Bayesian Additive Regression Trees (BART) model, proposed by Chipman et al. (2010) [2]. This model represents the response variable as the sum of multiple regression trees, allowing flexible non-linear relationships and interactions between covariates.

In recent years the BART model has been applied to spatial data [16], although without accounting for spatial autocorrelation. For example, Yankey et al. (2024) [20] applied the BART model to disaggregate census population data from district-level totals to high-resolution grid cells. Zhang et al. (2020) [22] applied the BART model to environmental data, for predicting the concentration of the air pollutant $PM_{2.5}$. This can lead to underestimated prediction uncertainty and reduced predictive performance when modelling spatial data [17]. Also, AGB in forested areas generally exhibits spatial heterogeneity and autocorrelation, and in some cases, non-stationarity. Moreover, remote sensing variables are themselves spatially and temporally correlated [6,19].

To overcome this limitation, Jiang et al. (2025) [7] developed the Bayesian Additive Regression Trees for Spatial INLA Modelling and Prediction (BART-SIMP) model. This model combines the BART model with spatial modelling of spatial autocorrelation. Geostatistical modelling provides an alternative approach by treating AGB as a continuous spatial process observed with error at discrete locations. These models are advantageous because they naturally account for spatial dependence, integrate covariates, and quantify uncertainty, which is essential for informed decision-making [5]. Unlike most ML methods, geostatistical models explicitly model spatial autocorrelation, providing interpretable uncertainty estimates.

Recognizing the strengths of both statistical and ML approaches, hybrid methods have been proposed, as the Random Forest Ordinary Kriging (RFOK), in which the deterministic component is modelled using the Random Forest (RF)

model, while Ordinary Kriging interpolates the residuals obtained from the RF model [21]. Inspired by this, we propose in this study a hybrid BART model to improve predictive performance and capture spatial patterns in AGB.

The main purpose of this study is to evaluate and compare different modelling strategies for predicting AGB using remote sensing data. Specifically, the goal is to assess the predictive performance of four approaches: a Bayesian Geostatistical Model, the BART Model, the hybrid BARTSIMP Model, and a new hybrid BART framework. This study applies the BART and BARTSIMP models to AGB estimation for the first time. The study area corresponds to the Forest Area of *Páramos-Valles*, located in the province of *Palencia*, within the autonomous community of *Castilla y León* in northwestern Spain.

The paper is organized as follows: Sect. 1 provides an introduction to the work; Sect. 2 describes the datasets; Sect. 3 presents the models considered, together with the evaluation metrics used for their assessment; Sect. 4 presents the results; and Sect. 5 discusses the main findings and concludes.

## 2 Data

In this study three datasets are used to estimate AGB(t/ha): field data from Fourth Spanish Forest Inventory (*Cuarto Inventorio Forestal Nacional* - IFN4); Forest Map of Spain (*Mapa Forestal de España* - MFE); and remote sensing data from GEOSAT-2 satellite, operated by the company GEOSAT.

### 2.1 Ground Data

The study area is depicted in Fig. 1.

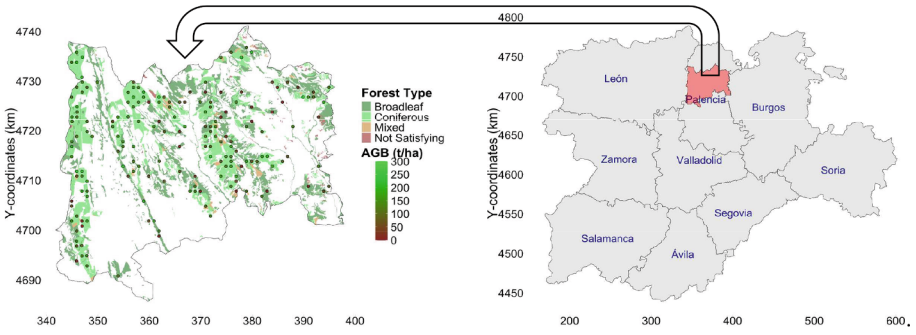



**Fig. 1.** Area of Interest: The left panel illustrates the forested areas along with the *estadillos*, the right panel depicts the provinces within *Castilla y León*.

Data from the IFN4 were used to calculate the response variable, AGB(t/ha). These field data were acquired and processed between 2008 and 2018, corresponding to one of the IFN 10-year cycles [3]. The inventory collects information

about trees within permanent sampling plots, the *estadillos*. Each *estadillo* is a circular plot of radius 25 m, georeferenced by its center coordinates.

Each *estadillo* is divided into four concentric circles of different radius [3]. For each tree, the IFN records information about its characteristics. Only trees within specific diameter at breast height classes are measured in each circle. AGB for individual trees is calculated using allometric equations, with an expansion factor used to extrapolate AGB to the *estadillo* level (t/ha). Ground data may contain uncertainty associated with measurement errors, which propagates into AGB estimates through allometric equations. The total AGB of an *estadillo* is obtained by adding the weighted contributions of all trees within the *estadillo*. Only *estadillos* within Forest Areas, as defined by the MFE (i) tree cover fraction exceeding 20%, ii) mature tree height greater than 3 m, and iii) a minimum area of 1 ha), were included. All others were excluded.

In the *Parámos-Valles* region, a total of 180 *estadillos* belong to the Forest Areas, of which 105 belong to Coniferous forests, 59 belong Broadleaf forests and the remainder belong to Mixed forests.

### 2.2 Remote Sensing Data

Remote sensing data used in this study were obtained from the GEOSAT-2 satellite, with imagery acquired between May and August 2020. The satellite collects information in five spectral bands: one panchromatic band and four multispectral bands (Red, Blue, Green, and Near Infrared (NIR)).
The multispectral bands have a spatial resolution of 3 m. Several covariates were derived from the multispectral bands, including the reflectance bands (NIR, Red, Green, and Blue), vegetation indices, such as the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Green Chlorophyll Index (GCI), and the Normalized Difference Water Index (NDWI) and texture metrics, specifically homogeneity. These variables were the covariates considered in this study.

In addition, topographic covariates as Elevation, Slope and Aspect were included. These variables were obtained from the Digital Terrain Model (DTM) produced by the *Centro Nacional de Información Geográfica* (CNIG) of Spain [13], developed as part of the *Plan Nacional de Ortofotografía Aérea* (PNOA) and based on LiDAR data from the First LiDAR Coverage.

### 2.3 Integrate Ground Data with Remote Sensing Data

To integrate the datasets from different sources, it was necessary to perform a spatial alignment between the remote sensing data and the field data. This process involves orthorectify the images and align them with the study coordinate reference system (EPSG:25830) using ground control points and a 5 m Digital Elevation Model (DEM) to compensate for terrain related distortions. For each *estadillo*, only pixels entirely within the boundaries were selected, and any partially overlapping pixels were excluded. The remote sensing metrics for each *estadillo* were then considered as the median values of the selected pixels.

# 3 Methods

## 3.1 Bayesian Geostatistical Model (BGM)

Geostatistical models are designed to capture spatial dependence in data observed at known locations [9]. Let $\mathcal{D} \subset \mathbb{R}^2$ denote a fixed and continuous spatial domain [5]. Spatial phenomena observed over $\mathcal{D}$ can be described as realizations of a stochastic process $\{U(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$. Geostatistical (point-referenced) data consist in observations collected at a finite set of locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \mathcal{D}$, yielding measurements $\{y(\mathbf{s}_i)\}_{i=1}^n$, which can be viewed as a partial realization of the process $U(\mathbf{s})$. This process is often modelled as a Gaussian Field (GF), where the assumptions of second-order stationarity and isotropy are assumed.

The spatial dependence can be analysed using a semivariogram, describing how variability between observations changes with distance, providing guidance for selecting appropriate covariance functions in spatial modelling [4]. In addition, Monte Carlo envelopes can be used to assess the statistical significance of spatial autocorrelation by comparing the empirical semivariogram with simulations under spatial independence hypothesis.

In geostatistical models, the linear predictor combines a deterministic component, which captures large-scale variation through covariates, with a spatially structured random effect accounting for residual spatial dependence [5]. The spatial random effect $u(\mathbf{s})$ is assumed to be a realization of a latent stochastic process $U(\mathbf{s})$. This process is commonly modelled as a GF with zero mean and covariance function $C(u(\mathbf{s}_i), u(\mathbf{s}_j))$. Under the assumptions of second-order stationarity and isotropy, the covariance structure depends only on the Euclidean distance between locations, that is, $\mathrm{Cov}\big(u(\mathbf{s}_i), u(\mathbf{s}_j)\big) = C\big(\|\mathbf{s}_i - \mathbf{s}_j\|\big)$, where $\|\mathbf{s}_i - \mathbf{s}_j\|$ denotes the Euclidean distance between $\mathbf{s}_i$ and $\mathbf{s}_j$.

Under this framework, the Generalized Geostatistical Model is defined hierarchically:

$$\begin{aligned} y(\mathbf{s}_i) &\sim \text{Exponential-Family}\big(\mu(\mathbf{s}_i), \sigma_\varepsilon^2\big) \\ h\big(\mu(\mathbf{s}_i)\big) &= \eta(\mathbf{s}_i) \\ \eta(\mathbf{s}_i) &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + u(\mathbf{s}_i), \;\; i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $y(\mathbf{s}_i)$ is the observed response at location $\mathbf{s}_i$, $\mathbb{E}[y(\mathbf{s}_i)] = \mu(\mathbf{s}_i)$ is its mean linked to the linear predictor $\eta(\mathbf{s}_i)$ through a link function $h(\cdot)$, $\beta_0$ the intercept, $\beta_j$ the regression coefficients for covariates $x_j$, $u(\mathbf{s}_i)$ the spatial random effect, and $\sigma_\varepsilon^2$ the variance of response variable, if required.

Taking a Bayesian approach for inference, Rue et al. (2009) [15] developed the deterministic methodology known as Integrated Nested Laplace Approximation (INLA), which efficiently approximates the posterior distributions of model parameters and hyperparameters in hierarchical Bayesian models. INLA is implemented in the `R` software through `R-INLA` package.

The Stochastic Partial Differential Equations (SPDE) approach can be used to model the spatial random effects [9]. In this approach, the exact stationary

solution of a given SPDE corresponds to a GF with a Matérn covariance structure:

$$\text{Cov}\left[u(\mathbf{s}_i), u(\mathbf{s}_j)\right] = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}}\left(\kappa||\boldsymbol{s_i} - \boldsymbol{s_j}||\right)^{\lambda} K_{\lambda}\left(\kappa||\boldsymbol{s_i} - \boldsymbol{s_j}||\right), \tag{2}$$

where, $\sigma^2$ denotes the marginal variance of the spatial process, $K_\lambda$ is the modified Bessel function of the second kind of order $\lambda$, which controls the smoothness of the process ($\lambda > 0$), and $\kappa > 0$ is a scaling parameter related to the spatial range $r$. The term $||\boldsymbol{s_i} - \boldsymbol{s_j}||$ represents the Euclidean distance between locations $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$.

Using the Finite Element Method, the GF is approximated by a discretely indexed Gaussian Markov Random Field (GMRF), through a basis function representation in $\mathcal{D}$. This approach, implemented in the `R-INLA` package, enables efficient estimation of spatial dependence in the data. A key step in this approach is the construction of a triangulated mesh that discretizes the spatial domain and defines the nodes for the approximation. The mesh design should follow the recommendations of Moraga et al. (2021) [12] and Bakka (2017) [1].

Within the Bayesian approach, prior distributions of the parameters are added to the hierarchical formulation. In this study, the default non-informative priors provided by `R-INLA` for precision parameters and fixed effects were considered. For the hyperparameters of the Matérn covariance function, Penalized Complexity (PC) priors were used to control model flexibility and avoid overfitting [9], where $P(r < 15500) = 0.3$ and $P(\sigma > 48) = 0.1$, having these values been selected from the semivariogram of the linear model residuals and the spatial distances between observation points.

### 3.2 Bayesian Additive Regression Trees (BART)

The Bayesian Additive Regression Trees (BART) model is a Bayesian model that uses non-parametric regression trees to approximate the function $f(x) = \mathbb{E}[y|x]$. Here $f(x)$ is represented as a sum of $m$ regression trees, each contributing a small portion to the overall prediction. Developed by Chipman et al. (2010) [2], this model has been applied to problems where the response variable is continuous, categorical, or time-to-event [8]. The BART model can be represented as:

$$\mathbf{y} \approx \sum_{j=1}^{m} g(x; T_j, M_j) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \text{Normal}(0_n, \sigma_{\varepsilon}^2 I_n), \tag{3}$$

where each function $g(x; T_j, M_j)$ represents the contribution of the $j^{th}$ tree, $j = 1, \ldots, m$, and $\boldsymbol{\varepsilon}$ is the error term. Each tree is defined by its structure $T_j$ and a set of terminal node parameters $M_j = \{\mu_{j1}, \ldots, \mu_{jb_j}\}$ assigned to its leaves [2]. The value $b_j$ is the number of terminal nodes for the $j^{th}$ tree.

The BART model has three main stochastic components: the residual error with variance $\sigma_{\varepsilon}^2$, the structures of the regression trees $T_1, \ldots, T_m$, and the values at the leaf nodes $M_1, \ldots, M_m$ [2]. A prior is placed over the ensemble by assuming

independence between trees, between leaf nodes conditional on their tree, and between the residual variance and the trees:

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma) = \prod_{j=1}^{m} p(T_j)\, p(M_j \mid T_j)\, p(\sigma). \tag{4}$$

The tree structure prior $p(T_j)$ favors shallow trees through a depth-dependent splitting probability of the form $\alpha(1+d)^{-\beta}$, where $\alpha \in ]0,1[$, $\beta \in [0,\infty[$, and node at depth $d \in \{0,1,2\}$, which regularizes the model and helps prevent overfitting [2,8]. At each internal node, splitting variables are selected uniformly, and cutpoints are chosen uniformly from the set of available values. Conditional on the tree structure, terminal node parameters are assumed independent and normally distributed, $\mu_{jl} \sim N(\mu_\mu/m, \sigma_\mu^2)$, $l = 1, \ldots, b_j$, shrinking individual tree contributions toward zero. The residual variance is assigned an inverse-gamma prior, $\sigma_\varepsilon^2 \sim \text{Inv-Gamma}\left(\frac{\nu}{2}, \frac{\nu\lambda^2}{2}\right)$, typically chosen to be weakly informative. The $\nu$ denotes the degrees of freedom and the scale parameter $\lambda$ is calibrated from the data so that, by default, there is a prior probability of 0.90 that the BART model improves upon the root mean square error (RMSE) of a linear regression model.

Model estimation is performed using Markov Chain Monte Carlo (MCMC), specifically a backfitting Gibbs sampler that iteratively updates each tree conditional on the others [2,8]. This procedure generates posterior samples for all model parameters as well as for predictions. Posterior predictions are obtained by averaging over the MCMC samples, while uncertainty is quantified using credible intervals, typically constructed from the 2.5th and 97.5th percentiles of the posterior distribution.

The parameter $m$ in the BART model can be treated as an unknown parameter with a prior distribution, following a fully Bayesian approach. This implementation has very high computational costs. Thus, an alternative and more computationally efficient strategy is to select the value of $m$ using cross-validation methods. However, this strategy often result in very large number of trees, which results in very high computational costs. Consequently we have chosen to follow the advice to reduce this number to a more moderate value [2].

In this study, the BART model was implemented using the `bartMachine` package in `R`. The optimal number of trees, $m = 50$, was selected via 10-fold cross-validation by minimizing RMSE. The default values of the hyperparameters $\alpha$ and $\beta$ are 0.95 and 2, respectively. In fitting this model, a total of $N = 4000$ MCMC iterations were performed, of which the first $B = 2000$ were discarded as burn-in, while the remaining $N - B = 2000$ were used for posterior inference. To interpret the fitted model, Partial Dependence Plots (PDP) were used. These plots represent the marginal effect of a covariate on the predicted response, averaging over the observed values of the remaining covariates [8]. PDP provide a useful tool for visualizing non-linear relationships and identifying regions where covariates have a stronger influence on the response variable.

### 3.3 Hybrid Bayesian Additive Regression Trees (HBART)

Using the BART model to analyse geostatistical data has the drawback of not explicitly accounting for the spatial autocorrelation present in the observations. To address this limitation, we propose a Hybrid Bayesian Additive Regression Trees (HBART) model, that is inspired in the traditional Hybrid ML models used in geostatistics, such as the Random Forest-Ordinary Kriging (RFOK) model [21].

Similarly to the RFOK hybrid framework, the proposed Hybrid BART model is estimated in two stages. In the first stage, a BART model is fitted to the response variable using the available covariates. This step captures potentially complex non-linear relationships between the covariates and the response variable. However, because BART does not explicitly model spatial dependence, the residuals obtained from this first stage may still exhibit spatial autocorrelation. Let $\varepsilon(\mathbf{s}_i) = y(\mathbf{s}_i) - \hat{y}_{BART}(\mathbf{s}_i)$ denote the residual of the BART model at location $\mathbf{s}_i$. In the second stage, these residuals are modelled using a BGM just with an intercept term and spatial random effects. The final prediction at location $\mathbf{s}_0$ is obtained by combining the BART prediction with the predicted spatial residual component:

$$\hat{\mathbf{y}}(\mathbf{s}_0) = \hat{\mathbf{y}}_{\text{BART}}(\mathbf{s}_0) + \hat{\varepsilon}_{\text{BGM}}(\mathbf{s}_0), \tag{5}$$

where $\hat{\mathbf{y}}_{\text{BART}}(\mathbf{s}_0)$ denotes the posterior predictive mean obtained from the BART model and $\hat{\varepsilon}_{\text{BGM}}(\mathbf{s}_0)$ represents the posterior predictive mean of the BART residual estimated using the BGM model.

Since the BART is Bayesian model, it provides posterior samples from the predictive distribution. These samples allow the construction of credible intervals that quantify the uncertainty associated with the prediction. In this method the prediction intervals are obtained by combining the posterior predictive uncertainty from the BART model with the uncertainty associated with the spatial prediction of the residuals. In practice, posterior samples from the BART predictive distribution are added with posterior samples from the spatial residual model, yielding a posterior predictive distribution for $y(\mathbf{s}_0)$. Credible intervals are then be derived from the corresponding quantiles. For the spatial residual model, the same PC-priors as in the BGM model were used.

### 3.4 BART for Spatial INLA Modelling and Prediction (BARTSIMP)

The BART model presented in Subsect. 3.2 offers a flexible way to capture complex non-linear relationships and interactions between covariates. However, it does not explicitly account for spatial dependence, which can play an important role when analysing spatially referenced data. To overcome this limitation, Jiang et al. (2025) [7] developed the BART for Spatial INLA Modelling and Prediction (BARTSIMP) model. In contrast to the HBART two stage approach, BARTSIMP defines a fully joint Bayesian hierarchical model in which covariate effects and spatial dependence are estimated simultaneously.

BARTSIMP combines the sum-of-trees structure of BART for modelling covariate effects with a latent spatial GF that captures residual spatial dependence. Let $y_i$ denote the response variable at spatial location $i$, $\mathbf{x}_i$ the vector of covariates, and $u_i$ the latent spatial effect. Let $\mathbf{T} = (T_1, \ldots, T_m)$ denote the collection of tree structures and $\boldsymbol{M} = (M_1, \ldots, M_m)$ the corresponding terminal node parameters of the BART model. The model can be expressed as

$$y_i \mid u_i, \mathbf{x}_i, \mathbf{T}, \boldsymbol{M}, \sigma_\varepsilon^2 \sim N\left(\sum_{l=1}^{m} g(\mathbf{x}_i; T_l, M_l) + u_i,\ \sigma_\varepsilon^2\right), \tag{6}$$

where $\sum_{l=1}^{m} g(\mathbf{x}_i; T_l, M_l)$ is the BART sum-of-trees contribution for the covariates, $u_i$ is the spatial random effect, and $\sigma_\varepsilon^2$ is the residual variance.

The BARTSIMP model assumes a prior independence between the tree parameters $(\mathbf{T}, \boldsymbol{M})$, the residual variance $\sigma_\varepsilon^2$, and the hyperparameters $\boldsymbol{\psi}$ of the spatial random effects modelled as a GF with the parameters of a Matérn covariance structure (marginal variance, $\sigma^2$, and spatial range, $r$, introduced in Subsect. 3.1):

$$p(\mathbf{T}, \boldsymbol{M}, \sigma_\varepsilon^2, \boldsymbol{\psi}) = \left[\prod_{j=1}^{m} p(T_j)\, p(M_j \mid T_j)\right] p(\sigma_\varepsilon^2)\, p(\boldsymbol{\psi})\,,$$

where the priors for $(\mathbf{T}, \boldsymbol{M}, \sigma_\varepsilon^2)$ follow the standard specification of Chipman et al. (2010) [2], as described in Subsect. 3.2. In particular, tree structures are regularized to act as weak learners, terminal node values are normally distributed, and the residual variance has a scaled inverse-gamma prior [7].

Penalized complexity (PC) priors are assigned to the spatial hyperparameters $\boldsymbol{\psi}$ of the spatial random effects $u_i$ to provide weakly informative, interpretable priors that control model complexity while allowing sufficient flexibility [7].

Computationally, direct MCMC sampling of the high-dimensional spatial random field is prohibitive. To address this, BARTSIMP uses an INLA-within-MCMC strategy, in which the spatial random effects are not sampled explicitly, but instead integrated out through an INLA-based approximation of the marginal likelihood [7]. This approximation is then used within a Metropolis-within-Gibbs MCMC algorithm to update the remaining model parameters, including the BART tree structures and their associated parameters. By avoiding the need to directly sample the spatial field, the approach becomes computationally more efficient while still enabling coherent Bayesian inference for the joint model.

In this work, the BARTSIMP model was fitted using the `BARTSIMP` package [7] in the `R` software. The number of trees considered was the same as in the BART model. To interpret the fitted model, PDP were also used. Default priors were adopted for model fitting. For the hyperparameters of the spatial random effects, the same PC-priors as in the BGM model were used. A total of $N = 4000$ MCMC iterations were performed, of which the first $B = 2000$ were discarded as burn-in, while the remaining $N - B = 2000$ iterations were retained for posterior inference.

### 3.5 Evaluation of the Models

Model performance was evaluated using spatial cross-validation to account for spatial autocorrelation, since spatial data violate the independence assumption of traditional ML validation methods, such as $K$-fold cross-validation. Ignoring spatial dependence can lead to overly optimistic accuracy estimates and an increased risk of over-fitting [17].

The spatial cross-validation method considered in this study was $K$-fold spatial block cross-validation [14]. This method divides the dataset into $K$ distinct, spatially separated folds, ensuring that training and testing data are not spatially autocorrelated. The folds were created by applying the $K$-means clustering algorithm [11] to the coordinates of the response variable. In this study, we set $K = 5$.

Model predictive performance was quantified using several metrics. The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used to measure overall prediction accuracy by quantifying the magnitude of prediction errors. The Nash–Sutcliffe Efficiency (NSE) was used to evaluate the proportion of the data variability explained by the model in terms of Mean Square Error. The 95% Coverage Probability ($CP_{95\%}$) metric was used to assess the predictive uncertainty of the model. This metric gives the proportion of observed values that fall within the 95% prediction interval.

Figure 2 shows the spatial distribution of training and testing locations for each fold generated using the 5-fold spatial cross-validation approach.

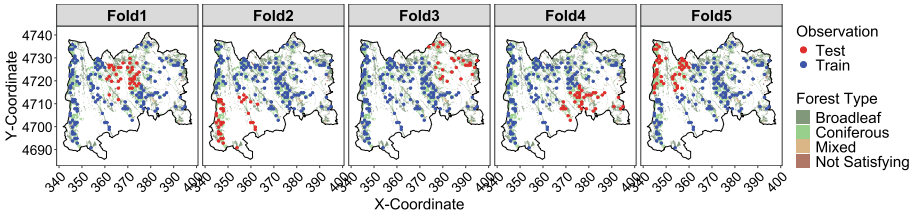



**Fig. 2.** Spatial distribution of training and testing locations using 5-fold spatial cross-validation.

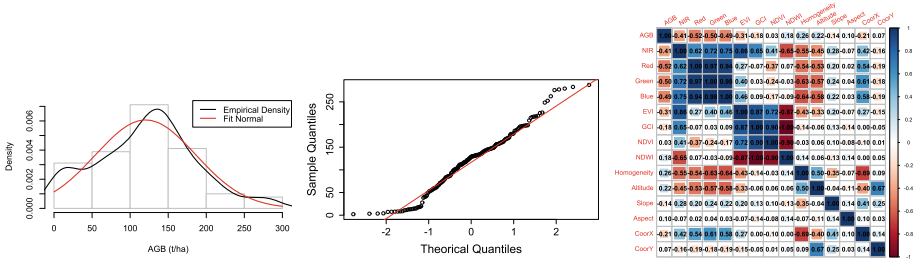
## 4 Results

In this Section, the results are presented. First, an explanatory analysis of the dataset used in this study is conducted, Subsect. 4.1. Subsections 4.2 and 4.3 present the predictive results obtained from the models described in Sect. 3 and the effects of each covariate in the considered models, respectively.

### 4.1 Explanatory Analysis

This Subsection begins with an assessment of the response variable distribution (AGB), an evaluation of multicollinearity among the covariates introduced in Sect. 2 and an analysis of residual spatial autocorrelation.

Figure 3 **a)** depicts the histogram along with the empirical distribution of AGB. Figure 3 **b)** shows a Normal Quantile–Quantile (Q–Q) plot, to evaluate the normality assumption. In general, the empirical quantiles closely follow the theoretical quantiles of the Normal distribution, suggesting that the normal approximation is reasonably appropriate, although small departures are noticeable in the distribution tails. Figure 3 **c)** presents the Spearman Correlation matrix between the response variable and the covariates, where high correlations can be observed among the reflectance bands as well as among the vegetation indices.




**a)** Histogram. **b)** Q-Q Plot. **c)** Spearman Correlation Matrix.

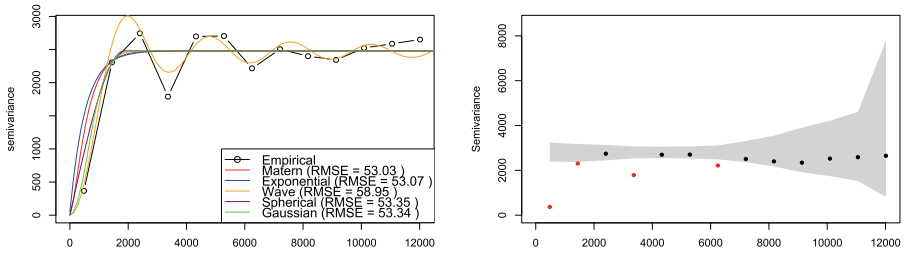
**Fig. 3.** Exploratory analysis: Histogram of AGB distribution (left), Q–Q plot for normality assessment (center), and Spearman correlation matrix (right).

To evaluate the normality of the response variable, a Kolmogorov-Smirnov test was done (p-value of 0.13). Thus, at a 10% significant level, it is reasonable to assume that the AGB follows a Normal distribution.

Due to the presence of high pairwise correlations among several variables, particularly the reflectance bands and vegetation indices, multicollinearity may be present, which can affect the performance of regression models. To address this issue, the Variance Inflation Factor (VIF) method [3], was applied. In this study, VIF was calculated iteratively using the `R` package `car`, with the covariate exhibiting the highest VIF removed at each step until all remaining covariates had acceptable VIF (below 10). The covariates retained were: the Blue reflectance band; NDVI and GCI vegetation indices; Homogeneity; Forest Type; and the topographic variables Altitude, Slope, and Aspect. These variables were included in all subsequent models.

To evaluate spatial autocorrelation in dataset, the empirical residual semivariogram from a linear regression model was analysed, along with its 95% Monte Carlo envelopes. Figure 4 **a)**, shows the empirical residual semivariogram

together with several theoretical models fitted to the semivariogram, including the Matérn, Exponential, Wave, Gaussian, and Spherical models. Figure 4 **b)** presents the empirical semivariogram together with its 95% Monte Carlo envelopes. From here, it can be concluded that residual spatial autocorrelation exists, since several points of the empirical semivariogram lie outside the 95% Monte Carlo envelopes.




**a)** Semivariogram. **b)** Envelopes.

**Fig. 4.** Empirical residual semivariogram and its 95% Monte Carlo envelopes: on the left, the semivariogram with fitted theoretical models (Matérn, Exponential, Wave, Gaussian, and Spherical); on the right, the empirical semivariogram with its 95% Monte Carlo envelopes.

In this study, the Matérn covariance function was chosen to model the spatial dependence structure because it provided the lowest RMSE when fitted to the empirical semivariogram, as shown in Fig. 4 **a)**, with RMSE values obtained through leave-one-out cross-validation procedure.

### 4.2 Predictive Results

To evaluate and compare the predictive performance of the all models considered, a 5-fold spatial block cross-validation was used. Table 1 summarizes the predictive results obtained for each of the models considered, reporting the RMSE, MAE, and NSE metrics, as well as the 95% prediction interval coverage metric ($CP_{95\%}$).

Analysing the results in Table 1, it can be observed that the BGM model achieves the lowest RMSE (23.09 t/ha) and MAE (8.45 t/ha), indicating superior performance in terms of average prediction error. Moreover, BGM also attains the highest NSE value (0.88), outperforming all other models according to this metric. Among the models based on BART, HBART shows the best performance (NSE = 0.79), followed closely by BART and BARTSIMP (both with NSE = 0.78). Regarding $CP_{95\%}$, values close to 95% indicate well-calibrated prediction intervals. HBART achieves the highest coverage (92.20%), suggesting better uncertainty representation. BART (89.03%) and BGM (88.89%) show similar performance, while BARTSIMP (76.11%) exhibits a substantially lower value, indicating a tendency to underestimate predictive uncertainty.

**Table 1.** Predictive results of all 4 models: BGM: Bayesian Geostatistical Model; BART: Bayesian Additive Regression Trees; HBART: Hybrid BART; and BARTSIMP: BART for Spatial INLA Modelling and Prediction.

| | Model | | | |
|---|---|---|---|---|
| Metric | BGM | BART | HBART | BARTSIMP |
| RMSE | 23.09 t/ha | 58.19 t/ha | 58.25 t/ha | 58.06 t/ha |
| $CP_{95\%}$ | 88.89% | 89.03% | 92.20% | 76.11% |
| MAE | 8.45 t/ha | 43.69 t/ha | 43.72 t/ha | 43.96 t/ha |
| NSE | 0.88 | 0.78 | 0.79 | 0.78 |

**Predictive Maps.** Figure 5 shows the predictive maps for each of the four models considered in this study.

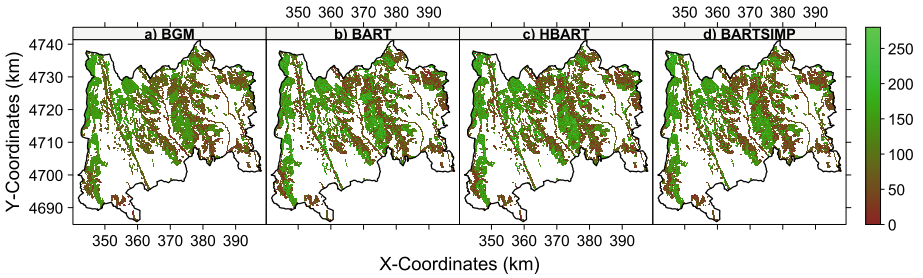



**Fig. 5.** Prediction maps of AGB(t/ha) in *Páramos-Valles*, considering the 4 models: BGM: Bayesian Geostatistical Model; BART: Bayesian Additive Regression Tree; HBART: Hybrid BART; and BARTSIMP: BART for Spatial INLA Modelling and Prediction.

Analysing the predictive maps AGB in *Páramos-Valles* (Fig. 5), it can be seen that all models show similar spatial patterns, with higher AGB values mainly in the north-eastern and central areas of the study region.

Figure 6 presents the mean GF estimated by the BGM and BARTSIMP models (Fig. 6 **a)** and Fig. 6 **c)**, respectively), and Fig. 6 **b)** shows the residuals of the BART model interpolated using the BGM model.

Analysing Fig. 6, it can be observed that all models exhibit clear and coherent spatial structures, indicating the presence of a well-defined spatial effect across the study region. The main differences between the maps are related to the scale and smoothness of the estimated effects rather than to their spatial configuration. The BGM model shows more pronounced spatial contrasts, with higher magnitude variations across space. In contrast, the HBART and BARTSIMP models are very similar in terms of the identified spatial pattern, differing mainly in scale, with both producing more attenuated and smoother representations of the spatial effect compared to BGM.

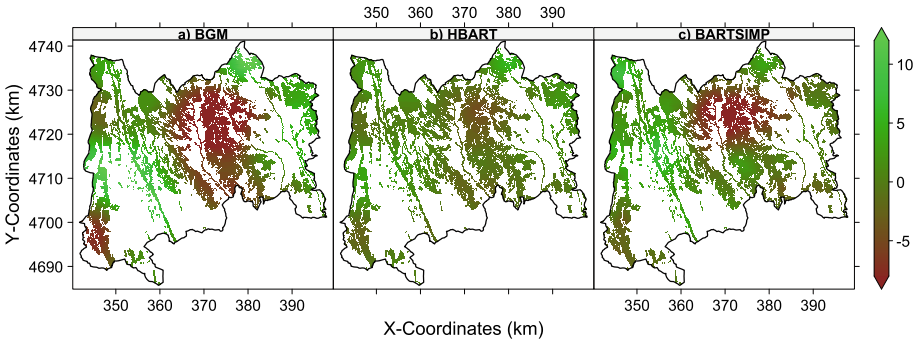



**Fig. 6.** Maps of spatial effect, considering the 3 models: BGM: Bayesian Geostatistical Model; BART: Bayesian Additive Regression Tree; and BARTSIMP: BART for Spatial INLA Modelling and Prediction.

### 4.3 Interpretation

**BGM:** Table 2 presents the posterior mean, standard deviation, and the corresponding Highest Posterior Density (HPD) intervals for the parameters and hyperparameters of the Bayesian Geostatistical Model. From Table 2, one observes that only covariates Aspect, GCI, and the two categories of the Forest Type variable (Broadleaf and Mixed) are statistically significant at the 5% level, as their HPD intervals do not include zero.

**Table 2.** Summary of the estimates of the parameters and hyperparameters of the BGM.

| | | | HDP | | | | | HDP | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | Mean | SD | Low | High | Variable | Mean | SD | Low | High |
| Blue | -0.00 | 0.04 | -0.08 | 0.08 | GCI | 23.96 | 10.31 | 3.70 | 44.11 |
| NDVI | 15.00 | 30.47 | -44.72 | 74.63 | Homogeneity | 4.33 | 25.98 | -46.63 | 55.18 |
| Altitude | 0.07 | 0.05 | -0.03 | 0.16 | Slope | 0.20 | 0.75 | –1.28 | 1.67 |
| Aspect | 0.11 | 0.04 | 0.03 | 0.19 | Broadleaf | –83.20 | 12.13 | –106.97 | –59.41 |
| Mixed | -27.74 | 13.60 | –54.40 | –1.08 | | | | | |
| Parameters of the Covariance Matrix | | | | | | | | | |
| r | 16454.13 | | 7949.40 | 25904.23 | $\sigma^2$ | 305.28 | | 37.81 | 782.33 |

Considering the four statistically significant covariates, the results indicate that Aspect and GCI are positively associated with AGB, whereas the Broadleaf and Mixed forest types show lower AGB values compared with Coniferous forests. Specifically, a one-unit increase in Aspect is associated with an average increase of 0.11 t/ha in AGB, while a one-unit increase in GCI corresponds to an average

increase of 23.96 t/ha, holding the remaining covariates constant. In contrast, Mixed forests present, on average, 27.74 t/ha lower AGB than Coniferous forests, whereas Broadleaf forests show an average decrease of 83.20 t/ha. Furthermore, the spatial component of the model revealed significant spatial dependence in AGB, with spatial correlation persisting up to approximately 16.5 km. The estimated marginal variance parameter also indicates considerable residual spatial variability not explained by the covariates included in the model.

**BART Model:** Figure 7 presents the Partial Dependence Plots (PDP) for the continuous covariates obtained from the BART model. These plots allow the evaluation of the marginal effect of each covariate on the predicted AGB values. Covariate NDVI shows a linear positive association with AGB, whereas covariates Blue and Slope exhibit a negative linear association with AGB (mild for the latter). In contrast, covariates CGI, Homogeneity, Altitude and Aspect display non-linear patterns, suggesting that their effects on AGB vary across different ranges of the covariates.

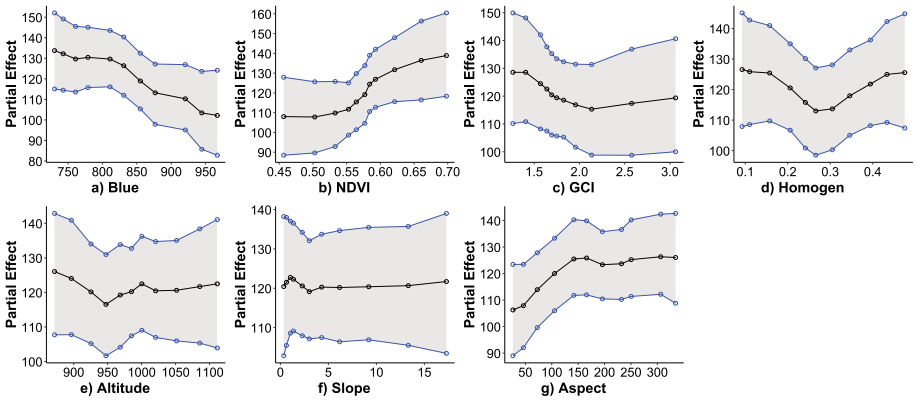



**Fig. 7.** Partial Dependence Plot (BART Model).

**BARTSIMP Model:** Figure 8 presents the PDP for the continuous covariates obtained from the BARTSIMP model. By inspecting the figure, it can be seen that the PDP are identical to those obtained from the BART model (Fig. 7), with the same interpretation.

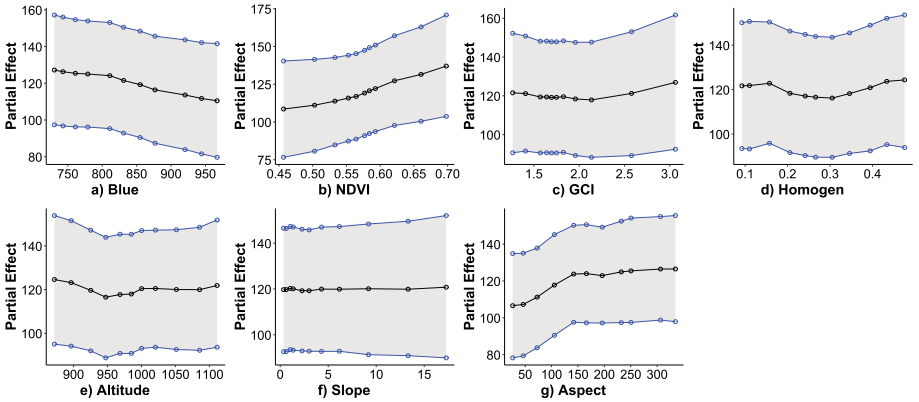



**Fig. 8.** Partial Dependence Plot (BARTSIMP Model).

## 5 Discussion and Conclusions

The findings of this study demonstrate discrepancies between the performances of the Bayesian ML models and the geostatistical model, particularly regarding predictive accuracy and uncertainty quantification. The BGM model demonstrated the best overall performance in terms of point prediction accuracy. In addition, BGM achieved the highest NSE value, indicating a superior ability to explain the variability of AGB compared to all other models considered.

In contrast, the models based on BART (BART, HBART, and BARTSIMP) produced noticeably higher RMSE and MAE values, indicating lower accuracy in point predictions. Among them, HBART demonstrated slightly superior performance, followed closely by BART and BARTSIMP. Even so, none of the models based on BART outperformed the BGM in any of the predictive accuracy metrics.

With respect to uncertainty quantification, the results were more nuanced. The HBART model demonstrated the highest coverage probability, approaching the nominal level and indicating well-calibrated prediction intervals. The BGM and BART models also provided reasonably adequate coverage, albeit still below the nominal level, suggesting a mild underestimation of uncertainty. In contrast, the BARTSIMP model demonstrated a markedly lower coverage, clearly indicating poorly calibrated predictive intervals and a systematic underestimation of uncertainty. The superior performance of the HBART model with respect to this metric may be related to its hybrid formulation, which combines the flexibility of BART with an explicit treatment of spatial structure in the residuals. This dual representation of uncertainty appears to improve interval calibration. However, this gain was accompanied by a decline in predictive accuracy, highlighting a trade-off between accurate point predictions and reliable uncertainty quantification.

Overall, these findings are partially aligned with those reported by Jiang et al. (2025) [7]. While the superior performance of the geostatistical model in terms of RMSE and MAE is consistent with their results, the advantage of the models based on BART in uncertainty quantification is not fully supported in this study. In the present study, only HBART demonstrated improved coverage, whereas BART and BARTSIMP failed to provide well-calibrated uncertainty estimates. These discrepancies may be attributed to dataset-specific characteristics, particularly the relatively small sample size and the extent to which the available covariates already explain spatial variability. In such circumstances, the advantages offered by highly flexible ML models may be restricted, while explicitly spatial approaches maintain greater robustness and efficiency.

Analysis of the predictive maps showed that all models exhibited analogous spatial patterns of AGB, with higher values concentrated in the northeastern and central regions of the study area. However, the analysis of spatial effects revealed some differences across models. The BGM model exhibited more pronounced spatial contrasts, whereas the HBART and BARTSIMP models were very similar in terms of the underlying spatial pattern, differing mainly in scale, with both producing smoother and more attenuated spatial effects. This finding indicates that the spatial component is more pronounced in the BGM model, while the models based on BART yield a more regularised representation. The interpolation of BART residuals using the BGM model further indicates that the BART model does not explicitly account for spatial dependence, thereby reinforcing the relevance of the hybrid approach to adequately represent this structure.

In terms of computational time, the models based on BART were substantially more demanding than the BGM model. The BGM model completed computations in approximately 15 min, whereas BART and HBART required around 30 min, and BARTSIMP approximately two days, which may represent a practical limitation in large datasets or when computational efficiency is a key concern.

The interpretation of covariate effects differs between the approaches considered. The BGM model identified only a few significant covariates (Aspect, GCI, and Forest Type), whereas the models based on BART exhibited more complex, non-linear relationships. Despite these, both approaches consistently identified Aspect and GCI as the most relevant variables, with a positive effect on AGB.

The results show that no single model dominates across all evaluation criteria. The BGM model demonstrated the highest predictive accuracy, while the HBART model provided the most reliable uncertainty quantification. The findings of this study indicate that hybrid approaches hold considerable promise, as they integrate the strengths of ML and geostatistical modelling. Therefore, the choice of model should be guided by the primary objective of the analysis, whether prioritising point prediction accuracy or the quality of uncertainty estimates.

**Acknowledgments.** This work is funded by national funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UID/00297/2025 (https://doi.org/10.54499/UID/00297/2025) and

UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025) (Center for Mathematics and Applications). National funding by FCT - Fundação para a Ciência e Tecnologia, I.P., through the individual research grant 2023.01166.BDANA (https://doi.org/10.54499/2023.01166.BDANA) of Ricardo Coelho.

## References

1. Bakka, H.: What is needed for a good mesh? (2017). https://haakonbakkagit.github.io/btopic104.html#31_What_is_needed_for_a_good_mesh, Accessed 7 January 2026
2. Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian Additive Regression Trees. Annals Appli. Stat. **4**(1), 266–298 (2010). https://doi.org/10.1214/09-AOAS285
3. Coelho, R., Natário, I., Fraile, S.: Bayesian geostatistical models for predicting above-ground biomass from remote sensing data. Remote Sensing Appli. Soc. Environ. **41**, 101855 (2026). https://doi.org/10.1016/j.rsase.2025.101855
4. Diggle, P.J., Giorgi, E.: Model-based geostatistics for global public health: methods and applications. Chapman & Hall/CRC (2019)
5. Diggle, P., Ribeiro, P.: Model-based Geostatistics. Springer Series in Statistics. Springer (Mar 2007)
6. Du, H., et al.: Spatial heterogeneity and carbon contribution of aboveground biomass of moso bamboo by using geostatistical theory. Plant Ecol. **207**, 131–139 (2010). https://doi.org/10.1007/s11258-009-9659-3
7. Jiang, A.Z., Wakefield, J.: BARTSIMP: flexible spatial covariate modeling and prediction using Bayesian Additive Regression Trees. Spatial Spatio-Temporal Epidemiol. **55**, 100757 (2025). https://doi.org/10.1016/j.sste.2025.100757
8. Kapelner, A., Bleich, J.: bartMachine: machine learning with bayesian additive regression Trees. J. Stat. Softw. **70**(4), 1–40 (2016). https://doi.org/10.18637/jss.v070.i04
9. Krainski, E., Gómez-Rubio, V., et al.: Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. Chapman and Hall/CRC (Dec 2018). https://doi.org/10.1201/9780429031892
10. Li, Y., Li, M., Li, C., Liu, Z.: Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Sci. Rep. **10**(1), 9952 (2020). https://doi.org/10.1038/s41598-020-67024-3
11. Likas, A., Vlassis, N., J. Verbeek, J.: The global k-means clustering algorithm. Pattern Recog. **36**(2), 451–461 (2003). https://doi.org/10.1016/S0031-3203(02)00060-2, biometrics
12. Moraga, P., Dean, C., Inoue, J., Morawiecki, P., Noureen, S.R., Wang, F.: Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: a case study predicting malaria risk in Mozambique. Spatial Spatio-Temporal Epidemiol. **39**, 100440 (2021). https://doi.org/10.1016/j.sste.2021.100440
13. National Geographic Institute (IGN): QSM - CNIG: National center for geographic information (2025). https://www.ign.es/web/qsm-cnig Accessed 12 December 2025
14. Roberts, D.R., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017). https://doi.org/10.1111/ecog.02881

15. Rue, H., Martino, S., Chopin, N.: Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. J. Royal Stat. Soc. Ser. B (Stat. Methodol.) **71**(2), 319–392 (2009). https://doi.org/10.1111/j.1467-9868.2008.00700.x
16. Tan, Y.V., Roy, J.: Bayesian additive regression trees and the General BART model. Stat. Med. **38**(25), 5048–5069 (2019)
17. Turton, A.E., Augustin, N.H., Mitchard, E.T.A.: Improving estimates and change detection of forest above-ground biomass using statistical methods. Remote Sensing **14**(19) (2022). https://doi.org/10.3390/rs14194911
18. Vashum, K.T., Jayakumar, S.: Methods to estimate above-ground biomass and carbon stock in natural forests-a review. J. Ecosyst. Ecography **2**(4), 1–7 (2012)
19. Wang, J., et al.: Remote sensing estimation of bamboo forest aboveground biomass based on geographically weighted regression. Remote Sensing **13**(15) (2021). https://doi.org/10.3390/rs13152962
20. Yankey, O., et al.: Disaggregating census data for population mapping using a Bayesian Additive Regression Tree model. Appl. Geogr. **172**, 103416 (2024). https://doi.org/10.1016/j.apgeog.2024.103416
21. Zhang, L., et al.: Mapping topsoil total nitrogen using random forest and modified regression kriging in agricultural areas of central China. Plants **12**(7) (2023). https://doi.org/10.3390/plants12071464
22. Zhang, T., Geng, G., Liu, Y., Chang, H.H.: Application of bayesian additive regression trees for estimating daily concentrations of PM2.5 components. Atmosphere **11**(11) (2020). https://doi.org/10.3390/atmos11111233