Ann. For. Res. 65(1): 47-63, 2022
https://doi.org/10.15287/afr.2022.2064

ANNALS OF FOREST RESEARCH
www.afrjournal.org

Simulated annealing in feature selection approach for modelling aboveground carbon stock at the transition between Brazilian Savanna and Atlantic Forest biomes

Laís Almeida Araújo[1], Isáira Leite e Lopes[1], Rafael Menali Oliveira[1], Sérgio Henrique Godinho Silva[2], Carolina Souza Jarochinski e Silva[1], Lucas Rezende Gomide[1]✉

Araújo L.A., Lopes I.L. e, Oliveira R.M., Silva S.H.G., Silva C.S.J. e, Gomide L.R., 2022. Simulated annealing in feature selection approach for modeling aboveground carbon stock at the transition between Brazilian Savanna and Atlantic Forest biomes. Ann. For. Res. 65(1): 47-63.

Abstract Forest ecosystems are important in the carbon storage process. Thus, the objective was to investigate the effectiveness of the Simulated Annealing metaheuristic analysis for selecting variables to maximize the accuracy of the aboveground carbon prediction at the tree level. We used data from uneven-aged forests located in the Rio Grande Basin - Minas Gerais, Brazil, where 227 trees had their carbon stock measured. The classic Spurr linear model, stepwise linear regression and pantropical coverage, Random Forest (RF), and the hybrid SARF method (Simulated Annealing and Random Forest) were used to estimate the carbon stock from the selection of variables for the different compartments of the tree (total, stem, branch, and leaf). The SARF consisted of the metaheuristic to select the variables to be used in the RF. These methods were evaluated by the root mean square error (RMSE), coefficient of determination (R^2), and residual graph. As a result, the pan-tropical equation demonstrated superior performance than the Spurr model due to its greater homogeneity of residues. The stepwise technique reduced the number of variables and the error of the estimates, mainly for the validation set. SARF showed better adjustments than RF, as it reduced in on average 99.2% of the number of variables and 9% of the error of estimates considering all compartments. In general, variables such as volume, basic wood density, canopy projection area, diameter at 0%, diameter at breast height, height, and latitude contributed strongly to the carbon independent of the tree compartment. Among the methods, SARF is an alternative to the traditional method, as it can extract accurate information from a large data set.

Keywords: data mining, biomass, simulated annealing, random forest.

Addresses: [1]Department of Forest Sciences, Federal University of Lavras, Brazil| [2]Department of Soil Science, Federal University of Lavras, Brazil.

✉ **Corresponding Author:** Lucas Rezende Gomide (lucasgomide@ufla.br).

Manuscript: received October 14, 2020; revised January 28, 2022; accepted March 06, 2022.

Introduction

Tropical forest ecosystems are an important pool of carbon sink, which partially regulates the exchanges flux of atmospheric CO_2 (Baker et al. 2010, Pechanec et al. 2018). However, the changes of tropical land use impact the biodiversity and this natural carbon cycle (Mendoza-Ponce et al. 2018). Deforestation and forest degradation contribute with 15% to 20% of global carbon emissions and tropical areas are highly correlated to this percentage (Vicharnakorn et al. 2014). In this context, the prediction of biomass and carbon stock of remnant forests is crucial to understand the global carbon cycle and possible actions to mitigate climate change (Heinrich et al. 2021, Chinembiri et al. 2013, Vicharnakorn et al. 2014). However, there is a lack of studies describing the direct factors affecting the pool of carbon in the trees.

An important condition for obtaining better estimates is the inclusion of appropriate structural variables that influence biomass (Goodman et al. 2014). Since diameter at breast height is often insufficient to produce good estimates (Guangyi et al. 2017), many authors have incorporated other variables in the models in order to increase their predictive capacity. Some examples are the variables related to the dimensions and architecture of the canopy (Goodman et al. 2014, Larsary et al. 2021), height (Feldpausch et al. 2012) and basic wood density (Chave et al. 2005). The use of these complementary variables can deliver more accurate estimates.

Carbon stock modeling at individual trees level has high contribution to forest science. The trees biomass is traditionally determined based on inventory data and allometric equations (Burrows et al. 2000). These equations assume the relation of the power-law between biomass and the diameter of the tree (Enquist & Niklas 2001). Based on this theory, several studies were made on allometric models considering the relationship between biomass and diameter (Zianis & Mencuccini 2004, Pilli et al. 2006, Návar 2009). Subsequently, the idea of a single explanatory variable (diameter) for biomass prediction was questioned, and the inclusion of other variables such as height and basic density were introduced (Chave et al. 2005, Vieilledent et al. 2012). Even so, several authors have studied the importance of other factors that influence the prediction of biomass (Burrows et al. 2000, Kuyah et al. 2012, Chave et al. 2014), mainly for tropical forests that represent a significant proportion of forests at global scope (Siddiq et al. 2021).

Historically, regression models have been well suited to predict many forest variables so far. This technique is widespread due to high accuracy and quality of estimations. Today, computational intelligence methods are recognized to overcome some limitations of the regression models (Drake et al. 2006, Were et al. 2015). They achieve a greater generalization of estimates and less susceptibility of noisy and outliers (Nunes & Görgens 2016, Vieira et al. 2018, Ou et al. 2019). These advantages have encouraged the application of computational intelligence in modelling the carbon stock, such as artificial neural networks (Vahedi 2016, Dantas et al. 2021) and support vector machine (Gleason & Im 2012, Vafaei et al. 2018). Promising data mining techniques have been used, including genetic algorithm (Hong et al., 2018), random forest (Silveira et al. 2019), simulated annealing or hybridization of these (Mafarja & Mirjalili 2017), to extract only relevant information for increasing the model prediction efficiency (Bagherzadeh-Khiabani et al. 2016).

The hybrid method consists of improving the algorithms performance for solving complex problems, combining a series of strategies. By combining at least two techniques, an increase in the performance of each technique is obtained. Since the search for regions not yet explored is expanded and the search in regions with the best solutions is intensified (Mafarja & Mirjalili 2017). Hong et al. (2018)

demonstrated the potential of hybridization in improving the performance of the random forest and support vector machine when using the genetic algorithm to obtain the ideal combination of variables. Other studies have also observed satisfactory performance when using hybrid methods, as in the modeling susceptibility to forest fires (Bui et al. 2017), classification of remote sensing images (Wang et al. 2017) and prediction of height and volume of trees (Reddy et al. 2017).

In prediction models, resource selection is an important technique and is constantly used in data pre-processing. This technique reduces irrelevant and redundant information from a data set, increasing predictive performance and interpretation of results (Liu & Yu 2005). However, although the search for an optimal subset is challenging, meta-heuristics have produced good results in optimization problems (Mafarja & Mirjalili 2017). In the context of biomass and carbon stock estimation, the search for ideal models using machine learning can contribute to the quantification of these attributes (Pham et al. 2020). Thus, more accurate estimates can guarantee and support effective mitigation actions on climate change (Scolforo et al. 2015, Vahedi 2016).

At this point, variables selection methods using computational intelligence have advantages for modeling any processes since they increase predictive performance, avoid overfitting and have higher computational speed (Day et al. 2020), when there is a huge sample space of possibilities and combination between variables. Recent studies have also focused on identifying and interpreting the variables that show the greatest influence on the behavior of the carbon stock in different tree compartments, such as stem, branch, leaves, and total. The main premise of the study is the approach of automatic selection of variables by the hybrid method, identifying advantages and disadvantages in its use. This may be particularly important when you have large amounts of data and need to turn it into useful information. In this scenario, data mining is an essential element in the knowledge discovery process, which consists of an iterative sequence of data pre-processing, data mining, pattern assessment and the presentation of knowledge (Han et al. 2011, Mafarja & Mirjalili 2017).

In fact, the main challenge relies on adding only variables with high potential to explain the carbon stock. This combinatorial problem explores high dimensional data with complex pattern and scales. Multivariate analysis techniques are well suited to reduce a data set to a lower level (Sun et al. 2014, Labani et al. 2018). However, using computational intelligence it is possible to work with large-scale and non-linear data (Anifowose et al. 2014). Hence, the main objective is investigating the effectiveness of the Simulated Annealing meta-heuristic for variables selection to maximize the accuracy of aboveground carbon prediction at tree level. Here we compare the regression model analysis and machine learning methods including the trees components of stem, branch, leaf and total carbon as variables. Finally, we also evaluate the pan-tropical equation (Chave et al. 2005) precision with our observed filed data.

Material and Methods

Study sites and tree sampling

The study site comprises the Rio Grande watershed located in the south of the state of Minas Gerais – Brazil with 86,110 km² or 14.7% of the state (Figure 1). Altitudes range from approximately 300 to 2,700 m. According to the climatic classification, the site has some classes as Humid B2, Humid B3, Humid B4 and Super Humid based on the Thornthwaite Moisture Index. The annual average temperatures are between 14ºC and 20ºC, and the annual average precipitation was found above 1,500 mm (Carvalho et al. 2008). The Rio Grande watershed is located in a transition area between two biomes, Cerrado (Brazilian savanna) and Atlantic

Forest, with high anthropization process. The characteristics of the predominant vegetation are Cerrado, Evergreen moist forest and Semideciduous tropical forests. The main types of soils are Argisol, Cambisol, Neosol and Latosol (Carvalho et al. 2008).

Data

A field survey of trees was conducted for aboveground carbon analysis at individual trees level. The division of the tree into at least stem, branches, leaves and total is almost universal, which defines our studied components. Further, we selected the municipalities of Comendador Gomes (north), Campanha (center) and Passa Quatro (south) for trees sampling procedure. These sites are the mid-point of each main vegetation class and presents the mean of the watershed tree diversity. The trees sample was carried out selecting systematically trees of each characteristic site. The phytosociological analysis highlighted the species with the highest IVI% (importance value index) in Comendador Gomes (*Pterodon emarginatus*: 65.1, *Copaifera langsdorffii*: 44.3, *Xylopia aromatica*: 31.1, *Cenostigma macrophyllum*: 26.4 and *Guapira venosa*: 12.5); Campanha (*Calyptranthes clusiifolia*: 15.6, *Mollinedia widgrenii*: 14.2, *Machaerium nyctitans*: 13.9, *Casearia sylvestris*: 12.6 and *Piptadenia gonoacantha*: 12.5) and Passa Quatro (*Eremanthus erythropappus*: 26.2, *Daphnopsis utilis*: 13.7, *Miconia trianae*: 11.3, *Machaerium dimorphandrum*: 11.2 and *Dalbergia villosa*: 10.8). A total of 227 trees was inventoried ranging in diameter classes of the most dominant species in each site. Given the three floristic groups: Cerrado (north), Semideciduous tropical forest (center) and Evergreen moist forest (south) (Table 1).

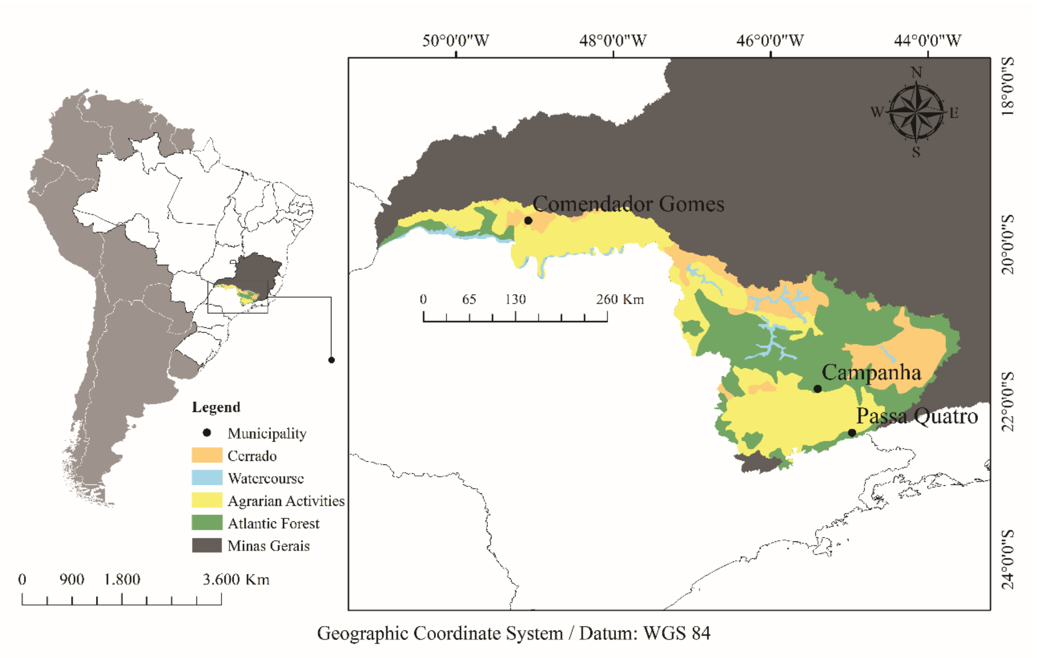

Figure 1 Location of the Rio Grande watershed in Minas Gerais state – Brazil.

Table 1 Diametric distribution of the individuals selected in the sampling for the rigorous cubing procedure and collection of other morphometric variables.

Diameter class	Floristic groups		
	Cerrado	Semideciduous tropical forest	Evergreen moist forest
5-10	10	15	10
10-15	9	10	9
15-20	10	10	12
20-25	9	10	10
25-30	11	10	11
30-35	8	9	10
35-40	7	4	8
>40	13	1	11
Total / group	77	69	81
Total		227	

The tree measure variables were *dbh* - diameter at breast height, *ht* - tree height, *cpa* - crown projection area, *hc* - commercial bole height at smallest merchantable diameter (3 cm), and *cbh* - crown base height. Field teams acquired latitude and longitude by GPS. Furthermore, the same portion of inventoried trees were submitted to rigorous cubing for quantifying volume and aboveground biomass of trees. To assist the wood density (*wd*) and carbon content analysis, field teams collected wood discs from different positions along tree stem (0, 25, 50, 75 and 100% of commercial height) and from branches (25 and 75% of length). Tree biomass is a function of wood volume, while carbon derives from biomass after laboratory analysis. This methodology followed the Food and Agricultural Organization of the United Nations - FAO (Picard et al. 2012).

Although allometric modeling has been used since last decade, the variables selection procedure is still the key point. In spite of the high variability of mathematical data transformation (logarithmic, inverse, square root, second and third powers) and combination (basic arithmetic operations multiplication and division), we defined 3 strategies of variable input usage for aboveground carbon modeling: S_1) *dbh* and *ht*, S_2) non-transformed variables (21 variables), and S_3) non-transformed, transformed and combination of variables (985 variables). In addition to the modeling tests and performance, we previously split the database randomly in two independent sets for training (80%) and validation (20%).

Modeling aboveground carbon strategies

A wide variety of techniques has been used to estimate trees attributes. In such case, regression models and computational intelligence techniques are well suitable for forest data. These techniques have been widespread and our objective were test their performance. Figure 2 presents a flowchart with the methods and strategies used to estimate aboveground carbon. The challenge of the study was to verify which strategy to follow to identify the method and the explanatory variables that produce better estimates of the aboveground carbon. In this sense, they were evaluated from the most usual variables (*dbh* and *ht*) through regression models to a high set of variables (including transformations and iterations between variables) using machine learning algorithms. Several studies have observed that machine learning algorithms outperform other methods such as stepwise regression, principal component regression and partial least squares regression (Mouazen et al. 2010, Guo et al. 2015, Wang et al. 2018).

Regarding the regression analysis, we take into account two strategies of modeling. At first, we associated the variables dbh and ht as inputs of the classical linear model (S_1). Additionally, we tested only pure variables (22) and pure together with transformed variables (132) within stepwise linear regression for selecting variables (S_2). We applied the pan-tropical coverage (Chave et al. 2005) only for modeling carbon total stock by using total dry aboveground biomass and 0.417 of carbon conversion factor. The coefficients of linear model were obtained by ordinary least squares (OLS) and multiple linear regression according the AIC criterion

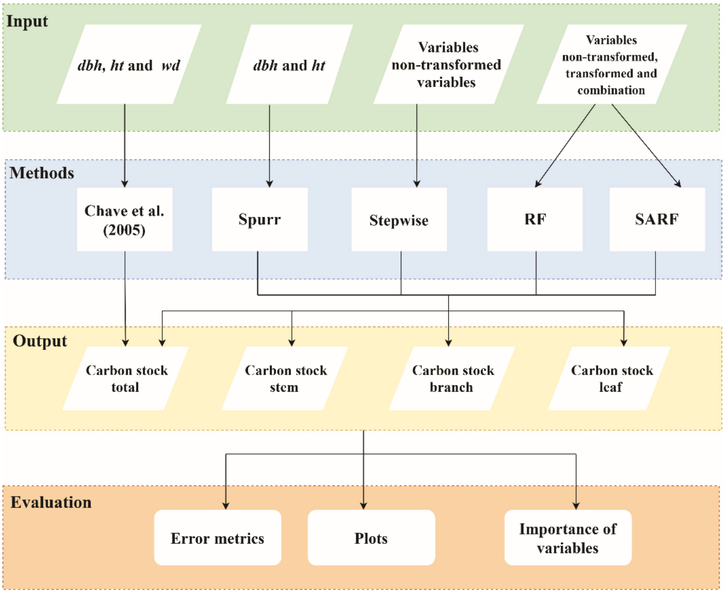

Figure 2 Methodological process with different strategies and methods. Where: RF: random forest; SARF: Simulated Annealing and random forest.

for stepwise model building. Therefore, the variance inflation factor (VIF<10) was adopted to avoid multicollinearity effects using R package car (Fox et al. 2019).

$$C = (0{,}0559*DBH^2*H_T*DB)*0{,}417 \qquad (1)$$

Random Forest (RF) algorithm (Breiman 2001) is a statistical method, non-parametric, very popular and effective for both regression and classification purposes, which can be attributed to its simple parameterization, high predictive performance and the ability to work with missing values, noise and high data dimension (Genuer et al. 2010, Hapfelmeier & Ulm 2014). In the context of the research, this algorithm was also chosen for aboveground carbon modeling and strategy S_3 of data. The tuning parameters were established from previous tests: ntrees: number of trees in 500 units; mtry: number of attributes to be chosen as 2; and, nodesize: number of observations at the end nodes of each tree as 5. The RF algorithm was performed 50 times and the best model was found. We implemented the RF in R package randomForest (Liaw & Wiener 2018).

The last method tested was a hybrid of the meta-heuristic Simulated Annealing (SA) and Random Forest algorithm (RF). Simulated Annealing (SA) is a stochastic local search algorithm that, from an initial randomly generated solution, iteratively searches the neighborhood of the current solution. The solution consists of a vector dimensioned according to the problem under study. The implementation of SA demands the selection of some control parameters: the initial temperature, the definition of the evaluation function, the cooling schedule, and the stopping criterion (Abbasi et al. 2011). In this study, the SA defines an optimized set of variables for RF run, composing a relatively new method for modeling data in several fields of science named SARF (Figure 3). Generally, this procedure is robust for variables selection, in huge database, extracting only relevant variables from the system. The negative impacts of large number of variables for random forest accuracy are well-known. Consequently, variable shrinkage methods are strongly requested for improving its performance. The SA only selects variables and the solution performance is quantified by RF errors. Actually, the variable selection technique is a framework for highlighting only relevant information (Kavzoglu & Mather 2002, Guyon & Elisseeff 2003).

The automatic variable selection considering artificial intelligence is possible after the union of these algorithms. They explore numerous alternatives of combinations for a well-fitted final model. The deep searching for the best

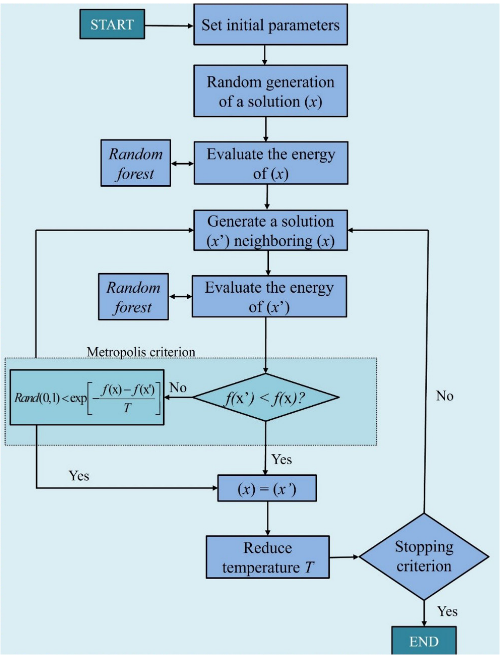
START
Set initial parameters
Random generation of a solution (x)
Random forest
Evaluate the energy of (x)
Generate a solution (x') neighboring (x)
Random forest
Evaluate the energy of (x')
Metropolis criterion
Rand(0,1) < exp[−(f(x)−f(x'))/T]
No
f(x') < f(x)?
No
Yes
Yes
(x) = (x')
Reduce temperature T
Stopping criterion
Yes
END

Figure 3 Methodological scheme of SARF algorithm.

subset reduces the algorithm efforts and escape from local optimum. The method does not test all possible combinations as brute force procedure. Therefore, the stop criterion is achieved under a certain number of iterations. In this context, the method assumptions seek a high-quality solution with low number of predictor variables. Consequently, the multi-objective optimization problem consists of two components: the first is associated with the high accuracy solution; and the second with the minimum number of variables. However, the component's weights are unbalanced due to distinct units and we standardized them to avoid such noises. The out-of-bag (OOB) is an error metric derive from RF performance, and the OOB_{max} is a utopic value from RF with all variables. The second part of the equation represents a ratio between the number of selected variables (n) and the total number of candidate variables tested (N). Unfortunately, the method limitation includes the selection of highly correlated variables and its stochastic procedure.

$$f(x) = \frac{OOB}{OOB_{max}} + \frac{n}{N} \qquad (2)$$

Initially, the SA tunning was set as initial (T_0 = 106) and final (T_F = 25) temperatures, 1% of the cooling rate, and the objective function was to minimize the mean square error and the number of variables (Equation 1). The SA algorithm starts with an initial solution (x) generated randomly, being represented by a dimensional vector with length according to the number of selected variables. At each iteration, a new candidate solution is explored with random values. Then a better value is allowed and horst will be accepted under a certain probability by Metropolis criteria. The current temperature is reduced over the iteration which affects the accepting probability (Mafarja & Mirjalili 2017). Considering any new solution, RF run by selected variable with SA returns the OOB value. Finally, we run 50 times the method due to stochastic algorithm whose outcome involves some randomness and uncertainty.

Methods performance analysis

The data processing was performed using R software (Version 3.5.3 - © 2019 RStudio, Inc.) with Intel (R) Core™ i3-2100, CPU @ 3.10 GHz processor, and memory of 8.0 GB (RAM). The quality of predictions can be defined as many metrics and indices. Therefore, according to the literature (Carreiras et al. 2012, Vahedi 2016, Corona-Núñez et al. 2017), the most recurrent and unbiased statistical metrics are root of the mean square error (RMSE) and coefficient of determination (R^2). We applied them for training and validation subsets analysis and formally they are defined as equations 3, 4 and 5, in which n: number of observations; i: sub-index of observations; Y: observed value of the carbon stock (kg), $\hat{Y}$: estimated value of the carbon stock (kg).

$$RMSE = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(Y_i - \hat{Y}_i)^2} \qquad (3)$$

$$\text{RMSE \%} = \frac{\text{RMSE}}{\bar{Y}} * 100 \qquad (4)$$

$$\text{RMSE} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2} \qquad (5)$$

Results

Carbon stock variability

The database presented a high heterogeneity of carbon stock values independent of the tree compartment (Table 2). This behavior is proven through the characteristics of the studied area, which exhibits a transition from the Cerrado to Atlantic Forest biomes. These biomes have regions ranging from high density forests and large trees to regions with low numbers of individuals and small trees. Therefore, we may attribute the high variation in carbon stock (more than 100%) to regional differences which encompass distinct forest species composition and structure. This characteristic of the study area indicates a difficulty in obtaining good estimates in the modeling of the carbon stock, in all evaluated compartments.

Carbon stock modeling

In this study, we compared all the tested methods according to their modeling accuracy of carbon stock and selection of variables. The latter showed significant improvements in the carbon stock estimates, regardless of the tree compartment. In this respect, the relevance of this process is notorious. Therefore, all variables selected for the new developed models (via stepwise) and the adjustment of the Spurr model are shown in Table 3. The VIF values for the models developed via stepwise varied between 1.04 and 7.37. The adjusted parameters for each model are shown in the table 3.

In general, according to statistical metrics for all equations tested (Table 4) is verified that the equations showed low performance to estimate the carbon stock of the compartments (kg tree^{-1}), not efficiently following the observed values. This result shows that only the use of *dbh* and *ht* cannot accurately estimate the carbon stock. The pan-tropical equation provided greater precision in estimates for the total carbon

Table 2 Analysis of the aboveground carbon stock (kg) in trees components and sites.

Comp	Statistics	Comendador Gomes (north)	Campanha (center)	Passa Quatro (south)
Total	Minimum (kg/tree)	2.63	3.38	3.96
	Mean (kg/tree)	198.59	93.22	174.76
	Maximum (kg/tree)	1005.23	755.28	787.73
	CV (%)	119.01	119.90	104.01
Stem	Minimum (kg/tree)	1.51	2.13	2.24
	Mean (kg/tree)	86.34	47.41	99.79
	Maximum (kg/tree)	376.54	406.67	652.67
	CV (%)	113.60	123.65	113.82
Branch	Minimum (kg/tree)	0.51	0.78	1.01
	Mean (kg/tree)	103.75	42.29	68.35
	Maximum (kg/tree)	658.37	331.73	290.99
	CV (%)	135.12	126.56	108.36
Leaf	Minimum (kg/tree)	0.29	0.14	0.25
	Mean (kg/tree)	8.50	3.52	6.63
	Maximum (kg/tree)	53.18	17.50	31.16
	CV (%)	100.78	112.11	100.98

Table 3 Equations adjusted by the linearized Spurr model and the stepwise method to estimate the total carbon stock, in the stem, branches and leaves.

Comp	Strategy S_1	*Strategy S_2
Total	$\ln(c) = -4.203 + 1.008 * \ln(dbh^2*ht)$	$c = -173.47 + 207.73 * vt + 277.96 * wd_t + 1.33*cpa + 1.13*d_{25}$
Stem	$\ln(c) = -4.299 + 0.942 * \ln(dbh^2*ht)$	$c = -114.86 + 229.27 * vt_{st} + 180.13 * wd_{st} + 2.01 * hc$
Branch	$\ln(c) = -6.360 + 1.143 * \ln(dbh^2*ht)$	$c = -94.19 + 219.92 * vt_b + 168.88 * wd_b + 0.82 * cpa$
Leaf	$\ln(c) = -4.228 + 0.649 * \ln(dbh^2*ht)$	$c = 130.20 + 0.14 * cpa + 4.01 * vt_b + 1.94 \times 10^{-5} * X + 1.85 \times 10^{-5} * Y - 0.97 * cbh + 0.70 * hc$

Where: S_1: *dbh* and *ht* input variable strategy; S_2: non-transformed input variable strategy; ln: natural logarithm; c: stock carbon; *dbh*: diameter at breast height; *ht*: total height; *vt*: total volume; wd_t: total wood basic density; *cpa*: crown projection area; d_{25}: diameter at 25% of the tree; vt_{st}: total volume of the stem; wd_{st}: stem basic density; *hc*: commercial bole height; vt_b: total volume of branches; wd_b: branches basic density; X: longitude: Y: latitude: *cbh*: crown base height. * all parameters were significant at 95% probability.

stock compared to the Spurr model. The Spurr model for the other compartments showed a high tendency to overestimate individuals with lower carbon stocks. The greatest performance of the equation suggested by Chave et al. (2005) can be attributed to the use of the basic density variable, which generally correlates with the carbon stock, improving the estimates.

Using the stepwise technique, we developed an equation for each tree compartment (leaf, branch, stem and total) in order to predict the carbon stock (kg) at the tree level. Furthermore, the objective was also to find the most explanatory variables on the behavior of the carbon stock in the trees. Thus, when looking at the statistics of the models generated for training and validation data (Table 4), it was found that the S2 strategy presented slightly more efficient models for estimating the carbon stock, mainly for validation. Regarding the selected variables, in general, it is observed that the volume and basic density of the tree compartments, and the crown projection area obtained a high explanatory capacity for all compartments.

Table 4 Statistics of the adjusted models to estimate the carbon stock (kg) for all compartments.

Comp	Data	Strategy	RMSE	RMSE%	R^2
Total	Training	S_1-regression	76.29	51.94	0.83
		Chave*	70.63	48.08	0.86
		S_2-regression	34.71	23.63	0.97
		RF	25.30	17.23	0.98
		SARF	21.24	14.46	0.99
	Validation	S_1-regression	88.42	57.87	0.81
		Chave*	90.61	59.31	0.80
		S_2-regression	37.84	24.77	0.96
		RF	67.17	43.96	0.89
		SARF	59.43	38.90	0.91
Stem	Training	S_1-regression	31.82	40.97	0.88
		S_2-regression	16.76	21.58	0.97
		RF	11.46	14.76	0.98
		SARF	10.79	13.89	0.99
	Validation	S_1-regression	82.35	94.01	0.48
		S_2-regression	22.31	25.46	0.96
		RF	67.24	76.76	0.65
		SARF	60.22	68.74	0.72
Branch	Training	S_1-regression	59.32	79.92	0.67
		S_2-regression	22.12	29.80	0.95
		RF	18.35	24.72	0.97
		SARF	16.80	22.63	0.97
	Validation	S_1-regression	52.69	71.76	0.71
		S_2-regression	14.40	19.61	0.98
		RF	28.65	39.03	0.92
		SARF	21.98	29.94	0.95
Leaf	Training	S_1-regression	5.62	88.12	0.39
		S_2-regression	3.69	57.80	0.74
		RF	2.30	35.97	0.90
		SARF	2.35	36.76	0.89
	Validation	S_1-regression	5.19	78.32	0.43
		S_2-regression	3.60	54.34	0.72
		RF	3.78	57.14	0.69
		SARF	4.13	62.43	0.63

Where: Comp: components; RMSE: root mean squared error; RMSE(%): root mean square percentage error; R^2: coefficient of determination; RF: Random Forest; SARF: Simulated Annealing with Random Forest; *only estimated values from pan-tropical equation Chave et al. (2005) plus carbon factor.

Importance of variables

The application of the RF algorithm in its pure form was used to estimate carbon stocks, to assess the contribution of each variable and, later, to make comparisons with the hybrid methodology. When analyzing the importance of the selected / used variables, it was found that volume and total basic density as well as volume and basic density of the stem showed greater importance values. In addition, among all compartments (leaf, branch, stem and total), transformed and combined variables were the most representative ones in the first positions.

The SARF selected optimized subsets with different variables' numbers according to each tree compartment (leaf - 7, branch - 8, stem - 9, and total - 8) (Figure 4). The greatest contributions are attributed to the diameter at 25% (d_{25}), the total volume of the branches (vt_b), the total volume of the stem squared (vt_{st^2}), and the commercial height times the total volume (hc * vt), for the respective compartments, leaf, branch, stem, and total carbon stock. Thus, it is observed that some variables have a higher frequency among the compartments, being the volume and the basic density both in its untransformed form and with the transformations and combinations with other variables. In addition, variables such as latitude, longitude, crown projection area, diameter at breast height, and diameter at 0%, height proved to be very representative in most of the analyzed compartments.

From these results, we noticed that the proposed methodology with the multiobjective approach, allowed the SA algorithm to satisfactorily reduce the number of variables (on average 99.2%) for the different compartments and to decrease the error (on average 9%) obtained by the prediction models (Table 3). The SARF presented the best results according to the RMSE (%) in relation to the other models for the training data, with the exception of the leaf compartment. SARF showed improvements in the estimates for the validation data but was inferior to the S_2 strategy. This result shows that only the variables *dbh* and *ht* cannot explain well the behavior of the carbon stock in the tree. For the R^2 values, the models presented similar responses for both training and validation data, except for the Spurr model. Although the RF showed a performance close to the SARF in the training set, the reduction in the number of variables provided a significant improvement in the estimates for most compartments of the tree.

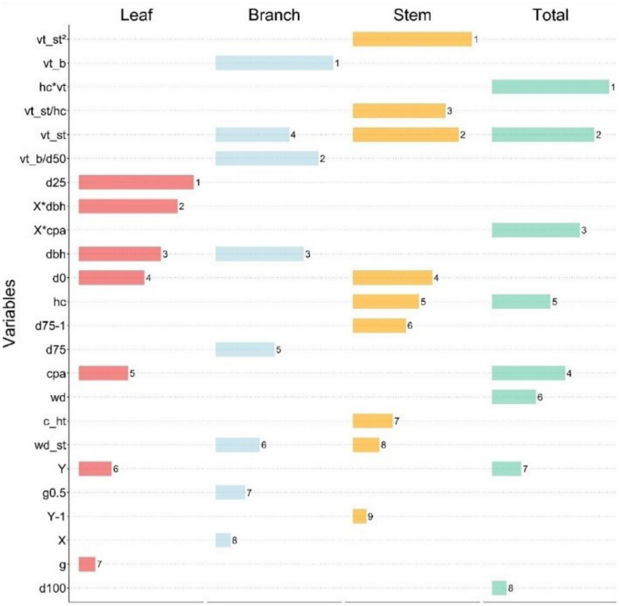

Figure 4 Ranking of selected variables for each tree compartment. Where: d_{25}: diameter at 25%, X * *dbh*: longitude times diameter at breast height, *dbh*: diameter at breast height, d_0: diameter at 0%, *cpa*: canopy projection area, Y: latitude, *g*: sectional area, *vt_b*: the total volume of the branches, vt_b/d_{50}: the ratio of the total volume of the branches and diameter at 50%, *vt_st*: total volume of the stem, d_{75}: diameter at 75%, *wd_st*: basic stem density, $g^{0.5}$: root of sectional area, X: longitude, vt_st^2: the total volume of the stem squared, *vt_st/hc*: the ratio of the total volume of the stem and commercial height, *hc*: commercial height, d_{75}^{-1}: the inverse of the diameter at 75%, *c_ht*: total height class, Y-1: the inverse of latitude, *hc* * *vt*: the commercial height times the total volume, X * *cpa*: longitude times the canopy projection area, *wd*: total basic density, d_{100}: diameter at 100%.

When analyzing the residual dispersion graphs, it was observed that all methods, including the model suggested by Chave et al. (2005), showed a tendency to overestimate individuals with a lower carbon stock, both for the training set and for validation (Figure 5). Among the methods, the RF and SARF exhibited a more homogeneous residual distribution for the compartments, except the leaf compartment.

The selection methods, stepwise and SARF, did not select the same set of variables for each compartment, however, some variables were the same: the total and stem basic density, crown projection area, total stem and branch volume, and commercial height. Most of these variables have high correlation values (Figure 6). This figure also shows the values of Pearson's linear correlations for the other selected variables, in which a similar behavior is observed between the correlation of the variables and the compartments. When comparing the models, S1 and SARF, for each compartment it appears that the modeling of the total carbon stock stood out in relation to the other compartments due to the more homogeneous distribution. In addition, the SARF method performs well in the selection of explanatory variables and in the estimation of carbon stock. Therefore, these results justify the application of the SARF to estimate the carbon stock, in addition to contributing to the understanding of the variables that most influence the carbon stock. In the context of the selection of explanatory variables, the analyzed techniques found a relatively similar set, however the methods differed more in the results of the estimates between the set of training and validation.

Discussion

Carbon stock estimates present in native forests are important

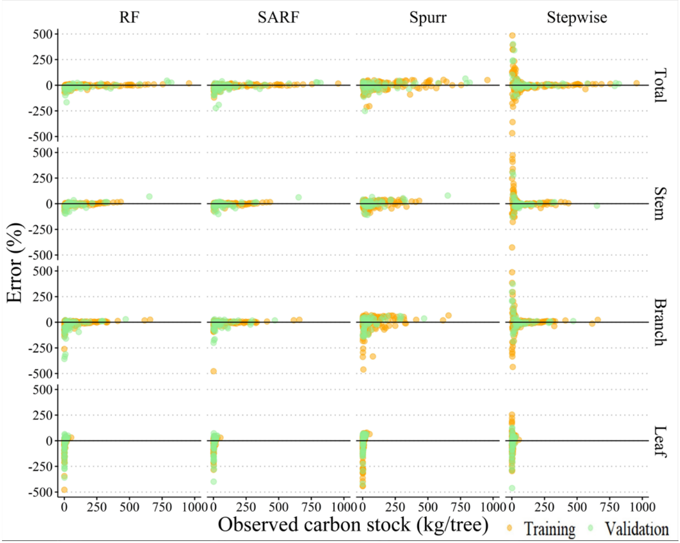

Figure 5 Residual distribution of the best models generated in each group to estimate the carbon stock (kg / tree).

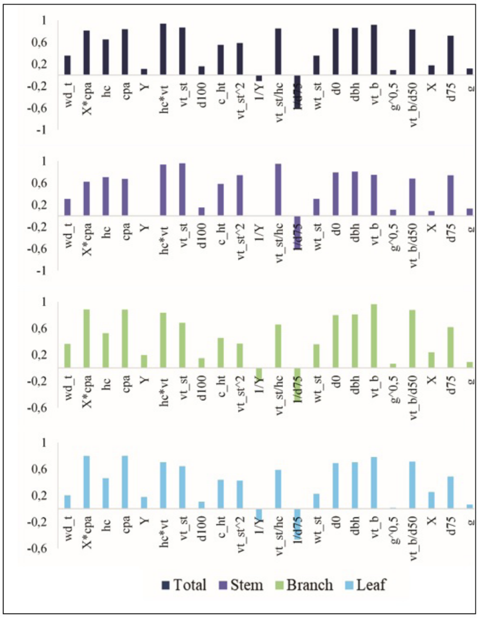

Figure 6 Pearson's linear correlation coefficients between the carbon stock of all compartment and the variables selected of the best selection methods used.

Where: *dbh*: diameter at breast height; *ht*: total height; *vt*: total volume; *wd_t*: total wood basic density; *cpa*: crown projection area; d_{25}: diameter at 25% of the tree; *vt_st*: total volume of the stem; *wd_st*: stem basic density; hc: commercial bole height; *vt_b*: total volume of branches; *wd_b*: branches basic density; X: longitude: Y: latitude: cbh: crown base height; d100: diameter at 100% of the tree; *c_ht*: ; d_{75}: diameter at 75% of the tree; d_0: diameter at 0% of the tree; *g*: sectional area; d_{50}: diameter at 50% of the tree.

to understand the role of ecosystems in the global carbon cycle, in addition to ensuring sustainable management of forest resources. In general, according to Marziliano et al. (2017), the quantification of the carbon stock is performed by means of biomass conversion standard factors. However, these values may vary according to the forest structure and environmental conditions, so their use could result in unreliable assessments. In this sense, the present study suggested testing different methodologies to select variables and estimate the aboveground carbon stock of native vegetation. According to this study's findings, the Spurr model did not present accurate carbon stock estimates for the compartments (Table 3). In contrast, the pan-tropical equation suggested by Chave et al. (2005) also evaluated for total carbon, stood out for showing greater homogeneity of residues. However, this equation had a high RMSE value. A similar result was obtained by Segura et al. (2018), who evaluated and compared the application of local and global models in tropical forests in the Colombian Pacific. The authors found that the equation proposed by Chave et al. (2005) and the other global equations analyzed showed a behavior of underestimating the biomass of trees. Thus, the authors point out the importance of developing tools for generating local models.

Besides these models, the use of stepwise and algorithms for the construction of new models were also evaluated, in order to estimate the carbon stock between the tree compartments. The results obtained (Table 3) indicated that, in general, the models developed using several explanatory variables showed very different responses. The use of stepwise and machine learning algorithms provided improvements in the adjustments, as they were able to explain the variance of the data by presenting R^2 values between 63% and 99%. However, it is noticed that for some compartments the models generated showed high values of RMSE%. This result can be attributed to the different phytophysiognomies present in the area, since the stocks of biomass and carbon vary between species and forest types. In line with this statement, Tetemke (2021) reported an indirect increase in aboveground carbon stock by species diversity due to stand structural diversity in a dry Afromontane forest. Because carbon variation reflects the morphological characteristics that differ in the way of using light, in the competition relationships between trees and local conditions that, generally, lead to a substantial variation in estimates (Sanquetta et al. 2011, Marziliano et al. 2013, Coletta et al. 2016). In this way, Henry et al. (2011) evidenced that the development of allometric models with stratification by ecological types of forests is a highly effective way to improve estimates.

Regarding the use of the RF algorithm, the robustness of the method was observed, which was not greatly affected by the inclusion of many variables related to each other. Wu et al. (2016) and Wu et al. (2018) evaluated the use of RF and other regression approaches to estimate the aboveground carbon stock. In these studies, the authors identified that RF provided accurate and satisfactory estimates with higher R^2 and low RMSE. Contrary to the results obtained in this study, in which the use of the stepwise approach presented better estimates, mainly for the validation set. However, it is emphasized that this technique has its performance affected by correlated and noisy variables (Gauchi & Chagnon 2001).

Despite this, the use of the variable selection technique made by the meta-heuristic SA managed to contribute considerably in reducing the number of variables and in the error of the estimates. This behavior was expected since a multi-objective condition was established. Manimala et al. (2011) evaluated the use of the genetic algorithm and Simulated Annealing as methods to select characteristics and optimize the parameters of the support vector machine. The authors found good results for both algorithms. Nevertheless, they emphasize that SA produces good solutions in a short time.

Regarding the variables that most contribute to estimate the aboveground carbon stock, there is a greater contribution of volume, basic density, crown projection area, diameter at 0%, diameter at breast height (*dbh*), height and latitude. Figure 6 shows that most of these variables have high correlation values with the carbon stock. The dry weight or biomass and the volume are directly related to the carbon stock, since they are used to generate conversion factors and expansion of biomass. Magalhães & Seifert (2015) evaluated the differences in estimates with the standard value for carbon content (50%), used by the Intergovernmental Panel on Climate Change. According to these authors, small differences in the estimates were found. However, although the differences are small, these errors can increase as the estimates are expanded, for example, at the stand level.

The *dbh* and *ht* variables are commonly used to estimate biomass and / or carbon stock due to their high relation (Vargas-Larreta et al. 2017, Sanquetta et al. 2018, Segura et al. 2018), this fact was also observed in this study. In addition to these findings, the wood basic density has great application in explaining the carbon stock, since this variable represents aspects related to the structure of the forest, such as the growth rate and the state of succession of the area (Ribeiro et al. 2011). According to Chave et al. (2005), these variables have great explanatory power since they reflect the different sizes between trees and species. Among the selected variables, the sectional area (*g*) and the diameter at 0% (d_0) were present for most of the analyzed compartments. The diameter at 0% was also found as a predictor of the biomass stock in the study by Henry et al. (2011). Regarding the sectional area and, consequently, the basal area, Burrows et al. (2000) indicate that this variable is a good predictor of biomass and carbon, since it includes the effect of the number and size of trees. These authors also emphasize that the relationship between biomass and basal area can be applied to assist in the estimates, as this variable can be quickly measured. Magalhães & Seifert (2015) state that the biomass stock, and consequently the carbon stock, is a function of the density of the stem and the *dbh*; therefore, it is also associated with the basal area. According to the authors, the higher the proportion of basal area, the higher the proportion of biomass stock.

The crown projection area (*cpa*) explains variations in individual tree growth, as their development is limited by the size of the crown and influenced by competition. Changes in the canopy of trees caused by interventions and variations in climatic and edaphic factors, are generally followed by an increase or decrease in biomass stock (Kuyah et al. 2012, Coletta et al. 2016). As an example, the tallest trees that stand out in the forest canopy have their growth reduced in height, while the growth of the branches horizontally increases. In this sense, adding variables related to crown can improve carbon stock estimates by capturing the tree's biomass variations (Goodman et al. 2014), as seen in this study.

Latitude is a geographical variable that helps explain different biomes and phytogeographic characteristics. Moreover, it is one of the geographical elements that most influences the climate (Mello et al. 2013, Scolforo et al. 2015). According to the study by Scolforo et al. (2015), it was found that latitude and variation of the carbon stock distribution of the arboreous vegetation have a strong association in some biomes in the state of Minas Gerais (Brazil). The authors explain that as the latitude decreases, greater carbon stocks are found. Therefore, although most of the variables selected in this study are not usually addressed in forest inventories, the gain obtained with the inclusion of these variables in the estimates must be evaluated, since they present an association of carbon stock of the trees with the ecophysiological and environmental processes. Future studies could broaden the present study scope by integrating variables from several data sources, such as remote sensing and meteorological data. These variables can assist in generating predictive tools applied on a large scale in monitoring carbon stocks and implementing carbon conservation programs.

Conclusion

In view of the various methods evaluated, these presented different adjustment results. The model by Chave et al. (2005) presented a superior result to the Spurr linear model, which shows the importance of including basic density. Thus, the relevance of the interaction between variables is emphasized, since the methods of selecting variables showed good predictive capacity. stepwise and SARF provided good adjustments, with similar R^2 and minor errors for stepwise with the validation set. Among these, SARF was an alternative to the traditional method, since the application of stepwise is limited when it involves a greater number of variables than that of observations. The hybrid SARF method also stood out in relation to RF in its pure form, as it managed to reduce the error of the estimates with the use of an optimized set of variables. Demonstrating the relevance of the interaction of biometric variables, ecophysiological and environmental factors in the estimation of the carbon stock. In general, the most important variables for modeling the carbon stock in each compartment were volume, basic density, crown projection area, sectional area, 0% diameter, diameter at breast height, height and latitude, these being pure, processed or combined. Furthermore, they denote the applicability of Simulated Annealing in the selection of variables for modeling the aboveground carbon stock with Random Forest.

Acknowledgements

This study was in part financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Brazil) under Finance Code 001. We also thank to Companhia Energética de Minas Gerais S.A (CEMIG) for financial support in the collection of forest data in the Rio Grande basin used in this work.

References

Abbasi B., Niaki S.T.A., Khalife M.A., & Faize Y., 2011. A hybrid variable neighborhood search and simulated annealing algorithm to estimate the three parameters of the Weibull distribution. Expert Systems with Applications, 38(1):700-708. https://doi.org/10.1016/j.eswa.2010.07.022

Anifowose F.A., Labadin J., Abdulraheem A., 2014. Non-linear feature selection-based hybrid computational intelligence models for improved natural gas reservoir characterization. J Nat Gas Sci Eng 21:397–410. https://doi.org/10.1016/j.jngse.2014.09.001

Bagherzadeh-Khiabani F., Ramezankhani A., Azizi F., Hadaegh F., Steyerberg E.W., Khalili D., 2016. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J Clin Epidemiol 71:76–85. https://doi.org/10.1016/j.jclinepi.2015.10.002

Baker D.J., Richards G., Grainger A., Gonzalez P., Brown S., DeFries R., Held A., Kellndorfer J., Ndunda P., Ojima D. et al., 2010. Achieving forest carbon information with higher certainty: A five-part plan. Environ Sci Policy 13:249–260. https://doi.org/10.1016/j.envsci.2010.03.004

Breiman L., 2001. Random forests. Machine Learning 45, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Bui D.T., Bui Q.T., Nguyen Q.P., Pradhan B., Nampak H., Trinh P.T., 2017. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. Agric For Meteorol 233:32–44. https://doi.org/10.1016/j.agrformet.2016.11.002

Burrows W.H., Hoffmann M.B., Compton J.F., Back P.V., Tait L.J., 2000. Allometric relationships and community biomass estimates for some dominant eucalypts in Central Queensland woodlands. Aust J Bot 48:707–714. https://doi.org/10.1071/BT99066

Carreiras J.M.B., Vasconcelos M.J., Lucas R.M., 2012. Understanding the relationship between aboveground biomass and ALOS PALSAR data in the forests of Guinea-Bissau (West Africa). Remote Sens Environ 121:426–442. https://doi.org/10.1016/j.rse.2012.02.012

Carvalho L.G. de, Oliveira M.S. de, Alves M. de C., Vianello R.L., Sediyama G.C., Neto P.C., Dantas A.A.A., 2008. Clima. In: Scolforo J.R.S., Carvalho L.M.T., Oliveira A.D. (eds.) Zoneamento ecológico - econômico do estado de Minas Gerais: comoponentes geofísico e biótico. pp 89–101.

Chave J., Andalo C., Brown S., Cairns M.A., Chambers J.Q., Eamus D., Fölster H., Fromard F., Higuchi N., Kira T. et al., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145:87–99. https://doi.org/10.1007/s00442-005-0100-x

Chave J., Réjou-Méchain M., Búrquez A., Chidumayo E., Colgan M.S., Delitti W.B.C., Duque A., Eid T., Fearnside P.M., Goodman R.C. et al., 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Glob Chang Biol 20:3177–3190. https://doi.org/10.1111/gcb.12629

Chinembiri T.S., Bronsveld M.C., Rossiter D.G., Dube T., 2013. The precision of C stock estimation in the Ludhikola watershed using model-based and design-based approaches. Nat Resour Res 22:297–309. https://doi.org/10.1007/s11053-013-9216-6

Coletta V., Menguzzato G., Pellicone G., Veltri A., Marziliano P.A., 2016. Effect of thinning on above-ground biomass accumulation in a Douglas-fir plantation in southern Italy. J For Res 27:1313–1320. https://doi.org/10.1007/s11676-016-0247-9

Corona-Núñez R.O., Mendoza-Ponce A., López-Martínez R., 2017. Model selection changes the spatial heterogeneity and total potential carbon in a tropical dry forest. For Ecol Manage 405:69–80. https://doi.org/10.1016/j.foreco.2017.09.018

Dantas D., Terra M. de C.N.S., Schorr L.P.B., Calegario N., 2021. Machine learning for carbon stock prediction in a tropical forest in southeastern brazil. Bosque 42:131–140. https://doi.org/10.4067/S0717-92002021000100131

Day P., Iannucci S., Banicescu I., 2020. Autonomic feature selection using computational intelligence. Futur Gener Comput Syst 111:68–81. https://doi.org/10.1016/j.future.2020.04.015

Drake J.M., Randin C., Guisan A., 2006. Modelling ecological niches with support vector machines. J Appl Ecol 43:424–432. https://doi.org/10.1111/j.1365-2664.2006.01141.x

Enquist B.J., Niklas K.J., 2001. Correction: Corrigendum: Invariant scaling relations across tree-dominated communities. Nature 425:741. https://doi.org/10.1038/nature02023

Feldpausch T.R., Lloyd J., Lewis S.L., Brienen R.J.W., Gloor M., Monteagudo Mendoza A., Lopez-Gonzalez G., Banin L., Abu Salim K., Affum-Baffoe K. et al., 2012. Tree height integrated into pantropical forest biomass estimates. Biogeosciences 9:3381–3403. https://doi.org/10.5194/bg-9-3381-2012

Fox J., Weisberg S., Price B., et al., 2019. car: Companion to applied regression. R package version 3(3).

Gauchi J.-P., Chagnon P., 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemom Intell Lab Syst 58:171–193. https://doi.org/10.1016/S0169-7439(01)00158-7

Genuer R., Poggi J.M., Tuleau-Malot C., 2010. Variable selection using random forests. Pattern Recognit Lett 31:2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Gleason C.J., Im J., 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. Remote Sens Environ 125:80–91. https://doi.org/10.1016/j.rse.2012.07.006

Goodman R.C., Phillips O.L., Baker T.R., 2014. The importance of crown dimensions to improve tropical tree biomass estimates . Ecol Appl 24(4):680-698. https://doi.org/10.1890/13-0070.1

Guangyi M., Yujun S., Saeed S., 2017. Models for predicting the biomass of *Cunninghamia lanceolata* trees and stands in Southeastern China. PLoS One 12:1–14. https://doi.org/10.1371/journal.pone.0169747

Guo P.-T., Li M.-F., Luo W., Tang Q.-F., Liu Z.-W., Lin Z.-M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma 237–238:49–59. https://doi.org/10.1016/j.geoderma.2014.08.009

Guyon I., Elisseeff A., 2003. An introduction to variable and feature selection. J of Machine Learn Res 3:1157–1182.

Han J., Kamber M., Pei J., 2011. Data mining: concepts and techniques. Elsevier.

Hapfelmeier A., Ulm K., 2014. Variable selection by Random Forests using data with missing values. Comput Stat Data Anal 80:129–139. https://doi.org/10.1016/j.csda.2014.06.017

Heinrich V.H.A., Dalagnol R., Cassol H.L.G., et al., 2021. Large carbon sink potential of secondary forests in the Brazilian Amazon to mitigate climate change. Nat Commun 12:1785. https://doi.org/10.1038/s41467-021-22050-1

Henry M., Picard N., Trotta C., Manlay R.J., Valentini R., Bernoux M., Saint-André L., 2011. Estimating tree biomass of sub-Saharan African forests: A review of available allometric equations. Silva Fenn 45:477–569. https://doi.org/10.14214/sf.38

Hong H., Tsangaratos P., Ilia I., Liu J., Zhu A.-X., Xu C., 2018. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. Sci Total Environ 630:1044–1056. https://doi.org/10.1016/j.scitotenv.2018.02.278

Kavzoglu T, Mather PM, 2002. The role of feature selection in artificial neural network applications. Int J Remote Sens 23:2919–2937. https://doi.org/10.1080/01431160110107743

Kazempour Larsary M, Pourbabaei H, Sanaei A, et al., 2021. Tree-size dimension inequality shapes aboveground carbon stock across temperate forest strata along environmental gradients. For Ecol Manage 496:1–10. https://doi.org/10.1016/j.foreco.2021.119482

Kuyah S, Muthuri C, Jamnadass R, Mwangi P, Neufeldt H, Dietz J, 2012. Crown area allometries for estimation of aboveground tree biomass in agricultural landscapes of western Kenya. Agrofor Syst 86:267–277. https://doi.org/10.1007/s10457-012-9529-1

Labani M, Moradi P, Ahmadizar F, Jalili M, 2018. A novel multivariate filter method for feature selection in text classification problems. Eng Appl Artif Intell 70:25–37. https://doi.org/10.1016/j.engappai.2017.12.014

Liu H., Yu L., 2005. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17:491–502. https://doi.org/10.1109/TKDE.2005.66

Mafarja M.M., Mirjalili S., 2017. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection. Neurocomputing 260:302–312. https://doi.org/10.1016/j.neucom.2017.04.053

Magalhães T.M., Seifert T., 2015. Estimation of tree biomass, carbon stocks, and error propagation in Mecrusse Woodlands. Open J For 05:471–488. https://doi.org/10.4236/ojf.2015.54041

Manimala K., Selvi K., Ahila R., 2011. Hybrid soft computing techniques for feature selection and parameter optimization in power quality data mining. Appl Soft Comput J 11:5485–5497. https://doi.org/10.1016/j.asoc.2011.05.010

Marziliano P.A., Lafortezza R., Colangelo G., Davies C., Sanesi G., 2013. Structural diversity and height growth models in urban forest plantations: A case-study in northern Italy. Urban For Urban Green 12:246–254. https://doi.org/10.1016/j.ufug.2013.01.006

Marziliano P.A., Menguzzato G., Scuderi A., Scuderi A., Scalise C., Coletta V., 2017. Biomass conversion and expansion factors in Douglas-fir stands of different planting density: Variation according to individual growth and prediction equations. For Syst 26(1):e003. https://doi.org/10.5424/fs/2017261-10239

Mello C.R., Viola M.R., Beskow S., Norton L.D., 2013. Multivariate models for annual rainfall erosivity in Brazil. Geoderma 202–203:88–102. https://doi.org/10.1016/j.geoderma.2013.03.009

Mendoza-Ponce A., Corona-Núñez R., Kraxner F., Leduc S., Patrizio P., 2018. Identifying effects of land use cover changes and climate change on terrestrial ecosystems and carbon stocks in Mexico. Glob Environ Chang 53:12–23. https://doi.org/10.1016/j.gloenvcha.2018.08.004

Mouazen A.M., Kuang B., De Baerdemaeker J., Ramon H., 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma 158:23–31. https://doi.org/10.1016/j.geoderma.2010.03.001

Návar J., 2009. Allometric equations for tree species and carbon stocks for forests of northwestern Mexico. For Ecol Manage 257:427–434. https://doi.org/10.1016/j.foreco.2008.09.028

Nunes M.H., Görgens E.B., 2016. Artificial intelligence procedures for tree taper estimation within a complex vegetation mosaic in Brazil. PLoS One 11(5):e0154738. https://doi.org/10.1371/journal.pone.0154738

Ou Q., Lei X., Shen C., 2019. Individual tree diameter growth models of larch–spruce–fir mixed forests based on machine learning algorithms. Forests 10:187. https://doi.org/10.3390/f10020187

Pechanec V., Purkyt J., Benc A., Nwaogu C., Štěrbová L., Cudlín P., 2018. Modelling of the carbon sequestration and its prediction under climate change. Ecol Inform 47:50–54. https://doi.org/10.1016/j.ecoinf.2017.08.006

Pham MH, Do TH, Pham VM, Bui QT, 2020. Mangrove forest classification and aboveground biomass estimation using an atom search algorithm and adaptive neuro-fuzzy inference system. PLoS One 15(5):e0233110. https://doi.org/10.1371/journal.pone.0233110

Picard N., Saint-André L., Henry M., 2012. Manual for building tree volume and biomass allometric equations: From field measurement to prediction. Food and Agricultural Organization of the United Nations, Rome, and Centre de Coopération Internationale en Recherche Agronomique pour le Développement.

Pilli R., Anfodillo T., Carrer M., 2006. Towards a functional and simplified allometry for estimating forest biomass. For Ecol Manage 237:583–593. https://doi.org/10.1016/j.foreco.2006.10.004

Reddy N., Gebreslasie M., Ismail R., 2017. A hybrid partial least squares and random forest approach to modelling forest structural attributes using multispectral remote sensing data. South African J Geomatics 6(3):377. https://doi.org/10.4314/sajg.v6i3.8

Ribeiro S.C., Fehrmann L., Soares C.P.B., Jacovine L.A.G., Kleinn C., Gaspar R.O., 2011. Above- and belowground biomass in a Brazilian Cerrado. For Ecol Manage 262(3):491–499. https://doi.org/10.1016/j.foreco.2011.04.017

Sanquetta C.R., Corte A.P.D., Da Silva F., 2011. Biomass expansion factor and root-to-shoot ratio for Pinus in Brazil. Carbon Balance Manag 6(6). https://doi.org/10.1186/1750-0680-6-6

Sanquetta C.R., Dalla Corte A.P., Behling A., Oliveira Piva L.R., Péllico Netto S., Rodrigues A.L., Sanquetta M.N.I., 2018. Selection criteria for linear regression models to estimate individual tree biomasses in the Atlantic Rain Forest, Brazil. Carbon Balance Manag 13:25. https://doi.org/10.1186/s13021-018-0112-6

Scolforo H.F., Scolforo J.R.S., Mello C.R., Mello J.M., Ferraz Filho A.C., 2015. Spatial distribution of aboveground carbon stock of the arboreal vegetation in Brazilian Biomes of Savanna, Atlantic Forest and Semi-arid woodland. PLoS One 10(6):e0128781. https://doi.org/10.1371/journal.pone.0128781

Segura M.A., Acuña L.M., Andrade H.J., 2018. Allometric models to estimate aboveground biomass of small trees in wet tropical forests of Colombian Pacific Area. Rev Árvore 42. https://doi.org/10.1590/1806-90882018000200009

Siddiq Z., Hayyat M.U., Khan A.U., et al., 2021. Models to estimate the above and below ground carbon stocks from a subtropical scrub forest of Pakistan. Glob Ecol Conserv 27:e01539. https://doi.org/10.1016/j.gecco.2021.e01539

Silveira E.M. de O., Silva S.H.G., Acerbi-Junior F.W., Carvalho M.C., Carvalho L.M.T., Scolforo J.R.S., Wulder M.A., 2019. Object-based random forest modelling of aboveground forest biomass outperforms a pixel-based approach in a heterogeneous and mountain tropical environment. Int J Appl Earth Obs Geoinf 78:175–188. https://doi.org/10.1016/j.jag.2019.02.004

Sun S., Peng Q., Shakoor A., 2014. A kernel-based multivariate feature selection method for microarray data classification. PLoS One 9(7):e102541. https://doi.org/10.1371/journal.pone.0102541

Tetemke B.A., Birhane E., Rannestad M.M., Eid T., 2021. Species diversity and stand structural diversity of woody plants predominantly determine aboveground carbon stock of a dry Afromontane forest in Northern Ethiopia. For Ecol Manage 500(15):119634. https://doi.org/10.1016/j.foreco.2021.119634

Vafaei S., Soosani J., Adeli K., Fadaei H., Naghavi H., Pham T.D., Bui D.T., 2018. Improving accuracy estimation of Forest Aboveground Biomass based on incorporation of ALOS-2 PALSAR-2 and Sentinel-2A imagery and machine learning: A case study of the Hyrcanian forest area (Iran). Remote Sens 10(2):172. https://doi.org/10.3390/rs10020172

Vahedi A.A., 2016. Artificial neural network application in comparison with modeling allometric equations for predicting above-ground biomass in the Hyrcanian mixed-beech forests of Iran. Biomass and Bioenergy 88:66–76. https://doi.org/10.1016/j.biombioe.2016.03.020

Vargas-Larreta B., López-Sánchez C.A., Corral-Rivas J.J., López-Martínez J.O., Aguirre-Calderón C.G., Álvarez-González J.G., 2017. Allometric equations for estimating biomass and carbon stocks in the temperate forests of North-Western Mexico. Forests 8(8):269. https://doi.org/10.3390/f8080269

Vicharnakorn P., Shrestha R.P., Nagai M., Salam A.P., Kiratiprayoon S., 2014. Carbon stock assessment using remote sensing and forest inventory data in Savannakhet, Lao PDR. Remote Sens 6:5452–5479. https://doi.org/10.3390/rs6065452

Vieilledent G., Vaudry R., Andriamanohisoa S.F.D., Rakotonarivo O.S., Randrianasolo H.Z., Razafindrabe H.N., Rakotoarivony C.B., Ebeling J., Rasamoelina

M., 2012. A universal approach to estimate biomass and carbon stock in tropical forests using generic allometric models. Ecol Appl 22:572–583. https://doi.org/10.1890/11-0039.1

Vieira G.C., de Mendonça A.R., da Silva G.F., Zanetti S.S., da Silva M.M., dos Santos A.R., 2018. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. Sci Total Environ 619–620:1473–1481. https://doi.org/10.1016/j.scitotenv.2017.11.138

Wang B., Waters C., Orgill S., Cowie A., Clark A., Li Liu D., Simpson M., McGowen I., Sides T., 2018. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. Ecol Indic 88:425–438. https://doi.org/10.1016/j.ecolind.2018.01.049

Wang M., Wan Y., Ye Z., Lai X., 2017. Remote sensing image classification based on the optimal support vector machine and modified binary coded ant colony optimization algorithm. Inf Sci (Ny) 402:50–68. https://doi.org/10.1016/j.ins.2017.03.027

Were K., Bui D.T., Dick Ø.B., Singh B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Indic 52:394–403. https://doi.org/10.1016/j.ecolind.2014.12.028

Wu C., Shen H., Shen A., Deng J., Gan M., Zhu J., Xu H., Wang K., 2016. Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery. J Appl Remote Sens 10:035010. https://doi.org/10.1117/1.jrs.10.035010

Wu C., Tao H., Zhai M., Lin Y., Wang K., Deng J., Shen A., Gan M., Li J., Yang H., 2018. Using nonparametric modeling approaches and remote sensing imagery to estimate ecological welfare forest biomass. J For Res 29:151–161. https://doi.org/10.1007/s11676-017-0404-9

Zianis D., Mencuccini M., 2004. On simplifying allometric analyses of forest biomass. For Ecol Manage 187:311–332. https://doi.org/10.1016/j.foreco.2003.07.007