PAB
Pesquisa Agropecuária Brasileira

Remote Sensing/ Original Article

ISSN 1678-3921

Journal homepage: www.embrapa.br/pab

For manuscript submission and journal contents, access: www.scielo.br/pab

Data mining applied to feature selection methods for aboveground carbon stock modelling

Abstract – The objective of this work was to apply the random forest (RF) algorithm to the modelling of the aboveground carbon (AGC) stock of a tropical forest by testing three feature selection procedures – recursive removal and the uniobjective and multiobjective genetic algorithms (GAs). The used database covered 1,007 plots sampled in the Rio Grande watershed, in the state of Minas Gerais state, Brazil, and 114 environmental variables (climatic, edaphic, geographic, terrain, and spectral). The best feature selection strategy – RF with multiobjective GA – reaches the minor root-square error of 17.75 Mg ha^{-1} with only four spectral variables – normalized difference moisture index, normalized burn ratio 2 correlation texture, treecover, and latent heat flux –, which represents a reduction of 96.5% in the size of the database. Feature selection strategies assist in obtaining a better RF performance, by improving the accuracy and reducing the volume of the data. Although the recursive removal and multiobjective GA showed a similar performance as feature selection strategies, the latter presents the smallest subset of variables, with the highest accuracy. The findings of this study highlight the importance of using near infrared, short wavelengths, and derived vegetation indices for the remote-sense-based estimation of AGC. The MODIS products show a significant relationship with the AGC stock and should be further explored by the scientific community for the modelling of this stock.

Index terms: forest management, genetic algorithm, random forest.

Mônica Canaan Carvalho[1] ✉, Lucas Rezende Gomide[1], José Roberto Soares Scolforo[1], Kalill José Viana da Páscoa[1], Laís Almeida Araújo[1] and Isáira Leite e Lopes[2]

[1] Universidade Federal de Lavras, Departamento de Ciências Florestais, Aquenta Sol, CEP 37200-900 Lavras, MG, Brazil.
E-mail: monicacanaan@gmail.com, lucasgomide@ufla.br, josescolforo@ufla.br, kalill.pascoa@ufla.br, la_sal@hotmail.com

[2] Eucatex S.A., Rua Ribeirão Preto, nº 909, Jardim Marilia, CEP 13323-010 Salto, SP, Brazil.
E-mail: isairaleite2010@gmail.com

✉ Corresponding author

Received
June 09, 2022

Accepted
August 30, 2022

How to cite
CARVALHO, M.C.; GOMIDE, L.R.; SCOLFORO, J.R.S.; PÁSCOA, K.J.V. da; ARAÚJO, L.A.; LOPES, I.L. e. Data mining applied to feature selection methods for aboveground carbon stock modelling. **Pesquisa Agropecuária Brasileira**, v.57, e03015, 2022. DOI: https://doi.org/10.1590/S1678-3921.pab2022.v57.03015.

Mineração de dados aplicada a métodos de seleção de variáveis para a modelagem de estoque de carbono acima do solo

Resumo – O objetivo deste trabalho foi aplicar o algoritmo "random forest" (RF) à modelagem do estoque de carbono acima do solo (CAS) de uma floresta tropical, por meio da testagem de três procedimentos de seleção de variáveis: remoção recursiva e algoritmos genéticos (AGs) uniobjetivo e multiobjetivo. Os dados utilizados abrangeram 1.007 parcelas amostradas na bacia hidrográfica do Rio Grande, no estado de Minas Gerais, Brasil, e 114 variáveis ambientais (climáticas, edáficas, geográficas, de terreno e espectrais). A melhor estratégia de seleção de variáveis – a RF com AG multiobjetivo – chega ao menor erro quadrático de 17,75 Mg ha^{-1} com apenas quatro variáveis espectrais – índice de umidade por diferença normalizada, textura de correlação do índice de queimada por razão normalizada 2, cobertura arbórea e fluxo de calor latente –, o que representa redução de 96,5% no tamanho do banco de dados. As estratégias de seleção de variáveis ajudam a obter melhor desempenho da RF, ao melhorar a acurácia e reduzir o volume dos dados. Embora a remoção recursiva e o AG multiobjetivo mostrem desempenho semelhante como estratégias de seleção de variáveis, esta último apresenta menor subconjunto de variáveis, com maior precisão. As descobertas deste trabalho destacam a importância do uso de infravermelho próximo, comprimentos de onda curtos e índices de

vegetação derivados para a estimativa de CAS baseada em sensoriamento remoto. Os produtos MODIS mostram relação significativa com o estoque de CAS e precisam ser melhor explorados pela comunidade científica para a modelagem deste estoque.

Termos para indexação: manejo florestal, algoritmo genético, floresta aleatória.

Introduction

Forest habitats are a notable carbon pool. Because of this importance, several scientific efforts seek to quantify the aboveground carbon (AGC) stock from native forests (Safari et al., 2017; Silveira et al., 2019), which is a crucial information to assess mitigation policies. Studies on this topic still face many challenges to predict AGC stocks, especially in large areas of moist tropical forests. For this reason, remote sensing techniques have been widely applied to the modelling of aboveground biomass and carbon stock, using a large set of spatial variables.

In the literature, there is a consensus on the use of spectral variables to support the attaining of satisfactory accuracy, since spectral and environmental variables (climate, soil, and surface relief) are commonly correlated with field data at a regional/global scale (Lu et al., 2016). In turn, surveys involving large areas show challenges compatible with their size, as dependent variable modelling demands the use of a special statistical or computational approach to get around these dimensional problems.

Machine learning methods can optimally assist the modelling complex task involving big data. Advances in machine learning techniques have contributed many valuable tools to the scientific community encompassing gain in novel insights within the temporal and spatial carbon variation (Mascaro et al., 2014). Random forest (RF) is a machine-learning algorithm that has been successfully used because it improves the modelling and accuracy of estimates of different ecological systems (Mascaro et al., 2014; Safari et al., 2017). Moreover, the RF can be used with the recursive removal method (Silveira et al., 2019) for feature selection in large datasets, boosting their final model performance. The feature selection is not a trivial task. Thus, computational methods help with the modelling task, mainly if the number of predictor variables exceeds the human analysis limit. A high-dimensional data set may lead to lower estimate accuracy, due to irrelevant and redundant variables, noise problems, and complexity to understand the pattern.

The feature selection procedure selects a subset according to mathematical methods and criteria (Rodriguez-Galiano et al., 2018). In this scenario, the integration between RF and genetic algorithm (GA) can bring benefits to feature selection, in which the use of the GA can guide the solution searching in this procedure (Kumar & Sahoo, 2017). GA derives from theories of the biological evolutionary process and natural selection to solve a series of combinatorial problems, providing an adaptive search engine for the optimal solution based on the principle of "survival and reproduction of the fittest". Generally speaking, these algorithms randomly create several solutions to a problem, from which those with the best performance will be selected to give rise to new solutions (by genetic operators: crossover and mutation), this process is repeated several times until a satisfactory solution is found (Kumar & Sahoo, 2017). Despite the robust performance achieved by GA in the search for optimized solutions to combinatorial problems, its application has not been explored yet for remote-sense-based AGC modelling.

This study evaluated the potential of database shrinkage, and its effects on the estimate accuracy, to define the best predictor subset of variables to explain the AGC stock in the Rio Grande watershed, located in the south of the state of Minas Gerais, Brazil. Therefore, the present study attempts to answer the following research questions: which feature selection procedure has the best predictive performance?; what are the variables that affect most the AGC stock pattern?

The objective of this work was to apply the RF algorithm to the modelling of the aboveground carbon (AGC) stock of a tropical forest by testing three feature selection procedures – recursive removal and the uniobjective and multiobjective genetic algorithms (GAs).

Materials and Methods

The study area is the Rio Grande watershed (86,110 km²), in the state of Minas Gerais, Brazil (Figure 1). This region has a highly varied ecological habitats,

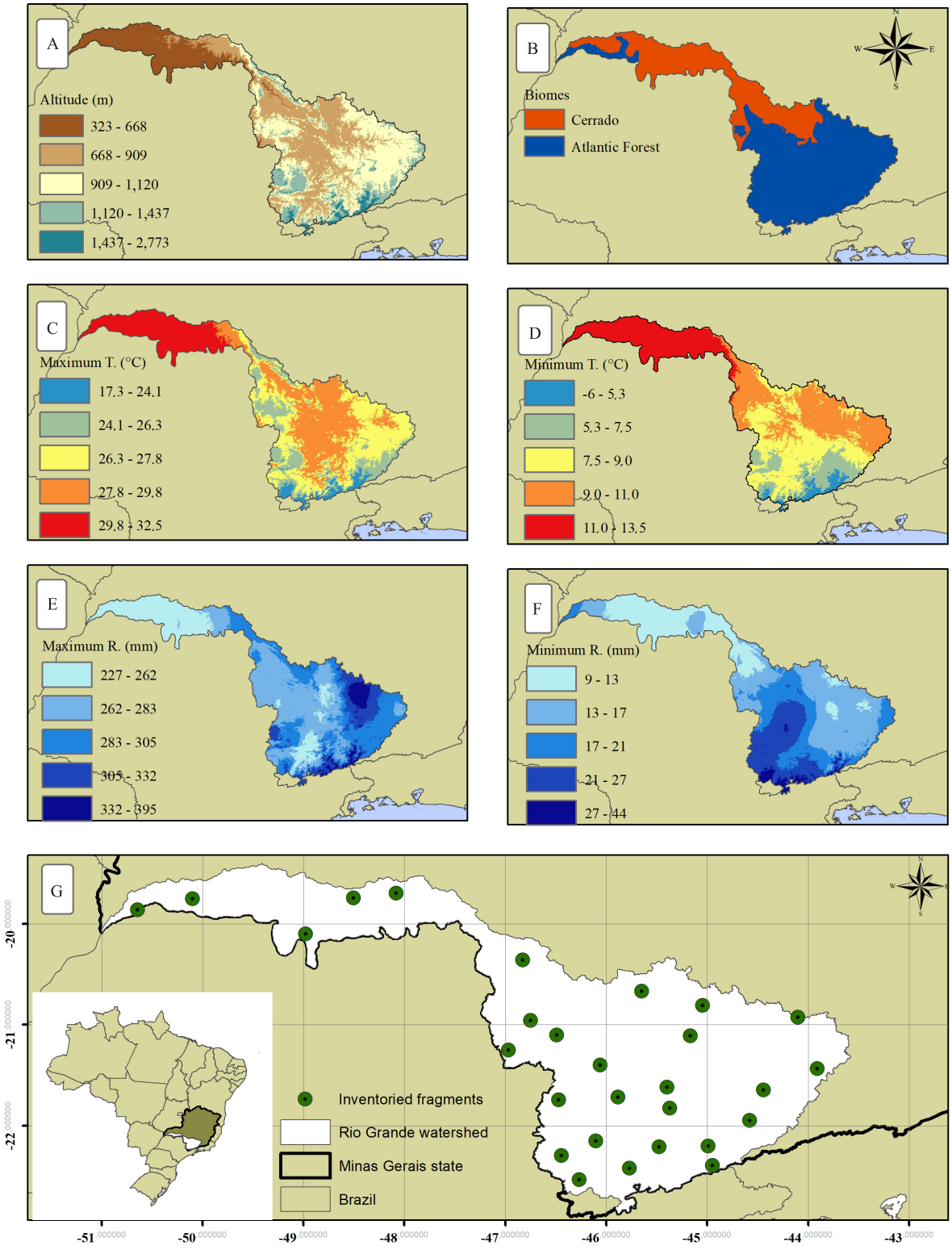

Figure 1. Rio Grande watershed, Minas Gerais state, Brazil. Description: A, altitude (SRTM – Shuttle Radar Topography Mission); B, biomes; C, monthly maximum temperature; D, monthly minimum temperature; E, precipitation in the wettest month; F, precipitation in the driest month (WorldClim version 1.4); G, location of inventoried fragments. August 18, 2022.

with a transition between Cerrado and Atlantic Forest biomes (Figure 1 B). Within the area, we highlight the existence of three types of vegetation: a rain forest (dense forest, with closed and continuous canopy, where the vegetation is highly influenced by high humidity and altitude); ii) a semideciduous forest (forest formation where part of the plants loses their foliage during the dry season, and where there is an understory formed by shrubs due to the existence of openings in the canopy; and an evergreen dry forest, locally known as "cerradão" (a forest formation with low density of individuals).

The forest inventory was performed by the cluster sampling methodology in the period 2014–2015. Rectangular clusters had three subplots with 250 m² (10 x 25 m). The arrangement was spatially distributed into systematic transects, 100 m apart, where the fragments had a minimum area to hold 10 conglomerates (30 subplots), totalizing 28 native vegetation fragments and 1,007 sampled subplots (Figure 1 G). We measured the diameter at breast height (DBH≥5 cm) and the height of all individuals in the subplots. Total height (H) was obtained by direct measurements using a telescopic ruler. The AGC quantification was based on the destructive sampling of 232 trees that were felled and their wood (trunk and branches at minimum 3 cm diameter) dimensions were measured using Huber method, to obtain the wood total volume (Scolforo & Thiersch, 2004). Discs of about 5 cm were removed in some relative heights from trunk and branches (> 3 cm) to determine the basic wood density (for dry mass conversion from volume) in laboratory. The crown compartment (leaves and branches with diameter less than 3 cm) had their fresh matter mass quantified still in the field, and their biomass was obtained by removing the moisture content (in the laboratory). Carbon content (percentage) of wood and crown compartments was obtained through the total organic carbon (TOC) analyzer Vario TOC cube (Elementar Analysensysteme GmbH, Langenselbold, Hesse, Germany).

Based on the wood carbon values (kg) plus crown compartments (total above ground carbon stock), we adjusted a multiple linear model showed below to estimate the AGC stock of individual trees (kg), using the forest inventory information (DBH and H). The adjusted coefficient of determination (R^2adjus%) and the residual standard error (Syx%) were respectively 96.2% and 52.5%. The tree AGC values were summed within each plot and normalized by the plot area in hectares (0.025 ha) (Mg ha^{-1}).

$$Ln(C) = \beta_0 + \beta_1 \times Ln(DBH) + \beta_2 \times Ln(H) \pm \varepsilon_i$$

in which: Ln is the natural logarithm; C is the aboveground carbon stock (kg); β_0, β_1, and β_2 are parameters; DBH is the diameter at breast height (cm); H is the total height (m); and ε_i is the error.

We investigated a wide range of variable types (climatic, topographic, geographic, spectral, and edaphic) for the AGC modelling, which constituted 114 environmental predictive variables (Figure 2). We downloaded 19 climatic variables from the WorldClim version 1.4 (Hijmans et al., 2005) with about 1 km² spatial resolution. This climatic dataset is widely applied in vegetation studies and have proven its validity in the determination of aboveground biomass (AGB) and AGC estimations (Silveira et al., 2019; Maia et al., 2020). We obtained 17 terrain variables from the digital elevation model (DEM) Shuttle Radar Topography Mission – SRTM (resampled to 100 m of spatial resolution), using the software SAGA GIS (Conrad et al., 2015). This software executes several algorithms in the DEM, producing different terrain variables, such as slopes, curvature, shading, place proximity to water channels, accumulation zones etc. Despite the application examples, the use of terrain variables for the modelling of AGC are rare in Brazil, and few studies indicate reliable contributions of these variables to the aboveground biomass/carbon modelling in tropical forests (Salinas-Melgoza et al., 2018; Silveira et al., 2019).

Spectral data were obtained from the Landsat 8 OLI satellite (30 m resolution) and MODIS (with a variable resolution between 250 and 1,000 m), in the same time interval of the forest inventory (2015). Seven vegetation indices were computed from the Landsat images: normalized difference vegetation index (NDVI); normalized difference moisture index (NDMI); enhanced vegetation index (EVI); soil-adjusted vegetation index (SAVI); modified soil-adjusted vegetation index (mSAVI; normalized burn ratio (NBR); and normalized burn ratio 2 (NBR2). Spectral indices have been used to estimate the AGB or AGC since two decades ago; however, there was no consensus for the best index applied to all vegetation types, once indices vary in their relationships with

biomass. Although vegetation indices – such as NDVI, EVI, SAVI, and mSAVI – have been proposed in previous studies to estimate biomass, some researchers on tropical forests found that spectral indices, including the near-infrared (NIR) wavelength, showed weaker relationships with biomass than those spectral indices including shortwave infrared (SWIR) (Lu et al., 2016; Silveira et al., 2019). NBR and NBR2 tend to incorporate the near-infrared (NIR) and shortwave infrared (SWIR) wavelengths because they often have strong relationships with observed AGB values (Nguyen et al., 2020).

For each vegetation index, the following textures measures were calculated: variance (var); homogeneity (homog); contrast (contrast); dissimilarity (dissim); entropy (entrop); second moment (secmom); and correlation (correl). For these calculations, a window (61 rows by 61 columns, 3721 pixels) was used by the grey level co-occurrence matrix methodology (Hamunyela et al., 2016). Textural measures have been used to produce new variables from multispectral data, enabling a reduction of impacts of data saturation in Landsat imagery on AGB estimation accuracy (Lu et al., 2016). Texture measures were applied to examine the relationships between biomass and textural images for secondary forest and mature forest in the state of Rondônia, Brazil (Lu & Bastistella, 2005). These authors found that spectral responses play roles that are more important for biomass modelling than the textural measures, when the forest stand structure is relatively simple; however, textural images are more important than spectral responses for complex forest stand structures.

The MODIS sensor derived other 12 variables for the year 2015: Earth's surface temperature (emis32, lstd, lstn), photosynthetic activity (fpar, lai), evapotranspiration (et, le, pet, ple), primary productivity (gpp, psnnet), and percentage of vegetation cover (treecover). The information provided by MODIS are highly valuable for forest management, such as forest biomass (Durante et al., 2019; Ploton et al., 2020). Although the spatial resolution of MODIS is coarse to the level of size plots, the integration of multiscale data from medium spatial resolution datasets, such as those from Landsat and radar, and coarse spatial resolution datasets, such as those from MODIS, are the direction for global/regional biomass estimation (Lu et al., 2016; Durante et al., 2019).

Soil physicochemical characteristics (organic matter content, pH, aluminum, clay, and the sum of bases) were determined at the first horizon soil depths (0-10 cm), at the midpoint of each subplot. Silicon dioxide (SiO_2) and total iron (Fe) were obtained from the portable X-ray fluorescence spectrometer (pXRF - Bruker model S1 Titan LE) (Silva et al., 2021). These data were interpolated with 100 m spatial resolution.

The AGC values were superimposed with environmental variables by using a standard grid size of 100 x 100 m to solve the scaling problem (Figure 2). Within each grid, the average values of carbon and environmental variables were extracted. Grids (671) were divided into two sets for assessing the random forest performance. The training set was used to adjust the models (70%) and the validation set for the analyses of the predictive ability of models (30%).

The RF algorithm was selected for the modelling of the AGC due to the following main characteristics: simple parameterization, robustness, and accuracy. The critical step in the production of accurate estimates is to identify an ideal subset of variables that reliably explains the pattern of the response variable. Variable selection has become essential to improve the modelling tasks, mainly when it is applied to high-dimensional data. This step allows of the removal of variables with low predictive power or autocorrelated variables. We tested three strategies to select a more efficient subset of variables and increase the predictive performance of the RF on carbon stock modelling, as follows: recursive removal of variables (RFrr); genetic algorithm with uniobjective function (GA-RFuni), and genetic algorithm with multiobjective function (GA-RFmulti). We also evaluated the model with the random forest algorithm with all variables to compare the results (RFall).

The first strategy uses the recursive removal of variables (RFrr) according to their ranking importance, in which the variable elimination is executed until the stop criteria (lowest mean square error) is attained. The iteration is associated with a single variable removal, and the remaining dataset starts a new algorithm cycle. The recursive feature elimination reduces potentially a very large number of variables to a more manageable subset. The initial parameter tests of RF algorithm suggested ntree = 1000 units and mtry = 10.67. The ntree is a parameter that defines the number of trees or algorithm divisions (a larger number of trees produces

more stable models, but requires more memory and a longer run time), and mtry is a parameter that defines the number of variables randomly sampled for splitting at each tree node (the default is the square root of the number of predictor variables).

The other two strategies of feature selection are hybrid methods. These methods consist of the GA managing the variable selection inserted within random forest. The procedure involved the dimensioning of the individuals in a defined length vector (114 genes). In this vector, each position represented a variable in the data set. AGC was the only fixed variable in this vector. Regarding the independent variables, a binary nature was assumed to activate (1) or not (0)

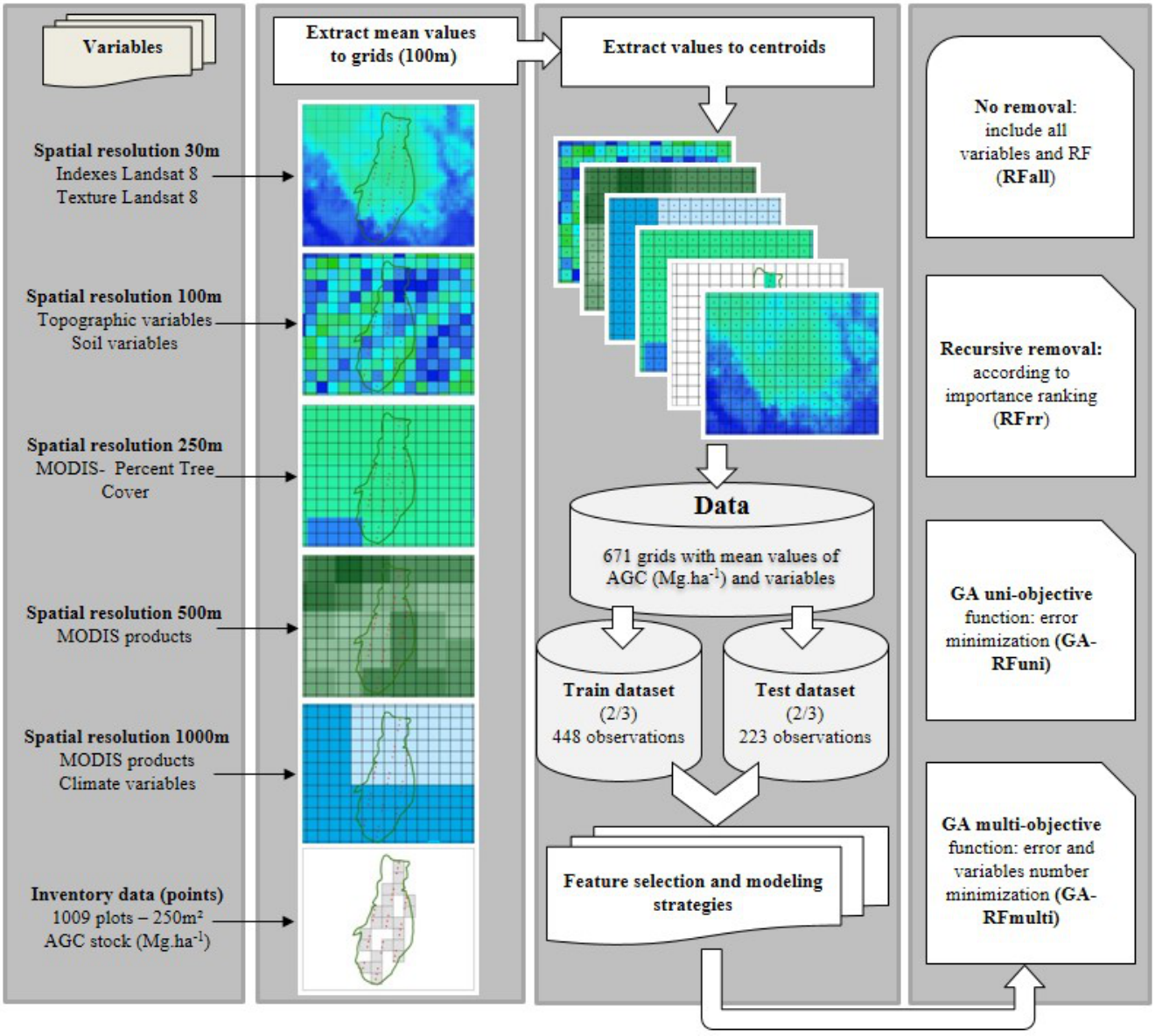

Figure 2. Flowchart of the procedures for data set arrangements and feature selection methods for AGC (aboveground carbon) modelling. RFall, random forest with all variables; RFrr, random forest with recursive removal feature selection; GA-RFuni, random forest with uniobjective genetic algorithm; AG-RFmulti, random forest with multiobjective genetic algorithm.

a determined variable inserted in the RF model. We used the fitness function to drive the search procedure and evaluated the effect of variable subset on the RF model accuracy. We tested two approaches of fitness functions in the GA-RF which differed by their goals. The first one is a uniobjective function (GA-RFuni) that sought the minimization of the out-of-bag (OOB) error. The other one is a multiobjective function (GA-RFmulti), by adding a second component responsible for minimizing the selected predictor variables. The OOB error is a method to measure the prediction error, in methodologies utilizing bootstrap aggregation, and the mean prediction error on each training sample is calculated only with the trees that did not make up the bootstrap sample (Mascaro et al., 2014). Besides, the normalized values of the multiobjective function were calculated for the equal weight of each characteristic. The denominator values are 1060 (maximum utopic OOB error) and 114 (total number of predictor variables), as follows:

$$\text{Fitness} = \frac{(\text{error OOB})}{1060} + \frac{\text{n}}{144}.$$

As the tested algorithms are stochastic techniques, each processing run time always results in new values. We run 50 times to achieve a consistent outcome due to the local optimal response problem (limitation of the algorithm's solution search space). To assess the performance of the methods, we used the following metrics: mean error (ME); root mean square error (RMSE); root mean square percentage error (RMSE%); residual plot analysis; and processing run time. The entire experiment was processed on a computer with an Intel Core i3-2100 processor at 3.10 MHz and 8 Gb of RAM. The RandomForest package (Liaw & Wiener, 2002) of R software was applied for the RF algorithm analysis. We coded the GA-RF algorithm in the same computational framework.

Results and Discussion

There was a clear pattern for which spectral variables (Landsat vegetation indices, its texture measures, and the MODIS products) are most useful for predicting AGC (Figure 3), specially, treecover and NDMI make up the most important predictors. Texture measures (homogeneity and correlation) of NBR, NBR2, and NDVI also showed a relevant influence on carbon stock. Treecover depicts a quantitative measure of woody cover and describes it as a percentage of ground cover. This variable plays an essential role in the present study, as it helps with distinguishing the vegetation types in an area, ranging from dense forests (Atlantic Forest) to fields with sparse trees (Cerrado), which causes a great variation in the carbon values. The importance values obtained by vegetation indices, such as NDMI, NBR, and NBR2, point to the relevance of using longer wavelengths (near infrared-shortwave infrared) in the AGC modelling, as suggested by some authors (Lu & Batistella, 2005; Campbell et al., 2021). Texture measures also showed be able to model AGC, featuring tree canopy cover and vegetation structure, surpassing the spectral data saturation as highlighted in other studies (Lu & Batistella, 2005; Lu et al., 2016).

Among the terrain variables, the direct insolation (direct_ins) and the vertical distance (vert_dist) stood out among the others (Figure 3). Isothermality (BIO3) and maximum temperature of the warmest month (BIO5) showed the highest-importance values among the climatic variables, and the sum of bases (SB), among the edaphic variables (Figure 3). Tropical forest C dynamics are tightly coupled with energy and water exchange between the biosphere and atmosphere (Wang et al., 2021). The vertical distance to a channel network base level is related with water availability, which drives strong differences in the biomass in deciduous upland and the semi-deciduous forests (Salinas-Melgoza et al., 2018). Earlier studies suggested that the tropical forest productivity could be limited more by solar radiation than by temperature and water (Seddon et al., 2016; Wang et al., 2021). Recent work suggests that temperature is also important in wet forests that operate close to a temperature optimum for their productivity (Huang et al., 2019).

The importance values (Increase MSE%) of variables were fewer than 8%, which shows a low explanatory power of the predictive variables (Figure 3). Local and regional scale studies that used similar variables obtained values of Increase MSE% above 20% and 30% (Silveira et al., 2019; Campbell et al., 2021). This fact can be attributed to the vegetation heterogeneity in the study area, and to the weak resolution of the predictive variables in relation to the subplot size (10 x 25 m) disturbing to directly link plot forest measures to satellite data due to coarser spatial resolution and positional uncertainty (Ploton et al., 2020). Even in

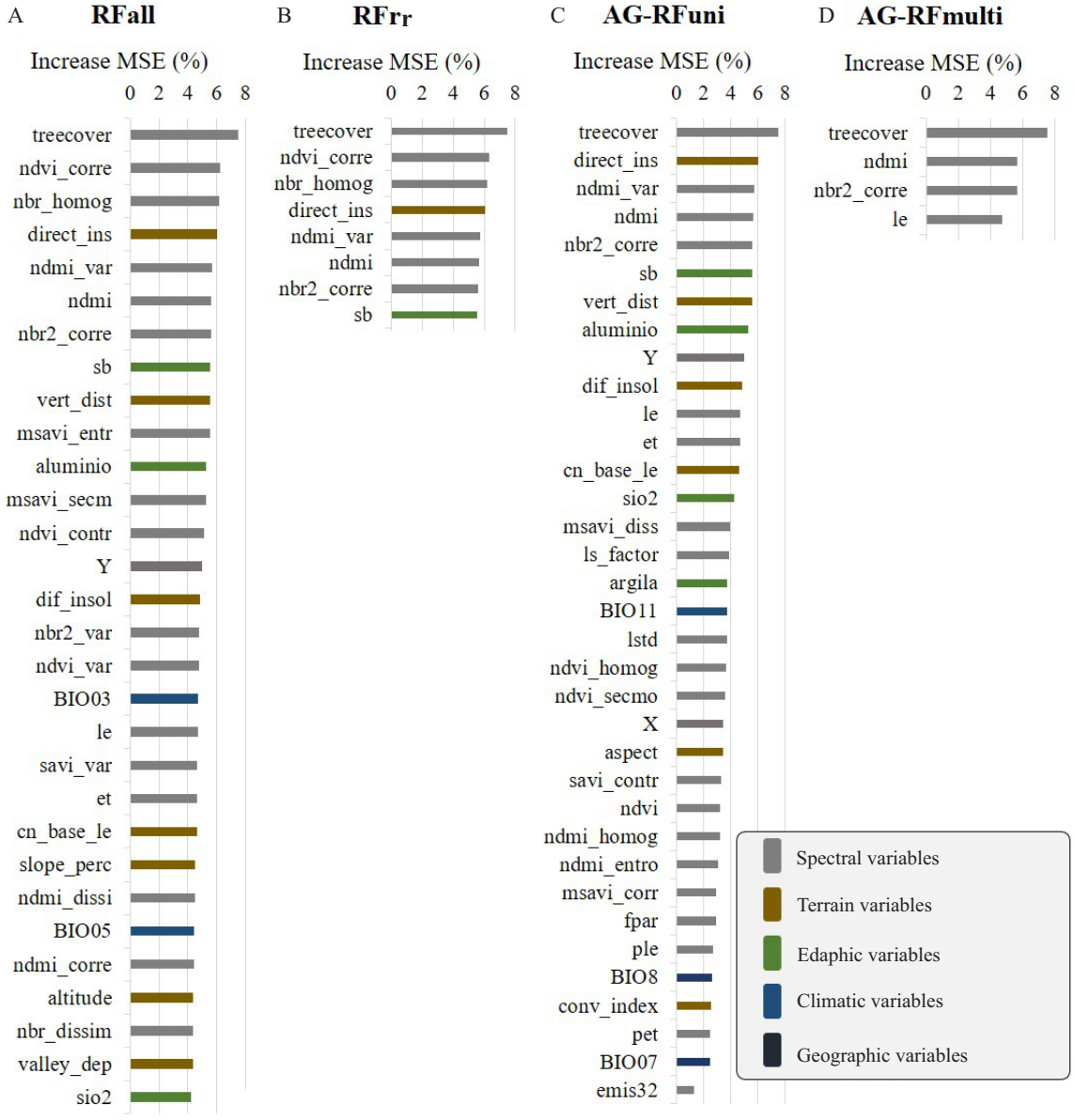

Figure 3. Variable importance ranking for the mean of Increase MSE (%) (mean square error), considering: A, RFall (random forest with all variables) with only 30 most important variables; B, variables selected by RFrr (random forest with recursive removal feature selection); C, GA-RFuni (random forest with uniobjective genetic algorithm); and D, GA-RFmulti (random forest with genetic algorithm multiobjective). Climatic variables: BIO03, isothermality; BIO05, maximum temperature of the warmest month; BIO07, annual temperature range; BIO08, mean temperature of the wettest quarter; BIO11, mean temperature of the coldest quarter. Terrain variables: direct_ins, direct insolation; dif_insol, diffuse insolation; cn_base_le, channel network base level; conv_index, convergence index; slope_perc, slope (%); valley_dep, valley depth; vert_dist, vertical distance. MODIS products: le, latent heat flux; et, global evapotranspiration; pet, potential global evapotranspiration; ple, potential latent heat flux; fpar, fraction of photosynthetically active radiation; lstd, land surface temperature day. Edaphic variables: sb, sum of bases; sio2, silicon dioxide. Vegetation indexes: EVI, enhanced vegetation index; NDVI, normalized difference vegetation index; NDMI, normalized difference moisture index; NBR, normalized burn ratio; NBR2, normalized burn ratio 2; SAVI, soil-adjusted vegetation index; MSAVI, modified soil-adjusted vegetation index. Geographic variables: X, latitude; Y, longitude. Texture measures: _corre, correlation; dissi, dissimilarity; var, variance; secm, second moment; _homog; homogeneity; _entrop, entropy; _contra, contrast.

the case of strong environmental contrasts monitored at fine scale, the environmental effects explain only a small fraction of variations in the AGB, and interact largely with the structural effects (Guitet et al., 2015; Silveira et al., 2019).

Regarding the modelling strategies based on different feature selections, the procedures did not show significant accuracy differences. However, it is noteworthy that they produced differences for selected variable quantities and variable subsets. Our best selection strategy (GA-RFmulti), with fewer variables and better accuracy, selected an optimized subset containing treecover, NDMI, NBR2_corre, and le. The treecover is the representation of the earth's surface vegetation cover, and le (latent heat flux) corresponds to the lost of water vapor from the Earth's surface to the atmosphere, which is called evapotranspiration. Even though these products are empirically related to forest AGC, they have not been applied to the estimation of remote-sense-based biomass. The NDMI detects the leaf water content and it is calculated by the ratio between NIR and SWIR1. The texture variable (nbr2_corre) reflects the correlation between vegetation water sensitivity, calculated by the ratio between SWIR1 and SWIR2, and the neighbor's pixels. These results highlight the significance of the SWIR and NIR wavelengths (which are less sensitive to atmospheric effects) for the AGC modelling and agree with the results of other studies (Silveira et al., 2019; Nguyen et al., 2020; Taddese et al., 2020). These selected variables confirm the relation between water availability and different AGC levels in the vegetation. As to the number of selected variables, the optimized subset of each tested feature selection resulted in 8 (RFrr), 35 (GA-RFuni), and 4 variables (GA-RFmulti) (Figure 3).

The strategies using GA were highly affected according to the enabled fitness function (uni/multi). The multiobjective function had 88.58% and 96.5% fewer variables than the uniobjective and RFall, respectively (Table 1). All selected variables in GA-RFmulti belonged to the spectral class (Figure 3 D). Furthermore, it is pertinent to highlight that the GA-RFuni strategy defined a set of 35 predictor variables, out of which 57% (20 variables) were spectral –, and some of them showed low values of IncMSE%. By selecting more variables, the GA-RFuni strategy allowed of the inclusion of predictors with IncMSE% (percentage increase in the mean square error) less than 4%, which means a low-explanatory power.

Overall, the prediction errors (RMSE%) had slight differences facing all tested strategies and datasets (Table 1). These variations were inferior to 1.4% (training) and 2.22% (validation). The validation of outcomes was satisfactory, with narrow differences for accuracy metrics in relation to training. The training phase adjusted the models efficiently, as their application to the validation set showed their generalization capacity for an independent base. This fact proved the applicability of the strategies. The comparison of the metrics ME, RMSE, and RMSE (%) showed a ranking based on accuracy decreasing for RFrr, GA-RFmulti, GA-RFuni, and RFall, in the training dataset, and a similar pattern in the validation dataset changed only RFrr and GA-RFmulti positions. Furthermore, it is worth noting that the residual plots showed a similar behavior for all strategies (Figure 4).

Table 1. Assessment metrics for training and validation data sets of the tested strategies (RFall, RFrr, GA-RFuni and GA-RFmulti) for the modelling of AGC stock.

Set	Strategy	Mean error	RMSE	RMSE (%)	Time[(1)] (s)	N
Training	RFall	-1.68	17.56	36.89	16.82	114
	RFrr	-1.26	16.89	35.49	204.41	8
	GA-RFuni	-1.33	17.03	35.79	1.656.00	35
	GA-RFmulti	-0.88	16.92	35.56	331.20	4
Validation	Rfall	-2.18	18.81	39.22	-	114
	RFrr	-1.61	18.11	37.74	-	8
	GA-RFuni	-1.61	18.34	38.24	-	35
	GA-RFmulti	-0.66	17.75	37.00	-	4

[(1)]Mean of the time-consuming (s) computation. RMSE, root mean square error; RMSE (%), root mean square percentage error; N, number of selected variables; RFall, random forest with all variables; RFrr, random forest with recursive removal feature selection; GA-RFuni, random forest with uniobjective genetic algorithm; AG-RFmulti, random forest with multiobjective genetic algorithm.

Model predictions showed a fairly linear relationship with the observed AGC, although the model tended to overestimate low AGC values and to underestimate high AGC values (Figure 4 A), which is a common bias pattern of the RF algorithm (Ploton et al., 2020).

Unfortunately, forest biomass/carbon estimates are associated with various errors and uncertainties (Guitet et al., 2015). Many studies have suggested that the relative errors (RMSE%) of the estimates can vary from 5% to 30%, depending on the forest ecosystems, topographic characteristics, remotely sensed data and their spatial resolutions, methods used etc. (Lu et al., 2016). An approach based on structural analysis of mixed pixels and the random forest model was proposed by Wang & Jiao (2020), in order to increase the accuracy of AGB estimated from coarse resolution data in broadleaf forest, mixed forest, and some coniferous forests. The results showed that the accuracy of AGB estimated from MODIS data was increased using this method, and RMSE decreased from 51.6 Mg ha^{-1} to 26.8 Mg ha^{-1}. Ploton et al. (2020) built an RF model to predict AGB from a combination of 9 MODIS products and 27 environmental variables in a tropical forest; their evaluation led to an estimated R^2 of 0.53 and RMSE of 56.5 Mg ha^{-1}. In the present study, the smallest error reached in validation was attained by the methodology GA-RFmulti (RMSE 17.75 Mg ha^{-1}). Our study area showed 47.59 Mg ha^{-1} average AGC stock, 25.05 Mg ha^{-1} standard deviation, and minimum and maximum values of 7.73 and 214.96 Mg ha^{-1}, respectively (Table 2). These results denote a high heterogeneity of AGC stock, which contributes to modelling uncertainties. The improvement of these estimates in extensive and heterogeneous landscapes requires a better understanding of the environmental system and its spatial variation (Guitet et al., 2015).

Even with all sampling and methodological efforts, the accuracy of the models did not reach the desired standard. This situation can be explained and even circumvented by taking some factors into account. The first one is the scale factor, in which variables with coarse resolution are linked to small plots areas. In the present study, AGC values obtained in 10 x 25 m plots were represented by variables with spatial resolution from 100 m to 1 km. The result is inaccurate because of the scaling effect, caused by nonlinearity in data representation, and because of the existence of mixed pixels containing different forest types and land uses (Ploton et al., 2020). An alternative to minimize the problem with scale is the use of object-based modelling instead of using the pixel or grid methodology (Silveira et al., 2019). Another point is that remotely sensed signals correlated to forest aboveground biomass– such as vegetation indices or surface reflectance in a particular wavelength – saturate when biomass reaches a threshold of 100–200 Mg ha^{-1} (Lu et al., 2016). Most recent efforts for the mapping forest aboveground biomass engage with remotely sensed data from multiple sensors, such as Lidar and SAR (Synthetic Aperture Radar) to get around this problem.

We suggest a similar performance for RFrr and GA-RFmulti (Table 1, Figure 4). These strategies with the smallest number of predictor variables obtained the best results concerning the accuracy of the models, confirming that it is better to use an optimal subset, rather than using all available variables. Feature selection is a complex task; therefore, the GA demands high time-consuming and computational efforts to search for the best variable subset. The processing time ranged between 98.45 (uniobjective function) and 19.69 more times than RFall (multiobjective function). This latter was similar to the RFrr (12.15 times of RFall).

Table 2. Descriptive statistics obtained for AGC stock from field inventory (real values), and the best predictive strategy – GA-RFmulti – random forest with multiobjective genetic algorithm – (predicted values), according to training and validation datasets.

Descriptive statistic	Training dataset – AGC stock (Mg ha^{-1})		Validation dataset – AGC stock (Mg ha^{-1})	
	Real (forest inventory)	Predicted (GA-RFmulti)	Real (forest inventory)	Predicted (GA-RFmulti)
Mean	47.59	48.47	46.69	48.58
Standard deviation	25.05	12.65	23.93	12.58
Maximum	214.96	120.99	178.71	98.12
Minimum	7.73	23.16	3.84	19.96

The RF algorithm efficiently performs information extraction even with a large data set. Our RF findings corroborate those by Speiser et al. (2019) suggesting the ability to handle datasets with a huge number of predictor variables. According to Rodriguez-Galiano et al. (2018), a large dataset with irrelevant features may affect the algorithm performance. It increases the model complexity and turns the replication impracticable in other areas. Identifying the optimum set of variables is also essential to tackle the variable redundancy problem (Speiser et al., 2019). A lower number of variables benefits the model application, optimizing the computational performance and processing time over large areas. Following the example of our study, we manipulated our dataset containing 114 variables from the Rio Grande watershed, which constituted a raster file with 160.2 GB size. Conversely, if we predict the AGC using the GA-RFmulti (4 variables), this file size will have only 5.5 GB or 3.43% of the total computational memory. In this context, the methodological efforts to reduce the dataset size are very important, mainly for mapping purposes in large areas. Nevertheless, the RFrr method has a questionable performance in extra-large databases. This fact is attributed to the application of unidirectional rules to remove undesired variables based on the increasing order of their importance values. This approach may result in a weak procedure, since the constraint order may affect or make impossible a positive combinatory behave of a variable set. It can lead RF to a greater probability of running with the same variables that are often highly correlated. Conversely, the GA-RF multiobjective function acted robustly, just focusing not only on the minimum error. The combination of two components (number of variable/error) guides for a better algorithm application. Kumar & Sahoo (2017) also applied the same meta-heuristic reducing 50% of the number of variables. Tavasoli & Arefi (2021) used Sentinel-2 optical data and GA-RF to estimate AGC stock. They reported GA-RF benefits, such as fast achievement of high accuracy and great capacity to reduce the number of predictors, improving the performance of the RF model. Therefore, our study encourages the application of the GA-RF multiobjective function, due to its valuable outcomes. The state of knowledge on the modelling of forest attributes has been changing due

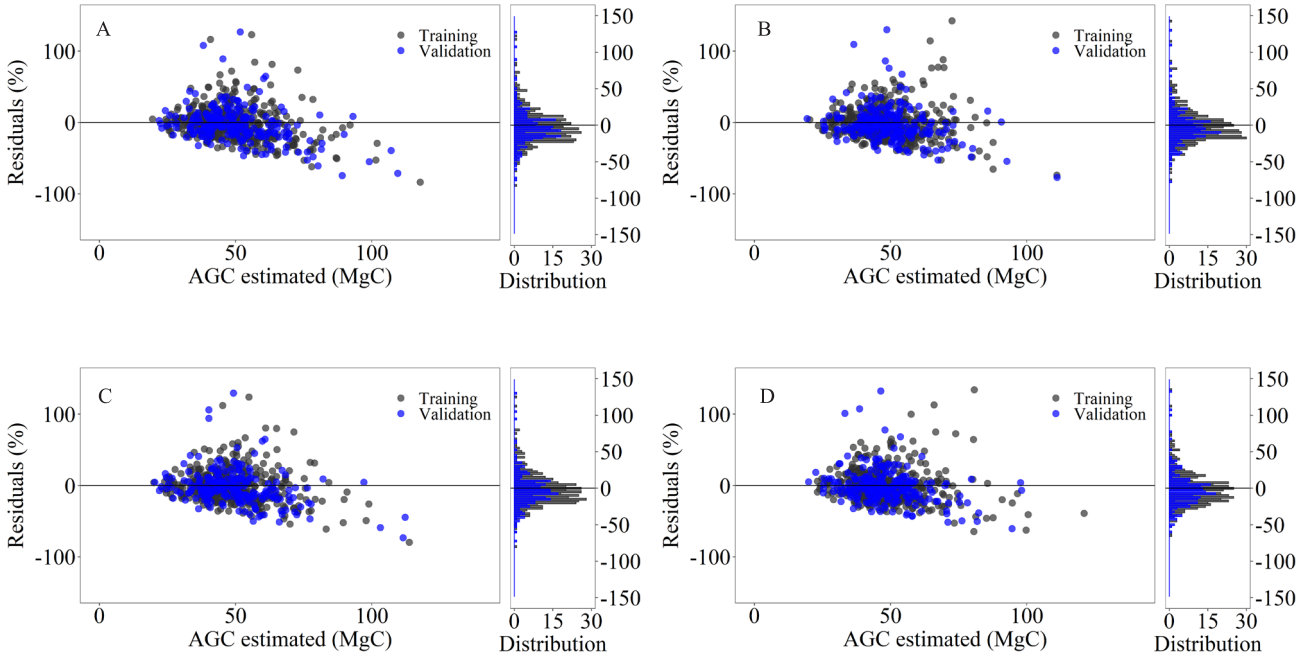

Figure 4. Residual plots of aboveground carbon stock (AGC, Mg ha^{-1}) for random forest strategies and datasets: A, RFall (random forest with all variables); B, RFrr (random forest with recursive removal feature selection); C, GA-RFuni (random forest with uniobjective genetic algorithm); and D, GA-RFmulti (random forest with multiobjective genetic algorithm).

to computational advances. The interest in results has moved from the consideration of statistical assessments only to the aggregation of more interpretative and robust meanings given to feature selection. Future studies combining big data with competing machine-learning models could broaden the insights of the present study. This approach may help to feed the systems of national forest service, as its findings would act decisively for forest management and planning.

Conclusions

1. Feature selection strategies assist in obtaining a better random forest (RF) performance, by improving the accuracy and reducing the volume of the data; although the recursive removal (RFrr) and multiobjective genetic algorithm (GA-RFmulti) showed a similar performance as feature selection strategies, the latter presents the smallest subset of variables, with the highest accuracy.

2. The best feature selection strategy– the random forest together with the multiobjective genetic algorithm – reaches the minor root-square error with only four spectral variables (the normalized difference moisture index, normalized burn ratio 2 – correlation texture, treecover, and latent heat flux), which represents a reduction of 96.5% in the size of the database.

3. Near infrared and short wavelengths are important for remote-sense-based aboveground carbon estimation; the vegetation indices derived these wavelength bands, since the normalized difference moisture index and normalized burn ratio prove their relevance in this task, even as its texture measures; the MODIS products, such as percent treecover and latent heat flux show a significant relationship with the aboveground carbon stock.

Acknowledgments

To Companhia Energética de Minas Gerais (Cemig) and to Universidade Federal de Lavras (Ufla), for the financial support of CEMIG GT-456 project; and to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes, Finance Code 001), for scholarship granted.

References

CAMPBELL, M.J.; DENNISON, P.E.; KERR, K.L.; BREWER, S.C.; ANDEREGG, W.R.L. Scaled biomass estimation in woodland ecosystems: testing the individual and combined capacities of satellite multispectral and lidar data. **Remote Sensing of Environment**, v.262, art.112511, 2021. DOI: https://doi.org/10.1016/j.rse.2021.112511.

CONRAD, O.; BECHTEL, B.; BOCK, M.; DIETRICH, H.; FISCHER, E.; GERLITZ, L.; WEHBERG, J.; WICHMANN, V.; BÖHNER, J. System for automated geoscientific analyses (SAGA) v.2.1.4. **Geoscientific Model Dev**elopment, v.8, p.1991-2007, 2015. DOI: https://doi.org/10.5194/gmd-8-1991-2015.

DURANTE, P.; MARTÍN-ALCÓN, S.; GIL-TENA, A.; ALGEET, N.; TOMÉ, J.L.; RECUERO, L.; PALACIOS-ORUETA, A.; OYONARTE, C. Improving aboveground forest biomass maps: from high-resolution to national scale. **Remote Sensing**, v.11, art.795, 2019. DOI: https://doi.org/10.3390/rs11070795.

GUITET, S.; HÉRAULT, B.; MOLTO, Q.; BRUNAUX, O.; COUTERON, P. Spatial structure of above-ground biomass limits accuracy of carbon mapping in rainforest but large scale forest inventories can help to overcome. **PLoS One**, v.10, e0138456, 2015. DOI: https://doi.org/10.1371/journal.pone.0138456.

HAMUNYELA, E.; VERBESSELT, J.; HEROLD, M. Using spatial context to improve early detection of deforestation from Landsat time series. **Remote Sensing of Environment**, v.172, p.126-138, 2016. DOI: https://doi.org/10.1016/j.rse.2015.11.006.

HIJMANS, R.J.; CAMERON, S.E.; PARRA, J.L.; JONES, P.G.; JARVIS, A. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v.25, p.1965-1978, 2005. DOI: https://doi.org/10.1002/joc.1276.

HUANG, M.; PIAO, S.; CIAIS, P.; PEÑUELAS, J.; WANG, X.; KEENAN, T.F.; PENG, S.; BERRY, J.A.; WANG, K.; MAO, J.; ALKAMA, R.; CESCATTI, A.; CUNTZ, M.; DE DEURWAERDER, H.; GAO, M.; HE, Y.; LIU, Y.; LUO, Y.; MYNENI, R.B.; NIU, S.; SHI, X.; YUAN, W.; VERBEECK, H.; WANG, T.; WU, J.; JANSSENS, I.A. Air temperature optima of vegetation productivity across global biomes. **Nature Ecology & Evolution**, v.3, p.772-779, 2019. DOI: https://doi.org/10.1038/s41559-019-0838-x.

KUMAR, S.; SAHOO, G. A random forest classifier based on genetic algorithm for cardiovascular diseases diagnosis. **International Journal of Engineering**, v.30, p.1723-1729, 2017.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R News**, v.2, p.18-22, 2002. Available at: <http://CRAN.R-project.org/doc/Rnews/>. Accessed on: Sept. 14 2022.

LU, D.; BATISTELLA, M.; MORAN, E. Satellite estimation of aboveground biomass and impacts of forest stand structure. **Photogrammetric Engineering & Remote Sensing**, v.71, p.967-974, 2005. DOI: https://doi.org/10.14358/PERS.71.8.967.

LU, D.; CHEN, Q.; WANG, G.; LIU, L.; LI, G.; MORAN, E. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. **International Journal of Digital Earth**, v.9, p.63-105, 2016. DOI: https://doi.org/10.1080/17538947.2014.990526.

MAIA, V.A.; SANTOS, A.B.M.; AGUIAR-CAMPOS, N. de; SOUZA, C.R. de; OLIVEIRA, M.C.F. de; COELHO, P.A.; MOREL, J.D.; COSTA, L.S. da.; FARRAPO, C.L.; FAGUNDES, N.C.A.; PAULA, G.G.P. de; SANTOS, P.F.; GIANASI, F.M.; SILVA, W.B. da; OLIVEIRA, F. de; GIRARDELLI, D.T.; ARAÚJO, F. de C. VILELA, T.A.; PEREIRA, R.T.; SILVA, L.C.A. da; MENINO, G.C. de O.; GARCIA, P.O.; FONTES, M.A.L.; SANTOS, R.M. dos. The carbon sink of tropical seasonal forests in southeastern Brazil can be under threat. **Science Advances** v.6, eabd4548, 2020. DOI: https://doi.org/10.1126/sciadv.abd4548.

MASCARO, J.; ASNER, G.P.; KNAPP, D.E.; KENNEDY-BOWDOIN, T.; MARTIN, R.E.; ANDERSON, C.; HIGGINS, M.; CHADWICK, K.D. A tale of two "forests": random forest machine learning aids tropical forest carbon mapping. **PloS ONE**, v.9, e85993, 2014. DOI: https://doi.org/10.1371/journal.pone.0085993.

NGUYEN, T.H.; JONES, S.; SOTO-BERELOV, M.; HAYWOOD, A.; HISLOP, S. Landsat time-series for estimating forest aboveground biomass and its dynamics across space and time: a review. **Remote Sensing**, v.12, art.98, 2020. DOI: https://doi.org/10.3390/rs12010098.

PLOTON, P.; MORTIER, F.; RÉJOU-MÉCHAIN, M.; BARBIER, N.; PICARD, N.; ROSSI, V.; DORMANN, C.; CORNU, G.; VIENNOIS, G.; BAYOL, N.; LYAPUSTIN, A.; GOURLET-FLEURY, S.; PÉLISSIER, R. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. **Nature Communications**, v.11, art.4540, 2020. DOI: https://doi.org/10.1038/s41467-020-18321-y.

RODRIGUEZ-GALIANO V.F.; LUQUE-ESPINAR, J.A.; CHICA-OLMO, M.; MENDES, M.P. Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. **Science of the Total Environment**, v.624, p.661-672, 2018. DOI: https://doi.org/10.1016/j.scitotenv.2017.12.152.

SAFARI, A.; SOHRABI, H.; POWELL, S.; SHATAEE, S. A comparative assessment of multi-temporal Landsat 8 and machine learning algorithms for estimating aboveground carbon stock in coppice oak forests. **International Journal of Remote Sensing**, v.38, p.6407-6432, 2017. DOI: https://doi.org/10.1080/01431161.2017.1356488.

SALINAS-MELGOZA, M.A.; SKUTSCH, M.; LOVETT, J.C. Predicting aboveground forest biomass with topographic variables in human-impacted tropical dry forest landscapes. **Ecosphere**, v.9, e02063, 2018. DOI: https://doi.org/10.1002/ecs2.2063.

SCOLFORO, J.; THIERSCH, C. **Biometria florestal**: medição, volumetria e gravimetria. Lavras: UFLA\FAEPE, 2004. 285p.

SILVA, S.H.G.; RIBEIRO, B.T.; GUERRA, M.B.B.; CARVALHO, H.W.P. de; LOPES, G.L.; CARVALHO, G.S; GUILHERME, L.R.G.; RESENDE, M.; MANCINI, M.; CURI, N.; RAFAEL, R.B.A.; CARDELLI, V.; COCCO, S.; CORTI, G.; CHAKRABORTY, S.; LI, B.; WEINDORF, D.C. pXRF in tropical soils: methodology, applications, achievements and challenges. **Advances in Agronomy**, v.167, p.1-62, 2021. DOI: https://doi.org/10.1016/bs.agron.2020.12.001.

SILVEIRA, E.M.O.; SILVA, S.H.G.; ACERBI-JUNIOR, F.W.; CARVALHO, M.C.; CARVALHO, L.M.T.; SCOLFORO, J.R.S.; WULDER, M.A. Object-based random forest modelling of aboveground forest biomass outperforms a pixel-based approach in a heterogeneous and mountain tropical environment. **International Journal of Applied Earth Observation and Geoinformation**, v.78, p.175-188, 2019. DOI: https://doi.org/10.1016/j.jag.2019.02.004.

SEDDON, A.W.R.; MACIAS-FAURIA, M.; LONG, P.R.; BENZ, D.; WILLIS, K.J. Sensitivity of global terrestrial ecosystems to climate variability. **Nature**, v.531, p.229-232, 2016. DOI: https://doi.org/10.1038/nature16986.

SPEISER, J.L.; MILLER, M.E.; TOOZE, J.; IP, E. A comparison of random forest variable selection methods for classification prediction modelling. **Expert Systems with Applications**, v.134, p.93-101, 2019. DOI: https://doi.org/10.1016/j.eswa.2019.05.028.

TADDESE, H.; ASRAT, Z.; BURUD, I.; GOBAKKEN, T.; ØRKA, H.O.; DICK, Ø.B.; NÆSSET, E. Use of remotely sensed data to enhance estimation of aboveground biomass for the dry Afromontane forest in South-Central Ethiopia. **Remote Sensing**, v.12, art.3335, 2020. DOI: https://doi.org/10.3390/rs12203335.

TAVASOLI, N.; AREFI, H. Comparison of capability of SAR and optical data in mapping forest above ground biomass based on machine learning. **Environmental Sciences Proceedings**, v.5, art.13, 2021. DOI: https://doi.org/10.3390/IECG2020-07916.

WANG, X.; JIAO, H. Spatial scaling of forest aboveground biomass using multi-source remote sensing data. **IEEE Access**, v.8, p.178870-178885, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.3027361.

WANG, J.; LI, W.; CIAIS, P.; BALLANTYNE, A.; GOLL, D.; HUANG, X.; ZHAO, Z.; ZHU, L. Changes in biomass turnover times in tropical forests and their environmental drivers from 2001 to 2012. **Earth's Future**, v.9, e2020EF001655, 2021. DOI: https://doi.org/10.1029/2020EF001655.