










MODELING ABOVEGROUND CARBON STOCK UNDER THE FOREST CANOPY INFLUENCE

Lucas Rezende Gomide^{2*}, Kalill José Viana da Páscoa², Caio Eduardo Vieira Alcantara Silva²,
Evandro Nunes Miranda², Mônica Canaan Carvalho², José Roberto Soares Scolforo² and
Carlos Rogerio de Mello²

1 Received on 21.02.2024 accepted for publication on 14.08.2024.

2 Universidade Federal de Lavras, Escola de Ciências Agrárias de Lavras, Departamento de Ciências Florestais, Lavras, Minas Gerais - Brasil. E-mail: <lucasgomide@ufla.br>, <kalill.pascoa@ufla.br>, <caioeduardovieira@gmail.com>, <evandromiranda.florestal@gmail.com>, <monicacanaan@gmail.com>, <josescolforo@ufla.br> and <crmello@ufla.br>.

*Corresponding author.

ABSTRACT

The dominance, size, and sociological position of tree species are essential components of natural forest patterns, providing critical information for understanding forest dynamics. These patterns can be influenced by a variety of factors. This study aimed to evaluate the effectiveness of using spectral, hydrological, and geographical variables to estimate aboveground carbon stock (AGC) across four canopy strata, defined by diametric percentiles, in a Brazilian Atlantic Forest remnant. Our methodology, which employed machine learning techniques (Random Forest - RF + Genetic Algorithm - GA) and Multiple Linear Regression (MLR) to model AGC, proved to be highly efficient, as evidenced by our results. We observed a wide range of AGC values, from 0.37 to 467.71 MgC.ha⁻¹, with an average of 77.4 MgC.ha⁻¹. Trees in the 30th, 60th, and 90th percentiles contributed, respectively, 97.32%, 87.74%, and 52.02% of the total AGC. Spectral and hydrological variables combined with basal area explain AGC stock. Our findings demonstrate the robustness of machine learning techniques and MLR methods in obtaining accurate carbon estimates and generating an optimized dataset. Trees within the 30th percentile represent a smaller portion of the total AGC, and their removal does not interfere with the relationship between AGC and spectral variables.

Keywords: Remote sensing; Forest hydrology; Forest management

How to cite:

Gomide, L. R., Páscoa, K. J. V. da, Silva, C. E. V. A., Miranda, E. N., Carvalho, M. C., Scolforo, J. R. S., & Mello, C. R. de. (2024). Modeling aboveground carbon stock under the forest canopy influence. *Revista Árvore*, 48(1). <https://doi.org/10.53661/1806-9088202448263778>



MODELAGEM DO ESTOQUE DE CARBONO ACIMA DO SOLO SOB INFLUÊNCIA DO DOSSEL DE UMA FLORESTA

RESUMO – A dominância, tamanho e posição sociológica das espécies arbóreas são aspectos cruciais para os padrões naturais das florestas, fornecendo informações de grande relevância. Esses padrões podem ser influenciados por uma série de fatores. Assim, o objetivo deste estudo foi avaliar o uso de variáveis espectrais, hidrológicas e geográficas para estimar o estoque de carbono acima do solo (AGC) em quatro estratos de dossel definidos por percentis diamétricos em um remanescente de Mata Atlântica brasileira. Utilizamos técnicas de aprendizado de máquina (Random Forest - RF + Algoritmo Genético - GA) e Regressão Linear Múltipla (MLR) para modelar o AGC. Nossos resultados indicam uma alta variação nos valores de AGC nas parcelas, variando de 0,37 a 467,71 MgC.ha⁻¹, com média de 77,4 MgC.ha⁻¹. Os conjuntos formados por árvores nos percentis 30, 60 e 90 contribuíram, respectivamente, com 97,32%, 87,74% e 52,02% do AGC total. As variáveis espectrais e hidrológicas se associam à área basal para explicar o estoque de AGC. Nossas descobertas comprovam a eficácia de ambos os métodos na obtenção de estimativas precisas de carbono e na geração de um conjunto de dados otimizado. Árvores no percentil 30 representam uma parte menor do AGC total, e a remoção dessas árvores não interfere na relação entre AGC e variáveis espectrais.

Palavras-Chave: Sensoriamento remoto; Hidrologia florestal; Manejo florestal

1. INTRODUCTION

The assessment of carbon stocks has emerged as a fundamental component in forest conservation and climate change mitigation (Seddon et al., 2020). Deforestation in tropical forests increases atmospheric carbon dioxide levels (Swamy et al., 2023). Projections suggest that measures to address these issues could significantly reduce carbon emissions, potentially contributing 84% of the mitigation in tropical regions and 66% of global mitigation by 2055 (Austin et al., 2020). Under these circumstances, tropical forests are significant for global carbon sink.

The growth rates of trees, species diversity, and wood density are directly influenced by local climatic factors (Blanc et al., 2009). According to Lambers and Oliveira (2019), air temperature, soil nutrient supply, water availability, and light/radiation duration control the physiological rules of trees. These environmental variables affect the ecological traits of trees over time and drive carbon accumulation dynamics in tropical forests. Inter-tree competition is another way to reduce carbon assimilation and plant carbon balance. The structure of the forest canopy may be a critical factor in underlining the dynamics of carbon accumulation. The forest dossel covers a group of strata that distinguishes a set of functional trees, maximum height, and light requirements (Fischer et al., 2014).

Due to the passive optical sensors suffering from saturation, the canopy structure associated with optical sensor images has not been applied to predict the aboveground carbon stock (AGC) in tropical forests so often (Zolkos, 2013). This limitation generally affects the accuracy of prediction methods over high-density forests (Fang et al., 2012). Knapp et al. (2018) applied three approaches to predict biomass changes for a tropical lowland rainforest over time. Their results provide insight into the relationship between canopy height and biomass change at large scales. Instead of using any specific research method, it is possible to associate optical sensor images with forest plots to estimate Aboveground Carbon (AGC) for large-scale areas in tropical forests. Therefore, hydrological variables such as soil water storage (SWS) and throughfall (TF) also play a fundamental role in canopy recovery, leaf growth during the dry season for seasonal deciduous trees, and species occurrence under the environmental filter theory, contributing to understanding the soil-plant-atmosphere system (Bonnesoeur et al., 2019). This ecological perspective corroborates that hydrological variables drive the forest occurrence and the richness of tree species distribution. According to Fonseca et al. (2024), the Atlantic Forest has changed over the last two decades in terms of carbon stock. Their modeling procedure revealed a positive correlation between average annual precipitation, successional stage, and carbon.

Models are a single representation of the real world, and ecological traits are difficult to summarize into a function. In this context, machine learning algorithms overcome the

complex behavior of the database nature, and most of them have high flexibility and accuracy (Chen et al., 2019; Blanco et al., 2020). These algorithms are robust and capable of handling any variable type, dataset size, and linearity (Rokach, 2016; Silveira et al., 2019). Zhang et al. (2019) and Dang et al. (2019) have applied a random forest algorithm to predict biomass/carbon stock in Asian tropical forests. They highlighted the high potential of method accuracy and the use of satellite imagery variables. Nowadays, there is a set of papers applying the combination of genetic algorithm and random forest (GARF) to improve the random forest performance, as seen in Chen et al., (2019) and Miranda et al. (2022). Previously to the random forest runs, the genetic algorithm contribution lies in the selection variables. This automatic selection procedure performs many variable combinations to improve the random forest work. In the current context, we have investigated the response of environmental and spectral variables to explain the variation of AGC across a secondary semi-deciduous Atlantic Forest. Our findings addressed two hypotheses: (i) the canopy strata levels may be a critical factor in underlining the carbon, and (ii) the behavior of the variable type is symmetric or asymmetric over the canopy strata levels.

2. MATERIAL AND METHODS

2.1 Study area and database

The study area covers 6.35 ha of a secondary Semi-deciduous Atlantic Forest located at 44°58'15" W and 21°13'42" S, an altitude of 900m, in Lavras, Minas Gerais state, Brazil (Figure 1). The predominant soil is Dystrophic Red Latosol, and Cwa is the climate according to the Köppen classification with rainy summers and dry winters (Rodrigues et al., 2021). The wet season ranges from October to March and the dry season extends from April to September. The mean annual precipitation and potential evapotranspiration are 1,462 mm and 1,254 mm, respectively, with a yearly mean temperature of 20.3°C, ranging from 16.9°C in July to 22.8°C in February (Rodrigues et al., 2021). The area has reached an advanced successional stage following complete protection in 1986. The forest canopy structure consists of three layers: an emergent layer (crowns of isolated trees over 20 meters high), the middle canopy (crowns

of trees between 10 and 20 meters high), and the understory layer, comprising small trees, seedlings, and bushes (Rodrigues et al., 2022).

We included all trees with DBH \geq 5cm and their respective heights to represent the forest canopy structure. In this forest census, the x-y coordinates of every single tree were mapped based on a Cartesian plane. We applied azimuth to recover their geographic coordinates within the sample's vertices. In the field, a continuous mesh of 100 m² (10x10m) was done to guide the forest inventory and the spatial location. It covers 86.5% of the study area (Figure 1). The AGB was predicted by the Chave et al. (2014) equation for tropical forests, which considers DBH of trees, wood density, and a parameter for environmental stress calculated from the geographical coordinates of each plot as inputs. The variable response of Wood density is derived from the getWoodDensity function. The BIOMASS package (Rejou-Mechain et al., 2017) for R version 4.0.3 was applied to calculate the final value of AGC. The carbon fraction of biomass had a default value of 0.471, corresponding to the individual tree carbon weight. The overall value of AGC from trees within each plot was extracted to hectares (MgC.ha⁻¹).

2.2 Independent variables

We derived AGC estimates from four distinct groups of independent variables: spectral, hydrological, geographic, and forest yield variables. Image data were acquired from the MSI/Sentinel-2A satellite. Two acquisitions were made annually (dry season - July and wet season - November) to mitigate the seasonal deciduous effects of some tree species. A total of ten spectral bands were applied with two spatial resolutions: 10m – encompassing Blue, Green, Red, and Near-infrared (NIR) bands and 20m – comprising Red Edges 1 to 4 and Shortwave Infrared (SWIR) bands 1 and 2. Subsequently, we have used eight vegetation indices: Simple Ratio (SR), Normalized Difference Vegetation Index (NDVI), Soil Adjusted Difference Vegetation Index (SAVI), Atmospherically Resistant Vegetation Index (ARVI), Soil and Atmospherically Resistant Vegetation Index (SARVI), Enhanced Vegetation Index (EVI), Triangular Vegetation Index (TVI), and Visible Atmospherically Resistant Vegetation Index (VARI).

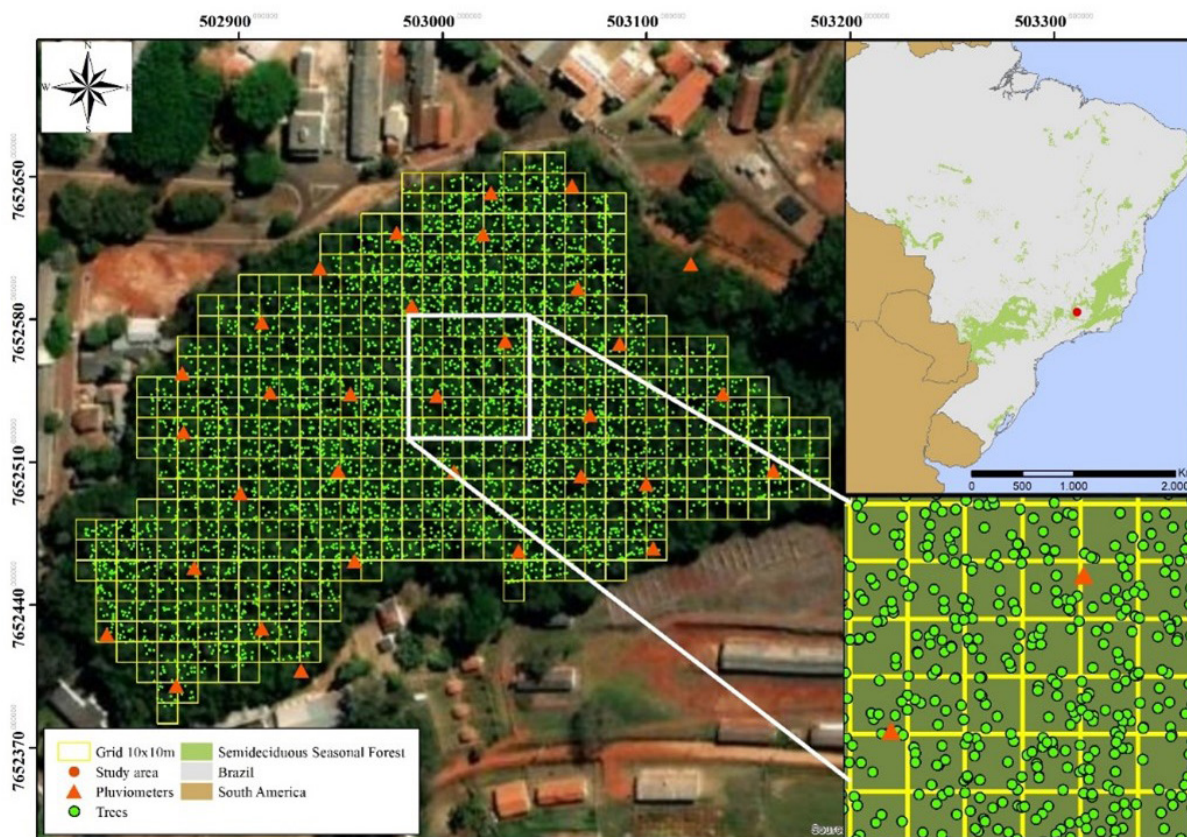


Figure 1. Study area localization and spatial distribution of grids, trees and pluviometers. Coordinate System SAD 1969 UTM Zone 23S

Figura 1. Localização da área de estudo e distribuição espacial de parcelas, árvores e pluviômetros. Sistema de Coordenadas SAD 1969 UTM Zona 23S

Throughfall (TF) was quantified by 32 “Ville de Paris” rain gauges strategically deployed across the study area. They were positioned above ground level at 1.50 m and possessed a catching area of 378.5 cm². Soil moisture storage (SWS) was evaluated employing 32 cylindrical tubes, each with a length of 1.0 m, coupled with a multi-sensor capacitance Profile Probe (PR2/6 capacitance probe, Delta-T Devices, Cambridge, UK). These measurements represent monthly averages for the two years preceding the census and underwent spatial interpolation. The spherical semivariogram and Ordinary Kriging techniques were used to predict these variables spatially within the geoR package (Ribeiro Junior et al., 2020). Basal area (G) (m².ha⁻¹), defined as the cross-sectional area of trees at breast height, served as a stand-level variable for each plot.

The forest yield selected variable was the basal area, the cross-sectional area of trees (g) that describes the stand density/stock or competition level. We have defined four

diametric percentiles (0, 30th, 60th, and 90th) to infer each portion of the forest canopy. We incorporated these diametric percentiles to investigate the influence of the canopy on biomass/carbon stock. In these canopy fractions, the upper accumulative percentile level represents the sum of all trees larger than the DBH position. The 0th percentile encompasses all trees, while the 90th percentile exclusively incorporates those exceeding the 90th position (i.e., the largest trees). The basal area associated with a particular percentile served as an independent variable within the model fitting corresponding to that percentile.

2.3 Aboveground carbon stock modeling strategies

The dataset was split systematically according to the forest structure and the original data distribution into training (80%) and validation (20%) sets. Later, Pearson’s correlation coefficient (r) was applied between

all variables to explore their pattern. A total of two modeling procedures were fitted for each percentile (0th, 30th, 60th, and 90th) to describe the relationship between the predictive variables and AGC stock. The first option was Multiple Linear Regression (MLR) within the Stepwise method, an iterative process that selects a better set of predictive variables to improve the model performance. Finally, the variance inflation factor ($VIF < 5$) was checked to identify the collinearity in the final model.

In addition, we have employed a hybrid methodology known as GA+RF (Genetic Algorithm + Random Forest) to address the complexities of AGC modeling and overcome the limitations of linear regression, particularly in managing multiple variables, noise, non-linear characteristics, and high-dimensional databases (Freitas et al., 2020). In this approach, we integrated the GA to select an optimal subset of variables (feature selection technique) to improve and automate the RF performance. The multi-objective fitness function (Equation 1) was the ratio between the Random Forest out-of-bag error (OOB error) and the maximum enthalpy of OOB error (errorOOBmax). The second term of this function relies on the ratio between the number of selected variables by GA (n) and the total number of tested variables (NVT).

Previously, the GARF tuning was applied to extend the algorithm's performance. The following parameters were set: a) Genetic Algorithm: population size (100), selection rate (0.5), mutation rate (0.1), selection operator (tournament), crossover operators (1 cutoff point), and stopping criteria (10 generations); b) Random Forest: number of trees (ntree: 50), number of predictor variables sampled randomly in each tree division (mtry: 2), and minimum number of samples within terminal nodes (nodesize: 5). The modeling processing was performed in R software and the randomForest package (Liaw & Wiener, 2002). The hardware was an Intel (R) Core™ i7-7500U with a processor running at 2.90 GHz and 8.0 GB of installed RAM.

$$fitness = \left(\frac{erroOOB}{erroOOB_{max}} + \left(\frac{n}{NVT} \right) \right) \quad (\text{Eq. 1})$$

3. RESULTS

The forest remnant contains 422.62 MgC across the 5.46 ha. The AGC values varied from 0.37 to 467.71 MgC.ha⁻¹ considering 546 contiguous samples. The mean AGC values for each percentile were 77.40 MgC.ha⁻¹ (0th) percentile, 75.33 MgC.ha⁻¹ (30th), 67.91 MgC.ha⁻¹ (60th), and 40.26 MgC.ha⁻¹ (90th) (Figure 2).

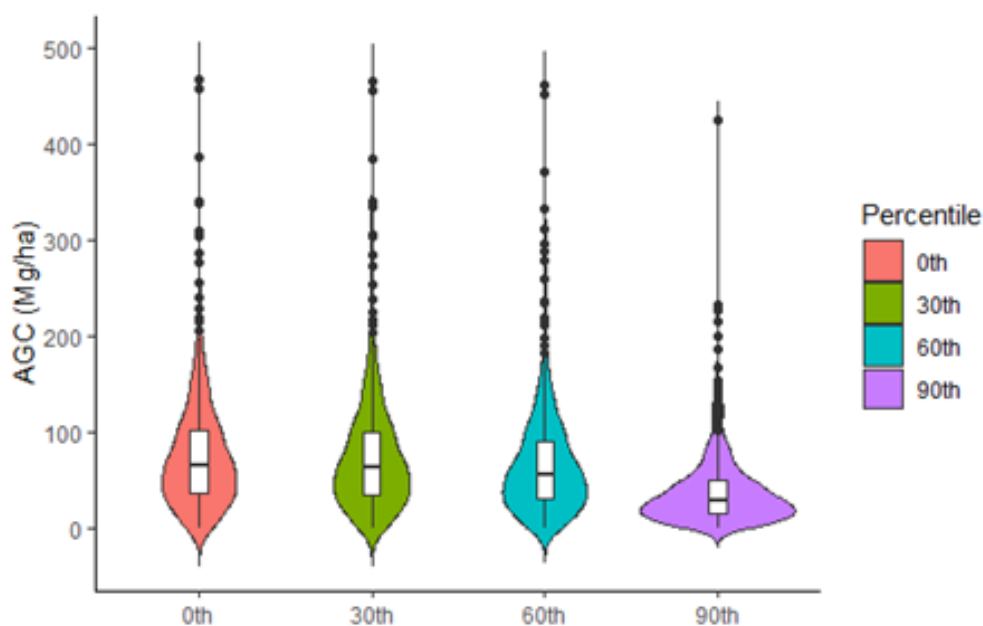


Figure 2. Boxplot with density of aboveground Carbon stock values (AGC) at each percentile (0th,30th, 60th and 90th)

Figura 2. Boxplot com a densidade dos valores de estoque de carbono acima do solo (AGC) em cada percentil (0°, 30°, 60° e 90°)

The relationship between independent variables and AGC stocks on the diametric percentiles is in Table 1. The values highlight a strong correlation between basal area (G) and AGC (0.97-0.98), as expected for biological reasons and the tree's size, and the remaining 43 variables showed low Pearson correlations (< 0.15). Spectral variables derived from the wet season (November) have slightly stronger correlations with AGC than those from the dry season (July) due to the forest seasonality. The SWIR1 (shortwave infrared) exhibited the highest correlation with AGB stock among all spectral variables (0th and 30th percentiles).

Furthermore, the spectral bands red, red edge 3 and 4 (RE3 and RE4), SWIR2, NIR, and all vegetation indices (excluding VARI) displayed significant Pearson correlations with AGC. This pattern was more consistent for the 0th and 30th percentiles during the wet season (0.12 - 0.14). This analysis revealed that the forest's spectral response presents a similar trend between the 0th and 30th percentiles. However, the correlations decreased slightly for the 60th and 90th percentiles. Hydrological (TF and SWS) and geographic variables exhibited no significant correlation with AGC.

Table 1. Pearson correlation's for aboveground Carbon (AGC) stock in tested percentiles and all independent variables

Tabela 1. Correlações de Pearson para estoque de carbono acima do solo (AGC) nos percentis testados e todas as variáveis independentes

Variables	AGCp0		AGCp30		AGCp60		AGCp90	
X	-0,09		-0,09		-0,09		-0,06	
Y	-0,08		-0,1		-0,01		-0,02	
Gp0	0,97***		-		-		-	
Gp30	-		0,97***		-		-	
Gp60	-		-		0,98***		-	
Gp90	-		-		-		0,98***	
Spectral	jul	nov	jul	nov	jul	nov	jul	nov
blue	-0,05	-0,1	-0,05	-0,1	-0,04	-0,09	0,01	-0,01
green	-0,01	0,04	-0,004	-0,04	-0,004	-0,03	0,03	-0,02
red	-0,09	-0,12**	-0,08	-0,12**	-0,07	-0,11	-0,02	-0,04
RE1	-0,06	-0,1	-0,06	-0,1	-0,05	-0,09	0	-0,03
RE2	-0,05	0,11	0,06	0,1	-0,06	0,1	0,04	0,04
RE3	-0,06	0,14**	0,06	0,14**	0,06	0,13**	0,04	0,05
RE4	-0,05	0,13**	0,06	0,13**	0,06	0,12**	0,03	0,03
SWIR1	-0,06	-0,15***	-0,05	-0,15***	-0,04	-0,13**	0,02	-0,03
SWIR2	-0,12**	-0,12**	-0,12**	-0,12**	-0,1	-0,1	-0,01	-0,02
NIR	-0,07	0,12**	0,07	0,11**	0,07	0,11	0,06	0,06
RS	0,1	0,14**	0,1	0,14**	0,098	0,12**	0,052	0,043
NDVI	0,13**	0,14**	0,12	0,14**	0,12**	0,13***	0,072	0,047
SAVI	0,13**	0,14**	0,12	0,14**	0,088	0,13**	0,072	0,047
ARVI	0,1	0,13**	0,1	0,12**	0,088	0,11**	0,031	0,038
SARVI	0,1	0,13**	0,1	0,12**	0,008	0,11**	0,031	0,038
EVI	0,1	0,14**	0,1	0,14**	0,099	0,13**	0,059	0,051
TVI	0,1	0,13**	0,1	0,13**	0,1**	0,12**	0,071	0,062
VARI	0,07	0,11**	0,067	0,11	0,055	0,097	-0,005	0,015
Hydrolo- gical	Year1	Year2	Year1	Year2	Year1	Year2	Year1	Year2
TF	0,089	0,04	0,088	0,038	0,085	0,039	0,11	0,051
SWS	0,029	0,054	0,027	0,051	0,028	0,043	-0,04	-0,02

Regardless of the modeling approach and forest canopy strata, feature selection procedures (whether based on statistical or computational criteria) consistently prioritize basal area (G) and at least one spectral variable for AGC stock estimation (Table 2). Despite lacking a significant correlation with AGC, hydrological variables (TF and SWS) and geographics variables (X and Y) were selected in some models. The high frequency of geographical variables was observed in the 0th, 30th, and 60th percentiles for MLR.

In contrast, GARF selected a hydrological variable for the 90th percentile. Regardless of the modeling process, the final models have high accuracy to predict the AGC. As expected, the GA+RF method performed superiorly for all sets of percentiles in the training dataset. Conversely, MLR has changed this behavior for validation datasets in 30th, 60th, and 90th percentiles. Additionally, the metrics RMSE (%) and R² (%) were lower as the percentile increases, meaning that the canopy strata gradient affects the model's predictability.

Table 2. The statistical of the models for training and validation set for each tree stratum level

Tabela 2. As estatísticas dos modelos para conjunto de treinamento e validação para cada nível de estrato arbóreo

Percentiles	Model	Database	RMSE (%)	R ² (%)	N	Selected variables	Time (s)
0th	MLR *	Training	17.63	94.92	6	G0 + VARI.11 + SWS2 + Y + X + RE2.11	1.87
		Validation	17.82	94.77			
	GA+RF	Training	8.02	98.81	3	G0 + re4.07+ ndvi.07	206
		Validation	17.97	95.63			
30th	MLR *	Training	16.63	95.48	5	G30 + swir.11 + Y + SWS2 + X	0.99
		Validation	16.76	95.36			
	GA+RF	Training	10.04	98.23	2	G30 + SArVi.11	246
		Validation	27.13	90.06			
60th	MLR *	Training	15.77	95.94	5	G60 + X + VARI.11 + SWS2 + Y	1.21
		Validation	15.99	95.77			
	GA+RF	Training	14.02	96.55	2	G60 + RE1.07	186
		Validation	28.85	88.77			
90th	MLR *	Training	38.72	75.49	2	G90 + swir.11	1.44
		Validation	38.82	74.66			
	GA+RF	Training	17.62	90.55	3	G90 + SAVI.07 + TF1	190
		Validation	39.88	69.25			

The spatial distribution of the best AGC estimation method is presented in Figure 3. This map reveals an overview of the predictive success rate for the tested methods, highlighting which method reaches a closer estimation of the observed AGC values. The absolute error showed that the GA+RF method achieved a greater accuracy in the three percentile levels (0th – 80%, 30th – 59%, and 90th – 62%).

The MLR exhibited superior performance for the 60th percentile (52%). These results are corroborated by the model residuals, which indicated a balanced distribution along the entire axis for training (Figure 4a-d) and validation (Figure 4e-h) datasets. The 90th percentile was an exception in both datasets.

Regardless of the method, the distribution error of the training dataset revealed a higher density of values in the range of -50 to 50 MgC.ha⁻¹, except for 90th and samples with high AGC stocks. The dataset for validation showed inferior accuracy for the canopy strata levels (from bottom to top gradient). Overall, there is a slight bias toward underestimation of AGC stock for both methods, being more visible in MLR.

4. DISCUSSION

The mean AGC stock exhibits high variability in secondary forests across tropical areas. Our results (77.40 MgC.ha⁻¹

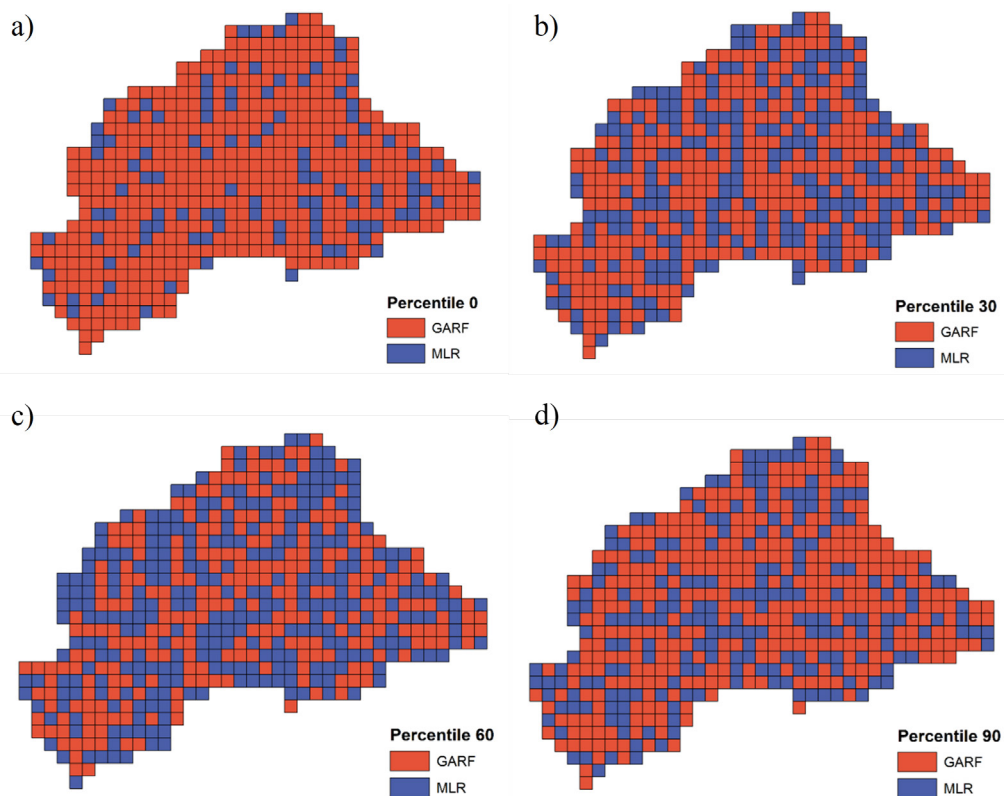


Figure 3. Spatial accuracy distribution within plots and percentiles for the best fitted model, being: a) Percentile 0th, b) Percentile 30th, c) Percentile 60th, and d) Percentile 90th

Figura 3. Distribuição da precisão espacial dentro das parcelas e percentis para o melhor modelo ajustado, sendo: a) Percentil 0°, b) Percentil 30°, c) Percentil 60° e d) Percentil 90°

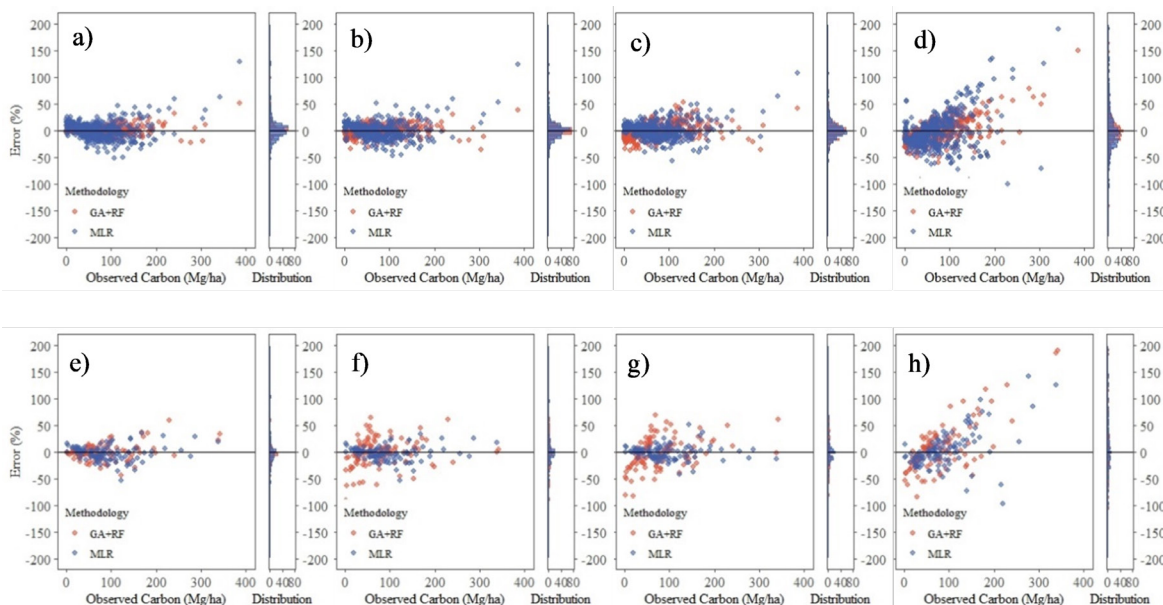


Figure 4. Residuals plots with marginal histograms for training (a-d) and validation (e-f) datasets, where a) and e) – 0th; b) and f) – 30th; c) and g) – 60th; d) and f) – 90th percentile

Figura 4. Gráficos de resíduos com histogramas marginais para conjuntos de dados de treinamento (a-d) e validação (e-f), onde a) e e) - 0°; b) e f) - 30°; c) e g) - 60°; d) e f) - percentil 90°

) are consistent with other studies on Semi-Deciduous Seasonal Forests, such as 83.34 MgC.ha⁻¹ (Ribeiro et al., 2009) and 71.81 MgC.ha⁻¹ (Figueiredo et al., 2015) in Viçosa-MG, and 55 MgC.ha⁻¹ in average for forest fragments located in Minas Gerais state (Scolforo et al., 2015). The canopy strata levels have a carbon pool rate of 97.32%, 87.74%, and 52.02% for the 30th, 60th, and 90th percentiles, respectively. The two intermediate canopy levels explain the greatest carbon stock in this forest. Conversely, the dominant trees of each sample may also reflect a significant portion of the carbon stock. According to Barros et al. (2022), these trees have a lower density in the forest distribution but contribute as the primary carbon sinks.

The correlation values between spectral variables and AGC stock were lower than other studies (Chen et al., 2019; Bucha et al., 2021; Macave et al., 2022). Most of these studies considered large areas with no continuous samples. In our study, we found that the neighborhood of trees, which may extend their sample limits (10x10m) into other samples, plays a significant role in the correlation values. This is because the canopy of the trees could inflate (positive/negative) the correlation values for each sample. The second factor is associated with the challenge of directly linking field-level vegetation data to satellite data, primarily due to positional uncertainty (Lu et al., 2016; Ploton et al., 2020). This uncertainty arises due to the collection of field and satellite data with varying geographic accuracies, potentially resulting in misalignment between the two data sources. Then, it becomes challenging to correlate and integrate the information accurately at the pixel level, impacting the reliability of analyses and conclusions.

The alternative to this integration is to improve the geospatial precision of field data with the use of high-precision GPS and the application of geometric corrections to satellite images involving the use of known control points. Consequently, pixel-based approaches often exhibit lower accuracy than object-oriented approaches (such as fragments, grids, or polygons derived from image segmentation) (Silveira et al., 2019). Furthermore, a well-recognized issue of pixel saturation exists, wherein pixel spectral reflectance values fail to capture changes in biomass for dense and multilayer canopy forests, leading to reduced accuracy in AGB estimation for values

exceeding 100-150 Mg.ha⁻¹ (Lu et al., 2016).

Considering the conversion of tree biomass to carbon content (0.471), the AGC values between approximately 50 and 70 MgC.ha⁻¹ are prone to pixel saturation. In our study area, 62% of samples exhibited values exceeding 50 MgC.ha⁻¹. In the future, efforts to reduce pixel saturation should focus on integrating additional variables such as texture measures and using active sensors like Synthetic Aperture Radar (SAR) and Light Detection and Ranging (Lidar) to establish a stronger relationship with AGC data (Lu et al., 2016; Gosh & Behera, 2018; Tadesse et al., 2020). The correlation values of spectral variables (the red bands, red edges 3 and 4, SWIR1, SWIR2, NIR, vegetation indices except for VARI) highlighted the importance of the red region and longer wavelengths (near-infrared and shortwave infrared bands) in AGC modeling (Lu et al., 2016). Shortwave infrared bands (SWIR1 and SWIR2) mitigate saturation effects by being less sensitive to atmospheric conditions, but the water level of biomass leaf may influence their responses (Zhu and Liu, 2015).

The hydrological variables (TF and SWS) have no significant correlation with AGC stock. However, they were selected at the percentile 90th under the GA+RF method. The canopy closure process of dominant trees may demand more interaction within the soil-water-plant system, which could justify these results. This assessment is supported by Slik et al. (2010), who stated that the water stress limit has a negative effect on biomass growth and correlations with annual rainfall and soil moisture storage. Our findings emphasize this ecophysiology pattern indirectly, where canopy heterogeneity leads to higher precipitation interception by dominant trees (Junqueira Junior et al., 2019). Conversely, light-demanding trees exhibit faster growth rates, develop larger crowns, intercept more precipitation, and allocate more aboveground biomass than shade-tolerant trees (Farrior et al., 2013; Jucker et al., 2014). Non-dominant trees display a high sensitivity to soil water availability, influencing the AGC stock of these trees.

Finally, the number of modeling variables increases the dimensionality of the dataset, which requests a feature selection method to reduce the size and keep the relevant variables to ensure maximum accuracy (Ahmadi, et

al., 2020). Variable selection is relevant in elucidating which factors influence the carbon stock and underlying ecological patterns. However, the machine-learning algorithm may also detect overfitting over the training process. Our findings also detected this issue with GA+RF, capturing the relevant patterns and the noise and random fluctuations in the data. The model generalization usually fails to predict over the training dataset, and the accuracy of the validation dataset is reduced. However, large-scale studies request lower-cost methodologies with high precision, and the integration of geographic, remote sensing, and hydrological variables is occasionally desirable. Furthermore, the multi-temporal and spatial data may also be applied to monitor the AGC in tropical forests.

5. CONCLUSION

Our research suggests that varying levels of canopy cover influence the accuracy of aboveground carbon stock (AGC) models in secondary Semi-deciduous Atlantic Forests. The 90th percentile showed lower precision compared to other tested percentiles. This gradient in accuracy is evident across the canopy layers (bottom to top) in the validation dataset. The geographic, spectral, and hydrologic variables have a minimal impact on AGC stock in our study area. However, both the testing methods and percentile levels highlighted the significance of spectral variables, suggesting that remote sensing indices could be valuable tools for modeling AGC stock.

6. ACKNOWLEDGEMENTS

We would like to thank all the students for their support during the field campaign, and UFLA for providing the institutional facilities. Lucas R. Gomide is supported by research fellowship (grant number 304572/2021-7).

AUTHOR CONTRIBUTIONS

GOMIDE LR: Writing, review, editing, discussion, and data analysis; PASCOA KJV: Writing, review, editing, discussion, and data analysis; ALCANTARA SILVA CEV: Writing, reviewing, editing, and obtaining the spectral database; MIRANDA EM: Model adjustment and data analysis; CARVALHO MC: Review, editing, results analysis and discussion;

SCOLFORO JRS: Review, and editing; DE MELLO CR: Obtaining the hydrological database, review, and editing.

7. REFERENCES

Ahmadi, K., Kalantar, B., Saeidi, V., Harandi, E. K., Janizadeh, S., & Ueda, N.. Comparison of machine learning methods for mapping the stand characteristics of temperate forests using multi-spectral sentinel-2 data. *Remote Sensing*, 2020; 12(18), 3019.

Austin, K G., Baker, J S, Sohngen, B L, Wade, C M, Daigneault, A, Ohrel, S B, Bean, A. The economic costs of planting, preserving, and managing the world's forests to mitigate climate change. *Nature communications*, 2020; 11(1), 5946. doi: 10.1038/s41467-020-19578-z

Barros, J H S, Ayres, F M, Chambó, E D, Constantino, M, Moraes, P M, Skowronski, L, Costa, R B. Aboveground carbon stock in phytophysionomies of the Southeast Pantanal, Brazil. *Brazilian Journal of Botany*, 2022; 45(2), 755-762. doi:10.14483/2256201X.14854

Blanc, L, Echard, M, Herault, B, Bonal, D, Marcon, E, Chave, J, & Baraloto, C. Dynamics of aboveground carbon stocks in a selectively logged tropical forest. *Ecological Applications*, 2009; 19(6), 1397-1404. doi:10.1890/08-1572.1

Blanco, A L T, Lizarazo Salcedo, I A, Rodríguez Eraso, N. Estimación de biomasa aérea de *Eucalyptus grandis* y *Pinus* spp usando imágenes Sentinel1A y Sentinel2A en Colombia. *Colombia forestal*, 2020; v.23, n.1. doi:10.14483/2256201X.14854

Bonnesoeur, V, Locatelli, B, Guariguata, M R., Ochoa-Tocachi, B F, Vanacker, V, Mao, Z, Mathez-Stiefel, S L. Impacts of forests and forestation on hydrological services in the Andes: A systematic review. *Forest Ecology and Management*, 2019; v.433, p.569-584. doi: 10.1016/j.foreco.2018.11.033

Bucha, T, Papčo, J, Sačkov, I, Pajtk, J, Sedliak, M, Barka, Feranec, J. Woody aboveground biomass estimation on abandoned agriculture land using sentinel-1 and sentinel-2 data. *Remote Sensing*, 2021; 13(13), 2488. doi:10.3390/rs13132488

- Chave, J, Réjou-Méchain, M, Búrquez, A, Chidumayo, E, Colgan, M S, Delitti, W B, Vieilledent, G. Improved allometric models to estimate the aboveground biomass of tropical trees. *Global change biology*, 2014; v.20, n.10, p.3177-3190. doi:10.1111/gcb.12629
- Chen L. et al. Assessment of multi-wavelength SAR and multispectral instrument data for forest aboveground biomass mapping using random forest kriging. *Forest Ecology and Management*, 2019; v.447, p.12-25. doi:10.1016/j.foreco.2019.05.057
- Dang, A T N, Nandy, S, Srinet, R, Luong, N V, Ghosh, S, & Kumar, A S. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. *Ecological Informatics*, 2019; v.50, p.24-32. doi:10.1016/j.ecoinf.2018.12.010
- Fang, H. et al. Leaf area index. In: Liang, S; LI, X; Wang, J. (Ed.). *Advanced remote sensing: terrestrial information extraction and applications*. [S.l.]: Academic Press, 2012. cap.11, p.347-381. doi:10.1080/01431161.2013.787707
- Farrior, C E, Dybzinski, R, Levin, S A, Pacala, S W. Competition for water and light in closed-canopy forests: a tractable model of carbon allocation with implications for carbon sinks. *The American Naturalist*, 2013; 181(3), 314-330.
- Figueiredo, L T M D, Soares, C P B, Sousa, A L D, Leite, H G, Silva, G F D. Dinâmica do estoque de carbono em fuste de árvores de uma floresta estacional semidecidual. *Cerne*, 2015; 21, 161-167. doi:10.1590/01047760201521011529
- Fischer, R, Armstrong, A, Shugart, H H., Huth, A. Simulating the impacts of reduced rainfall on carbon stocks and net ecosystem exchange in a tropical forest. *Environmental modelling & software*, 2014; 52, 200-206. doi:10.1016/j.envsoft.2013.10.026
- Fonsêca, N C, Cunha, J S A, Albuquerque, E R D, Lins-E-Silva, A C. B. Carbon stock in aboveground biomass and necromass in the Atlantic Forest: an analysis of data published between 2000 and 2021. *Anais da Academia Brasileira de Ciências*, 2024. 96(1), e20220761. doi: 10.1590/0001-3765202420220761
- Freitas, R, Cavalcanti, J M, Cleger, S, Higuchi, N, Celes, C H, Lima, A. Estimating Amazon carbon stock using AI-based remote sensing. *Communications of the ACM*, 2020; v.63, n.11, p.46-48. doi:10.1145/3416957
- Ghosh, S M, Behera, M D. Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest. *Applied Geography*, 2018; 96, 29-40. doi:10.1016/j.apgeog.2018.05.011
- Jucker, T, Bouriaud, O, Avacaritei, D, Dănilă, I, Duduman, G, Valladares, F, Coomes, D A. Competition for light and water play contrasting roles in driving diversity-productivity relationships in Iberian forests. *Journal of Ecology*, 2014; 102(5), 1202-1213. doi: 10.1111/1365-2745.12276
- Junqueira Junior, J A J, Mello, C R, Mello, J M, Scolforo, H F, Beskow, S, McCarter, J. Rainfall partitioning measurement and rainfall interception modelling in a tropical semi-deciduous Atlantic forest remnant. *Agricultural and Forest Meteorology*, 2019; 275, 170-183. doi:10.1016/j.agrformet.2019.05.016
- Knapp, N, Fischer, R, Huth, A. Linking lidar and forest modeling to assess biomass estimation across scales and disturbance states. *Remote Sensing of Environment*, 2018; 205, 199-209. doi:org/10.1016/j.rse.2017.11.018
- Koppen, W. Klassifikation der klimare nach Temperatur, Niederschlag und Jahreslauf. *Pet. Mitt.*, 1918; 64, 193-203.
- Lambers, H.; Oliveira, R.S. *Plant Physiological Ecology*. 755p. 2019.
- Liaw, A, Wiener, M. Classification and regression by randomForest. *R news*, 2002); 2(3), 18-22.
- 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*, 2016; 9(1), 63-105. doi:10.1080/17538947.2014.990526
- Macave, O A, Ribeiro, N S, Ribeiro, A I, Chauque, A, Bandeira, R, Branquinho, C, Washington-Allen, R. Modelling aboveground biomass of miombo woodlands in Niassa Special Reserve, Northern Mozambique. *Forests*, 2022; 13(2), 311. doi:10.3390/f13020311
- Miranda, E N, Barbosa, B H G, Silva, S H G, Monti, CAU, Tng, DYP, Gomide, LR. Variable selection for estimating individual tree height using genetic algorithm and random forest. *Forest Ecology and Management*, 2022; 504, 119828. doi:10.1016/j.foreco.2021.119828

Ploton, P, Mortier, F, Réjou-Méchain, M, Barbier, N, Picard, N, Rossi, V, Péliissier, R. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 2020; 11(1), 4540. doi:10.1038/s41467-020-18321-y

Réjou-Méchain, M, Tanguy, A, Piponiot, C, Chave, H, Hérault, B. biomass: an R package for estimating above-ground biomass and its uncertainty in tropical forests. *Methods in Ecology and Evolution*, 2017; v.8, n.9, p.1163-1167. doi: 10.1111/2041-210X.12753

Ribeiro Junior, P J, Diggle, P J. geoR: A package for geostatistical analysis. *R-news*, v.1, n.2, p.15-18, 2001.

Ribeiro, S C, Jacovine, L A G, Soares, C P B, Martins, S V, Souza, A L D, Nardelli, A M B. Quantificação de biomassa e estimativa de estoque de carbono em uma floresta madura no município de Viçosa, Minas Gerais. *Revista Árvore*, 2009; v.33, n.5, p.917-926. doi: 10.1590/S0100-67622009000500014

Rodrigues, A F, Mello, C R D, Terra, M D C N S, Beskow, S. Water balance of an Atlantic forest remnant under a prolonged drought period. *Ciência e Agrotecnologia*, 2021; v.45. doi: 10.1590/1413-7054202145008421

Rodrigues, A F, Terra, M C, Mantovani, V A, Cordeiro, N G, Ribeiro, J P, Guo, L, Mello, C R. Throughfall spatial variability in a neotropical forest: Have we correctly accounted for time stability?. *Journal of Hydrology*, 2022; v.608, p.127632. doi: 10.1016/j.jhydrol.2022.127632

Rokach L. Decision forest: Twenty years of research. *Information Fusion*, 2016; v.27, p.111-125. doi:10.1016/j.inffus.2015.06.005

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

Seddon, N, Chausson, A, Berry, P, Girardin, C A, Smith, A, & Turner, B. Understanding the value and limits of nature-based solutions to climate change and other global challenges. *Philosophical Transactions of the Royal Society B*, 2020; 375(1794), 20190120. doi: 10.1098/rstb.2019.0120

Silveira, E M, 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*, 2019; v.78, p.175-188. doi: 10.1016/j.jag.2019.02.004

Slik, J W F, Aiba, S I, Brearley, F Q, Cannon, C H, Forshed, O, Kitayama, K, van Valkenburg, J L. Environmental correlates of tree biomass, basal area, wood specific gravity and stem density gradients in Borneo's tropical forests. *Global ecology and biogeography*, 2010; 19(1), 50-60. doi:10.1111/j.1466-8238.2009.00489.x

Swamy, S L, Darro, H, Mishra, A, Lal, R, Kumar, A, & Thakur, T K. Carbon stock dynamics in a disturbed tropical forest ecosystem of Central India: Strategies for achieving carbon neutrality. *Ecological Indicators*, 2023; 154, 110775. doi:10.1016/j.ecolind.2023.110775

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*, 2020; 12(20), 3335. doi: 10.3390/rs12203335

Zhang, M, Du, H, Zhou, G, Li, X, Mao, F, Dong, L, He, S. Estimating forest aboveground carbon storage in hang-jia-hu using landsat TM/OLI data and random forest model. *Forests*, 2019; v.10, n.11, p.1004, 2019. doi:10.3390/f10111004

Zhu, X, Liu, D. Improving forest aboveground biomass estimation using seasonal Landsat NDVI time-series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015; 102, 222-231. doi:/10.1016/j.isprsjprs.2014.08.014

Zolkos, S G, Goetz, S J, Dubayah, R. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment*, 2013, 128, 289-298. Doi: 10.1016/j.rse.2012.10.017