ABOVEGROUND BIOMASS ESTIMATION IN THE BRAZILIAN AMAZON USING COMBINED LIDAR AND HYPERSPECTRAL DATA

Catherine Torres de Almeida[1], Lênio Soares Galvão[1], Luiz Eduardo de Oliveira Cruz e Aragão[1], Jean Pierre Henry Balbaud Ometto[1], Aline Daniele Jacon[1], Francisca Rocha de Souza Pereira[1], Luciane Yumie Sato[1], Camila Valéria de Jesus Silva[2], Jefferson Ferreira-Ferreira[3] and Marcos Longo[4]

[1]Instituto Nacional de Pesquisas Espaciais - INPE, Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil, E-mails: catherine.almeida@inpe.br, lenio.galvao@inpe.br, luiz.aragao@inpe.br, jean.ometto@inpe.br, alinejacon@hotmail.com, franrspereira@gmail.com, lucianesato@gmail.com; [2]Lancaster Environment Centre, Lancaster University - Bailrigg, Lancaster LA1 4YW, camilaflorestal@gmail.com; [3]Instituto de Desenvolvimento Sustentável Mamirauá, Caixa Postal 38 - 69.553-225 - Tefé - AM, Brasil, jefferson.ferreira@mamiraua.org.br; and [4]Jet Propulsion Laboratory, California Institute of Technology 4800 Oak Grove Dr Pasadena CA 91109 United States, mlongo@jpl.nasa.gov

ABSTRACT

Active Light Detection And Ranging (LiDAR) and passive Hyperspectral Imaging (HSI) remote sensing provide complementary information that can be combined to improve the estimation of vegetation properties, such as aboveground biomass (AGB). Thus, the main objective of this study is to evaluate the combined use of LiDAR and HSI data for estimating AGB in the Brazilian Amazon, by using six regression methods, a high range of remote sensing metrics, and feature selection. To assess the prediction ability of the remote sensing data, single and combined LiDAR and HSI metrics were regressed against AGB from 147 sample plots across the Brazilian Amazon Biome. Overall, the results showed a similar model performance for both LiDAR and HSI single datasets, and for the regression methods used. However, the combination of LiDAR and HSI data improved the AGB estimation accuracy.

Key words **-** imaging spectrometry, laser scanning, machine learning, biomass, tropical forest.

1. INTRODUCTION

Aboveground biomass (AGB) is a key component of the global carbon cycle. Given its importance, there is a growing interest in improving estimates of AGB, allowing accurate monitoring at the landscape scale. Tropical ecosystems, in particular the Amazon forest, have been receiving particular attention because of their critical but still highly uncertain carbon balance [1].

Remote sensing has been used to estimate AGB with different sensors by establishing relationships between field data and optical metrics [2]. Light Detection And Ranging (LiDAR) is promising to characterize complex forest structure because it is less sensitive to signal saturation than passive optical sensors [3]. In contrast, LiDAR has restricted spectral resolution [2], generally covering a single spectral range in the near infrared.

In contrast with LiDAR, hyperspectral imaging (HSI) acquires data in narrow and continuous bands, and thus enables detection of absorption features that are useful for distinguishing different vegetation types and tree species. In addition, hyperspectral instruments provide information on plant stress and biochemical properties [4]. However, their ability to detect vertical structure is limited since the reflectance comes mostly from the upper canopy [5].

Combining structural information provided by LiDAR and spectral information provided by HSI can improve the accuracy of AGB models [3]. Several studies have investigated the potential of combining LiDAR and HSI data for classifying land cover types or tree species [e.g. 6, 7]. However, only a few studies were conducted aiming AGB estimates [e.g. 5, 8, 9, 10]. From these, only Clark et al. [9] and Vaglio Laurin et al. [10] carried out studies in tropical areas from Costa Rica and Sierra Leone, respectively.

Several factors influence the prediction accuracy of remote sensing based AGB models. Examples include the number and type of metrics calculated from remote sensing data and the regression methods. Most of the AGB modeling studies use linear regression statistical models. However, the complex relationship between biomass and remote sensing metrics, sometimes non-linear and other times linear, can be better addressed by non-parametric methods [5]. These models include machine learning techniques such as Support Vector Regression (SVR), Gradient Boosting Machine (GBM), and Random Forest (RF).

This study aims to evaluate the combined use of LiDAR and HSI for estimating AGB in the Brazilian Amazon, by using different regression methods and distinct remote sensing metrics submitted to feature selection procedures.

2. MATERIAL AND METHODS

2.1. Study areas and field data

This study used a dataset from 13 sites in the Brazilian Amazon, distributed along the states of Amazonas, Pará,

Rondônia and Mato Grosso. The sites, composed of primary forests and secondary successions, span different climate conditions, soil types, forest structures, species compositions, and land use histories. Each site has been surveyed with forest inventories, airborne LiDAR and HSI. The field data comprised of a total of 147 plots, of which 124 plots have approximately 0.25 ha and 23 plots have 0.16 ha. Forest inventory data were collected between 2011 and 2017. The AGB for each living tree with DBH (diameter at breast height) ≥ 10 cm was estimated using the pantropical allometric equation of Chave et al. [11] (equation 4). We also took into account the uncertainties of measurements and allometric equation by propagating their errors in a Monte Carlo approach (see details in [12]).

2.2. Remote sensing data

Both LiDAR and HSI data are transects of approximately 300 m x 12.5 km, collected as part of the EBA project (Estimativa de Biomassa na Amazônia, http://www.ccst.inpe.br/projetos/eba-estimativa-de-biomassa-na-amazonia/). LiDAR data were collected between January 2016 and April 2017 using the same airborne discrete-return system (HARRIER 68i Trimble©). The LiDAR sensor recorded multiple returns with a minimum point density of 4 returns m^{-2}, small footprint and a scan angle of 45º. Horizontal accuracy varied among sites from 0.035 m to 0.185 m. Vertical accuracy ranged from 0.07 m to 0.33 m. Firstly, each point cloud was preprocessed by identifying and removing isolated noisy points with the *lasnoise* function (LAStools software [13]). Then, ground points were filtered and interpolated into a 1-m digital terrain model (DTM), using the FUSION/LDV software [14]. To obtain the height above ground for each point, the DTM was subtracted from point elevations. After that, the normalized point clouds were clipped according to each plot spatial extent to further calculate the LiDAR-derived metrics. These metrics included height statistics (maximum, mean, standard deviation, percentiles, coefficient of variation, skewness, and kurtosis), proportion of first returns, point density and leaf area density [15] at different height intervals, structural complexity (Shannon and Simpson indices), and topography (mean DTM elevation). LiDAR metrics (except topography) were calculated using just the first returns, above a 2 m height threshold, to remove near ground points.

Airborne HSI data were collected between September and October 2017, using the AISAFenix sensor (Spectral Imaging©) at an average altitude of 800 m. The sensor acquired images in 361 bands positioned between 380 and 2500 nm. Bandwidth ranged from 5.7 nm to 6.8 nm. The spatial resolution was 1 m. The radiance images were converted into atmospherically corrected surface reflectance data using the Atmospheric/Topographic Correction for Airborne Imagery (ATCOR-4 version 6.3). Data provided by a GPS onboard the aircraft were used for geometric correction. Noisy bands around the two major spectral intervals of water vapor absorption (1400 and 1900 nm) were removed from the analysis. From the 361 bands, 230 bands were left for subsequent statistical analysis.

HSI metrics included the original reflectance bands, 38 vegetation indices, continuum-removal absorption features (band depth, width, area, and asymmetry) centered at five wavelengths (487, 667, 980, 1200, and 2100 nm), and endmember fractions (green vegetation, shade, and non-photosynthetic vegetation/soil) obtained from linear Spectral Mixture Analysis (SMA). All metrics were firstly obtained on a pixel-basis and then converted to the plot-level by calculating the average of all pixels values within the field plot. We also calculated the proportion of pixels with a certain SMA-derived shade fraction (<30%, 30-60%, and >60%) as potential metrics related to canopy structure.

To avoid redundancy, we removed highly correlated metrics (absolute Pearson's correlation ≥ 0.98), and also linear combinations, using the R package *caret* [16]. After this procedure, the number of metrics for modeling was reduced from 45 to 34 for LiDAR and from 296 to 64 for HSI.

2.3. Modeling framework

We tested three predictor datasets for AGB modeling: 34 LiDAR metrics, 64 HSI metrics, and their combination (98 predictors). Six regression algorithms were used, including a parametric model, a Linear Model with elasticnet Regularization (LMR); and five non-parametric machine learning approaches: SVR with linear kernel (SVR_{linear}), polynomial kernel (SVR_{poly}), and radial basis kernel (SVR_{radial}), GBM, and RF. The six techniques were applied to the three datasets using the *rfe* function (*caret* package), which combines feature selection, parameter tuning, and cross-validation.

To evaluate models performance, a 5-fold cross-validation, repeated 10 times (total of 50 resamples), was used to quantify the Root Mean Squared Error (RMSE), the relative RMSE, and the coefficient of determination (R^2). For each training set, the models were trained using all predictors. We estimated the model performance on the k-fold test sets and we ranked the predictors according to an importance criterion [16]. Model parameters were optimized by using a 4-fold cross-validation inside the training set and selecting the parameters with lowest RMSE. Less important features were sequentially removed from the models, until only the most important variable remained. An average cross-validation test error (of the 50 resamples) for each feature subset size was obtained. The optimal subset size selected was the one with the lowest number of predictors and with a low RMSE (whose difference did not exceed 2.5% of the lowest RMSE). This approach selects a simpler model without sacrificing too much performance, by considering a tolerance in the RMSE value. The resulting 18 models (6 regression techniques x 3 data sources) with

reduced feature size were compared based on the distribution of cross-validated RMSE, RMSE% and R^2.

Furthermore, a robust one-way ANOVA (function *t1way* of the R package *WRS2* [17]) for trimmed means of RMSE and R^2, followed by its corresponding post hoc test (function *lincon*), was applied in order to determine significant differences between models at a significance level of 5%. This statistical analysis does not require homoscedasticity, being more robust than the classical ANOVA.

3. RESULTS AND DISCUSSION

Field plots covered a broad range of AGB, from 0 Mg ha^{-1} (no living tree with a DBH ≥ 10 cm) to 542.9 Mg ha^{-1}, with a mean value of 193.3 Mg ha^{-1}. The AGB uncertainty reached a maximum of 65 Mg ha^{-1}, with a mean value of 20.9 Mg ha^{-1} (11% of the AGB mean). This uncertainty is consistent with that reported by Chave et al. [18].

Considering the optimal feature size selected for each model, we compared the different models to assess the effect of the data source and regression methods on the AGB estimation performance (Table 1). The AGB models based on LiDAR data showed mean RMSE around 68 Mg ha^{-1} (RMSE% of 35%) and mean R^2 around 0.67, with no significant difference between the regression methods used. Models based on HSI data performed equally well (no significant difference from LiDAR-models), with the exception of the GBM method, which presented lower performance (RMSE of 71.52 Mg ha^{-1}, RMSE% of 37%, and R^2 of 0.63) than the GBM model with LiDAR data (RMSE of 67.38 Mg ha^{-1}, RMSE% of 35%, and R^2 of 0.68).

The combination of LiDAR and HSI data produced better model performances (RMSE decrease of 7-14 Mg ha^{-1}, relative RMSE decrease of 4-7%, and R^2 increase of 6-14%) than the use of only one data source (solely LiDAR or HSI), for all regression methods tested. The RMSE and R^2 of the models using combined datasets performed consistently better when compared to models using LiDAR-only or HSI-only datasets (p-value < 0.05). No statistically significant difference was observed between the different regression methods with combined data sources, meaning that any tested method used with the combination of LiDAR and HSI data yielded good results.

Previous studies [5, 9] have found lower prediction ability of the HSI-derived AGB models when compared to the LiDAR-derived estimates, and thus the combined LiDAR-HSI AGB models showed slight or no improvements in comparison with LiDAR models. Here, most of the regression techniques showed similar performances when using LiDAR or HSI data, and a significant improvement was observed when using the combined dataset. The broader range of HSI metrics used in this study, in addition to the selection of better features, contributed to the good performance of the HSI and combined models, by exploring the synergy between different vegetation properties (such as canopy structure, water content, leaf biochemical, and plant stress). Hyperspectral data have a large amount of information for AGB modeling, but its potential may be underestimated if only few metrics are considered in the analysis, which was not the case here.

Table 1. Mean cross-validated RMSE, RMSE%, and R^2 for each of the six regression methods and three data sources.

Method	Data	Mean RMSE [Mg ha^{-1}]	[%]	[A]	Mean R^2 [B]
LMR	LiDAR	68.84	35.62	ab	0.66 ab
	HSI	70.66	36.55	ab	0.64 ab
	LiDAR + HSI	60.59	31.35	c	0.74 c
SVR_{linear}	LiDAR	68.93	35.66	ab	0.66 ab
	HSI	67.90	35.13	a	0.67 a
	LiDAR + HSI	57.62	29.81	c	0.77 c
SVR_{poly}	LiDAR	67.81	35.08	a	0.67 a
	HSI	68.73	35.56	ab	0.66 ab
	LiDAR + HSI	58.40	30.21	c	0.76 c
SVR_{radial}	LiDAR	67.12	34.72	a	0.68 a
	HSI	68.93	35.66	ab	0.66 ab
	LiDAR + HSI	58.62	30.32	c	0.75 c
GBM	LiDAR	67.38	34.86	a	0.68 a
	HSI	71.53	37.00	b	0.63 b
	LiDAR + HSI	59.39	30.72	c	0.75 c
RF	LiDAR	67.86	35.11	a	0.67 ab
	HSI	69.26	35.83	ab	0.65 ab
	LiDAR + HSI	58.43	30.23	c	0.76 c

Different letters mean significant difference (p-value < 0.05) between trimmed [A] RMSE mean or [B] R^2 mean.

4. CONCLUSIONS

By optimizing the number of predictors and the model parameters, we found that different regression methods could perform equally well in estimating AGB. Therefore, the prediction method generally did not have a significant effect on the model's performance. Results showed that the performance of the AGB models was improved when LiDAR and HSI data were combined into the data analysis, in relation to the use of only one type of data (LiDAR or HSI). The gain of information observed in the analysis indicated the importance of the synergistic use of both data sources for the AGB estimation in the Brazilian Amazon.

5. ACKNOWLEDGMENTS

This study was partially funded by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) [grant numbers 140502/2016-5 and 305054/2016-3]. Data

from 92 field plots were acquired by the Sustainable Landscapes Brazil project supported by the Brazilian Agricultural Research Corporation (EMBRAPA), the US Forest Service, and USAID, and the US Department of State. We thank the FATE project, the IDSM (Instituto de Desenvolvimento Sustentável Mamirauá), and the LMF/INPA (Laboratório de Manejo Florestal do Instituto Nacional de Pesquisas da Amazônia) for providing part of the field data and the EBA project for the remote sensing data. We also acknowledge Aline P. Lopes for her valuable suggestions.

6. REFERENCES

[1] Ometto, J.P.; Aguiar, A.P.; Assis, T.; Soler, L.; Valle, P.; Tejada, G.; Lapola, D.M. and Meir, P. "Amazon forest biomass density maps: tackling the uncertainty in carbon emission estimates," Clim. Change, v.124, n.3, pp.545–560, 2014.

[2] Lu, D.; Chen, Q.; Wang, G.; Liu, L.; Li, G. and Moran, E. "A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems," Int. J. Digit. Earth, v.9, n.1, pp.63–105, 2016.

[3] Koch, B. "Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment," ISPRS J. Photogramm. Remote Sens., v.65, n.6, pp.581–590, 2010.

[4] Chadwick, K.D. and Asner, G.P. "Organismic-Scale Remote Sensing of Canopy Foliar Traits in Lowland Tropical Forests," Remote Sens., v.8, n.2, pp.1-16, 2016.

[5] Fassnacht, F.E.; Hartig, F.; Latifi, H.; Berger, C.; Hernández, J.; Corvalán, P. and Koch, B. "Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass," Remote Sens. Environ., v.154, pp.102–114, 2014.

[6] Dalponte, M.; Bruzzone, L. and Gianelle, D. "Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data," Remote Sens. Environ., v.123, pp.258–270, 2012.

[7] Zhang, C.; Selch, D. and Cooper, H. "A Framework to Combine Three Remotely Sensed Data Sources for Vegetation Mapping in the Central Florida Everglades," Wetlands, v.36, n.2, pp.201–213, 2015.

[8] Anderson, J.E.; Plourde, L.C.; Martin, M.E.; Braswell, B.H.; Smith, M.L.; Dubayah, R.O.; Hofton, M.A. and Blair, J. B. "Integrating waveform lidar with hyperspectral imagery for inventory of a northern temperate forest," Remote Sens. Environ., v.112, n.4, pp.1856–1870, 2008.

[9] Clark, M.L.; Roberts, D.A.; Ewel, J.J. and Clark, D.B. "Estimation of tropical rain forest aboveground biomass with small-footprint lidar and hyperspectral sensors," Remote Sens. Environ., v.115, n.11, pp.2931–2942, 2011.

[10] Vaglio Laurin, G.; Chen, Q.; Lindsell, J.A.; Coomes, D.A.; Frate, F.D.; Guerriero, L.; Pirotti, F. and Valentini, R. "Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data," ISPRS J. Photogramm. Remote Sens., v.89, pp.49–58, 2014.

[11] 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.; Henry, M.; Martínez-Yrízar, A.; Mugasha, W.A.; Muller-Landau, H.C.; Mencuccini, M.; Nelson, B.W.; Ngomanda, A.; Nogueira, E.M.; Ortiz-Malavassi, E.; Pélissier, R.; Ploton, P.; Ryan, C.M.; Saldarriaga, J.G. and Vieilledent, G. "Improved allometric models to estimate the aboveground biomass of tropical trees," Glob. Chang. Biol., v.20, n.10, pp.3177–3190, 2014.

[12] Réjou-Méchain, M.; Tanguy, A.; Piponiot, C.; Chave, J. and Hérault, B. "Biomass: an R Package for Estimating Above-Ground Biomass and Its Uncertainty in Tropical Forests," Methods Ecol. Evol., v.8, n.9, pp.1163–1167, 2017.

[13] Isenburg, M. "LAStools - efficient LiDAR processing software" (version 171030, unlicensed), obtained from http://rapidlasso.com/LAStools

[14] McGaughey, R.J. FUSION/LDV: Software for LIDAR Data Analysis and Visualization, Manual, USFS Pacific Northwest Research Station, Seattle, Wash. 2014.

[15] Bouvier, M.; Durrieu, S.; Fournier, R.A. and Renaud, J.P. "Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data," Remote Sens. Environ., v.156, pp.322–334, 2015.

[16] Kuhn, M. "Building Predictive Models in R using the caret package," Journal of Statistical Software, v.28, n.5, pp.1-26, 2008.

[17] Mair, P. and Wilcox., R. WRS2: A Collection of Robust Statistical Methods, 2017. URL https://cran.r-project.org/web/packages/WRS2/WRS2.pdf. R package version 0.10-0.

[18] Chave, J.; Condit, R.; Aguilar, S.; Hernandez, A.; Lao, S. and Perez, R. "Error propagation and scaling for tropical forest biomass estimates," Philos. Trans. R. Soc. B Biol. Sci., v.359, n.1443, pp.409–420, 2004.