COMPARISON BETWEEN RANDOM FOREST AND LINEAR REGRESSION FOR TROPICAL FOREST ABOVEGROUND BIOMASS ESTIMATION

Franciel Eduardo Rex [1], Ana Paula Dalla Corte[1], Carine Klauberg[2], Andrew Thomas Hudak[3], Pâmela Suélen Käfer[4], Vanessa Sousa da Silva[5] Carlos Alberto Silva[6-7]

[1]Federal University of Paraná – UFPR, Curitiba, PR, Brazil - 80210-170; francielrexx@gmail.com; anapaulacorte@gmail.com; [2]Federal University of São João Del Rei – UFSJ, Sete Lagoas, MG, Brazil - 35701-970; carine_klauberg@hotmail.com; [3]USDA Forest Service, Rocky Mountain Research Station, 1221 South Main Street, Moscow, ID 83843, USA; ahudak@fs.fed.us; [4]Federal University of Rio Grande do Sul, Brazil, pamelaskafer@gmail.com; [5]Department of Forest Sciences, Federal Rural University of Pernambuco, Rua Dom Manoel de Medeiros, s/n, Dois Irmãos, Recife, PE, Brazil – 52171-900; vsousads@gmail.com; [6]Department of Geographical Sciences, University of Maryland, College Park, Maryland, MD 20740, USA [7]NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA; carlos_engflorestal@outlook.com

ABSTRACT

The objective was to compare two methods for estimating aboveground biomass (AGB) in tropical rainforest using airborne LiDAR data. The study was conducted at Fazenda Cauxi in northern Brazil. Data from LiDAR and field inventory collected in 2014 were used. A total of 85 plots were used for the modeling. In the R environment, Random Forest (RF) and Linear Regression (lm) were compared in terms of RMSE, Bias and adj.R^2 through a LOOCV process with 500 replicates. The best performance was verified for the LM algorithm.

***Key words* —** LiDAR, Machine Learning, Remote Sensing.

1. INTRODUCTION

Among terrestrial ecosystems, forests sequester and store more carbon than any other ecosystem and are an important natural 'brake' on climate change [1-2]. Tropical forests are known as large carbon sinks [3]. Among these Amazon rainforest stores one fifth of the total carbon of global terrestrial vegetation [4], thus representing the largest carbon reservoir in the form of biomass of the planet [4-5].

Due to the high growing potential of tropical forests in converting atmospheric carbon into biomass, especially in comparison to other terrestrial ecosystems, an accurate estimate of the rainforest structure and biomass is essential for the understanding and management of the global carbon cycle [6-7]. However, forest monitoring in tropical regions is a challenge, and field surveys are resource demanding and very limited in extent and frequency [8]. In remote sensing studies, the use of airborne LiDAR data has rapidly become prominent in estimating forest biophysical characteristics, such as canopy height and basal area [9-10]. In addition, airborne LiDAR has been successfully used to estimate above-ground biomass in a number of forest ecosystems [10-13].

Regarding biomass modeling, [14] highlight that there are several methods used to estimate biomass / tree volume, which are varied in their assumptions and complexity, such as regression techniques. The authors also comment that several studies have demonstrated that regression of variables derived from LiDAR and field data is an effective method to estimate biomass. Nevertheless, there is a large set of premises-specific assumptions and considerations that must be made for each study.

In this sense, Machine Learning techniques may be more effective than traditional regression techniques, since ML is a rapidly growing predictive modeling area that is concerned with identifying structures in complex, often non-linear data [15]. This paper proposes to compare one machine-learning method, Random Forests (RF), to multiple linear regression (lm) for tropical forest aboveground biomass estimation by using airborne LiDAR data.

2. MATERIAL AND METHODS

2.1 Study area

The study was conducted at the Fazenda Cauaxi in the Paragominas Municipality of Pará state, Brazil. Pará state is located in the eastern Amazon, where deforestation and logging have been integral parts of the economy for decades [16-17]. The climate on Fazenda Cauaxi is humid tropical, and the total annual precipitation average is 2200 mm [18-19].

2.2. Field Data

A forest inventory was conducted from 18 February to 25 April 2014 [20]. A total of 85 plots of 50 × 50 m (0.25 ha) were spaced at intervals of 100 m along transects. At each plot, all trees with a diameter at breast height (DBH) equal to or greater than 35 cm were measured. Inside each plot, a sub-plot along one side of the plot with dimensions of 5 × 50 m (250 m2) was also demarcated. The Equation 1 was used for estimating AGB at tree level [21].

$$AGB\ (kg) = \exp[a1 - b2 + b2\ln(\rho) + c3\ln(dbh) - d4[\ln(dbh)]^2)] \quad \text{Eq.1}$$

where AGB (kg) is the live tree aboveground biomass in Kg; a1 = -1.803; b2 =0.97; c3 = 2.673; d4 = 0.0299; dbh is the diameter at breast height (1.30 m); ρ is the wood density, and E is a measure of environmental stress. In this study area location E = −0.103815.

2.3. LiDAR data

The airbone LiDAR data used in this study were acquired as part of Sustainable Landscapes Brazil, a joint project of the Brazilian Corporation of Agricultural Research (EMBRAPA) and the United States Forest Service (USFS). For the development of the present research we used field data collected in the same year. Table 1 presents the flight parameters.

Table I. Details of lidar data acquisitions.

Specifications	2014
Acquisition date	December 26th to 27th
Datum	Sirgas 2000
Mean point density	37.5 ppm²
Flying height	850 m
Field of view	12 °
Measurement rate	83,0 Hz
Overlay Percentage	65%

2.4. LiDAR data processing

LiDAR data processing was carried out with the following sequence of steps using the FUSION / LDV toolkit version 3.60 software, which allows the analysis and visualization of LiDAR data, besides being an efficient processing tool [22].

First, the catalog command was used to produce the descriptive report of the LiDAR dataset. Then, the groundfilter command was used to classify the ground points, which is grounded on the filtering algorithm, based on [23]. Afterwards, the Digital Terrain Models (DTM) were generated using the product of the previous step (the classified soil points). In this procedure the grid surface create command was used.

Normalization of the heights was carried out with the ClipData command, while the PolyClipdata command was used to perform the cut of the measured plots in the field. From the CloudMetrics command, the LiDAR metrics derived from the plots were extracted. We selected some metrics for the modeling, the metrics are shown in Table II.

LiDAR metrics are often highly interrelated [24-25], hence we selected potential original metrics extracted from LiDAR point clouds using Principal Components Analysis (PCA). The PCA is based on the variance and covariance of the data set [26].

Table II: LiDAR-derived canopy height metrics considered as potential candidate variables for predictive imputation machine learning models.

Description	
Height Maximum	Height 40th percentile
Height Mean	Height 50th percentile
Height standard deviation	Height 60th percentile
Height skewness	Height 70th percentile
Height kurtosis	Height 80th percentile
Height coefficient of variation	Height 90th percentile
Height mode	Height 95th percentile
Height 25th percentile	Height 99th percentile
Height variance	Canopy Cover (Percentage of first return above
Height Interquartil distance	

2.5. Model Accuracy and Assessment

We used the “lm” linear model function and "RF" random forest algorithm, both available in R [27]. RF is an ensemble learning technique that uses multiple decision trees on a validation set to generate a statistically prediction based on a set of independent variables [28]. While "lm" is used to fit linear models, it can be used to carry out regression, single stratum analysis of variance and analysis of covariance.

The methods presented were tested in the R statistical environment [27]. Root-mean-square error (RMSE), Bias, and adjusted coefficient of determination (adj.R2) were calculated to compare the prediction performance of different approaches. In our case, the coefficients were obtained in a process of Leave-One-Out Cross-Validation (LOOCV).

3. RESULTS

PCA indicated that 97.55% of the total variance in the 22 LiDAR metrics can be explained by the first 6 PCs. Therefore, the following metrics were selected for taking part of machine learning models: (PC1: Elev. Mean; PC2: Elev. CV; PC3: Elev kurtosis; PC4: Canopy Cover; PC5: Elev. Mode; PC6: Elev. skewness). Methods used in this study presented values of adj. R^2 close to 70% through the validation process (LOOCV). The results of the comparison for the methods are seen in (Figure 2).

Although the LM algorithm had the best performance, the two methods compared presented similar performances in terms of RMSE and Bias. The largest difference between the algorithms is verified in adj.R^2, in which RF produced a value close to 0.60% while the LM produced a value of 70%.

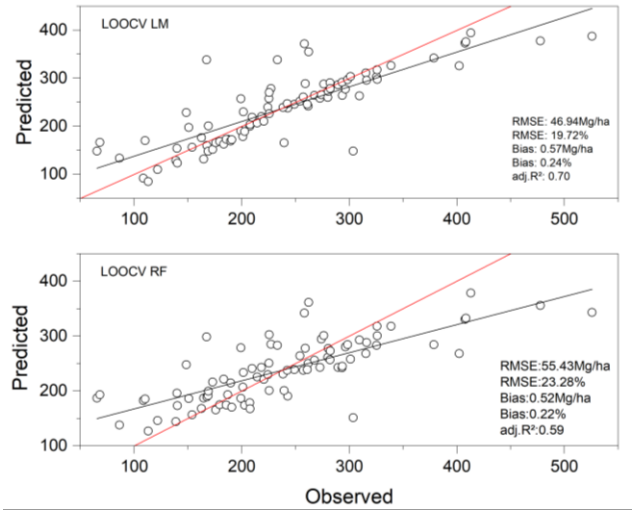

Figure 2: Scatter plots of adj.R2, RMSE and bias for the AGB leave-one-out cross validation – LOOCV models.

4. DISCUSSION

In this paper, the metrics selected to compose the models are in agreement with previous studies [30-31]. In general, the great majority of biomass prediction studies indicate that the mean canopy height metric is the most significant, which corroborates with the PCA result.

Among biomass prediction factors, the prediction method is one of the most important in most cases [30]. RF is powerful for empirical modeling using complex data [32] such as airborne LiDAR data. However, here RF was not able to present superior results in relation to LM. Despite the lower RF performance, the algorithm was similar in Bias values; according to [31] the bias rarely is considered as a parameter to judge the quality of the models in previous studies. It is strongly recommended that this be done, since both r^2 and RMSE can suffer notable offset errors when bias is not considered [33].

A number of studies in different ecosystems found RF to be superior to other methods [8,30,34,35]. Thus, the lower RF performance in this work might have been influenced not only by the number of field plots, but also by other factors, such as a robust bootstrapping of data to avoid overly optimistic r^2 values [30].

5. CONCLUSIONS

The results found in this work show that dealing is a powerful and practical tool for rainforest studies. Among the methods compared, the LM model provided the highest estimation accuracy in terms of higher adj.R^2, the lowest RMSE and the lowest bias.

6. ACKNOWLEDGEMENTS

We thank the USAID, the US Department of State, and EMBRAPA for the support Sustainable Landscapes Brazil program that provided lidar data from Brazilian sites. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES), finance code 001, and also by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

7. REFERENCES

[1] Gibbs, H.K., Brown, S., Niles, J.O. and Foley, J.A. Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environmental Research Letters*, v. 2, n. 4, pp. 045023, 2007.

[2] Thomson, A.M., Calvin, K.V., Chini, L.P., Hurtt, G., Edmonds, J.A., Bond-Lamberty, B., Frolking, S., Wise, M.A. and Janetos, A.C. Climate mitigation and the future of tropical landscapes. *Proceedings of the National Academy of Sciences*, v. 107, n. 46, pp. 19633-19638, 2010.

[3] Hooijer, A., Page, S., Canadell, J.G., Silvius, M., Kwadijk, J., Wösten, H. and Jauhiainen, J. Current and future CO 2 emissions from drained peatlands in Southeast Asia. *Biogeosciences*, v.7, n. 5, pp.1505-1514, 2010

[4] Malhi, Y., Roberts, J. T., Betts, R. A., Killeen, T. J., Li., W., Nobre, C. A. Climate Change, Deforestation, and the Fate of the Amazon. *Science*. 319, n. 5860, p. 169-172, 2008.

[5] Betts, R.A. Forcings and feedbacks by land ecosystem changes on climate change. *Journal de Physique IV*. v. 139, pp. 119-142, 2006.

[6] Marvin, D. C., Asner, G. P., Knapp, D. E., Anderson, C. B., Martin, R. E., Sinca, F., & Tupayachi, R. Amazonian landscapes and the bias in field studies of forest structure and biomass. *Proceedings of the National Academy of Sciences*, v. 111, n. 48, p. E5224-E5232, 2014.

[7] Molina, P. X., Asner, G. P., Farjas Abadía, M., Ojeda Manrique, J. C., Sánchez Diez, L. A., & Valencia, R. Spatially-explicit testing of a general aboveground carbon density estimation model in a western Amazonian forest using airborne LiDAR. *Remote Sensing*, v. 8, n. 1, p. 9, 2015.

[8] Laurin, G.V., Chan, J.C.W., Chen, Q., Lindsell, J.A., Coomes, D.A., Guerriero, L., Del Frate, F., Miglietta, F. and Valentini, R., 2014. Biodiversity mapping in a tropical West African forest with airborne hyperspectral data. *PLoS One*, v. 9, n. 6, p. e97910, 2014.

[9] Hudak, A.T., Evans, J.S. and Stuart Smith, A.M. LiDAR utility for natural resource managers. *Remote Sensing*, v.1, n.4, pp.934-951, 2009.

[10] Asner, G. P., Hughes, R. F., Varga, T. A., Knapp, D. E., & Kennedy-Bowdoin, T. Environmental and biotic controls over aboveground biomass throughout a tropical rain forest. *Ecosystems*, v.12, n.2, pp.261-278, 2009.

[11] 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 Sensing of Environment*, v.115, n.11, pp.2931-2942. 2011

[12] Næsset, E., Gobakken, T., Solberg, S., Gregoire, T.G., Nelson, R., Ståhl, G. and Weydahl, D., 2011. Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: a case study from a boreal forest area. *Remote Sensing of Environment*, v. *115*, v. 12, pp.3599-3614, 2011.

[13] Zolkos, S.G., Goetz, S.J. and Dubayah, R. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment*, v. *128*, pp.289-298, 2013.

[14] Gleason, C.J. and Im, J. Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment*, v. 125, pp.80-91, 2012.

[15] Olden, J. D., Lawler, J. J., & Poff, N. L. Machine learning without tears: a primer for ecologists. *The Quarterly Review of Biology*, v.83, n. 2, 171–193, 2008.

[16] Fearnside, P.M. Deforestation in Brazilian Amazonia: History, Rates, and Consequences. *Conservation Biology*, v.19, n.3, pp.680–688, 2005.

[17] Fearnside, P. M. Deforestation in Amazonia Encyclopedia of Earth ed CJ Cleveland (Washington, DC: Environmental Information Coalition, National Council of Science and the Environment), 2007.

[18] Costa, M. H., & Foley, J. A. A comparison of precipitation datasets for the Amazon Basin. *Geophysical Research Letters*, v. 25, n.2, 155–158. 1998

[19] IBGE – Instituto Brasileiro De Geografia E Estatística. Mapa de Vegetação do Brasil. Ministério da Agricultura, Brasília, Brasil. 1988.

[20] Dos-Santos, M.N., Keller, M., Scaranello, M., Longo, M., Pioto, D. 2016. 18 Years of Recovery: Spatial Variation and Structure of a Secondary Forest Analyzed with Airborne Lidar Data in the Brazilian Atlantic Forest. In: *AGU Fall Meeting*, San Francisco. AGU Abstract, 2016.

[21] Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chambers, J.Q., Eamus, D., Fölster, H., Fromard, F., Higuchi, N., Kira, T., et al. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia*, v. 145, pp. 87–99. 2005.

[22] Mcgaughey, R. J. M. 2016. FUSION/LDV: software for LiDAR data analysisand visualization (version 3.60). Seattle, WA. Disponível em: <http://forsys.cfr.washington.edu/fusion/FUSION_manual.pdf>

[23] Kraus, K; Pfeifer, N. Determination of terrain models in wooded areas with airborne laser scanner data. *ISPRS Journal Of Photogrammetry And Remote Sensing*, v. 53, n. 4, p. 193-203, 1998.

[24] Næsset, E., Bollandsås, O.M. and Gobakken, T. Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. *Remote Sensing of Environment*, v. 94, n. 4, p. 541-553, 2005.

[25] Li, Y., Andersen, H.E. and McGaughey, R. A comparison of statistical methods for estimating forest biomass from light detection and ranging data. *Western Journal of Applied Forestry*, v. 23, n. 4, p. 223-231, 2008.

[26] Smith, L. I. *A tutorial on principal component analysis.* Retrieved April 15, 2007. 2002.

[27] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2015. Available online: https://www.r-project.org/ (accessed on 15 Agust 2018).

[28] Breiman, L.Random forests. *Machine Learning*. v. 45, n. 1, p. 5-32, 2001.

[29] Latifi, H., Fassnacht, F. and Koch, B. Forest structure modeling with combined airborne hyperspectral and LiDAR data. *Remote Sensing of Environment*, v. 121, p. 10-25, 2012.

[30] Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P. and Koch, B., 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, v. 154, p. 102-114, 2014.

[31] Lopatin, J., Dolos, K., Hernández, H.J., Galleguillos, M. and Fassnacht, F.E. Comparing generalized linear models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. *Remote sensing of environment*, v.173, pp.200-210, 2016.

[32] J. Friedman, T. Hastie, R. Tibshirani The elements of statistical learning Springer Series in Statistics, *Springer*, Berlin. 2001.

[33]N. Bennett, B. Croke, G. Guariso, J. Guillaume, S. Hamilton,A . Jakeman, ..., V.Andreassian Characterising performance of environmental models. *Environmental Modelling & Software*, v. 40, pp. 1-20, 2013.

[34] Fayad, I., Baghdadi, N., Guitet, S., Bailly, J.S., Hérault, B., Gond, V., El Hajj, M. and Minh, D.H.T. Aboveground biomass mapping in French Guiana by combining remote sensing, forest inventories and environmental data. *International Journal of Applied Earth Observation and Geoinformation*, v. 52, pp.502-514. 2016.

[35] Greaves, H.E., Vierling, L.A., Eitel, J.U., Boelman, N.T., Magney, T.S., Prager, C.M. and Griffin, K.L. High-resolution mapping of aboveground shrub biomass in Arctic tundra using airborne lidar and imagery. *Remote sensing of environment*, v. 184, pp.361-373. 2016.