Remote Sensing of Environment 232 (2019) 111323

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Combining LiDAR and hyperspectral data for aboveground biomass modeling in the Brazilian Amazon using different regression algorithms

Catherine Torres de Almeida[a,*], Lênio Soares Galvão[a], Luiz Eduardo de Oliveira Cruz e Aragão[a,b], Jean Pierre Henry Balbaud Ometto[a], Aline Daniele Jacon[a], Francisca Rocha de Souza Pereira[a], Luciane Yumie Sato[a], Aline Pontes Lopes[a], Paulo Maurício Lima de Alencastro Graça[c], Camila Valéria de Jesus Silva[d], Jefferson Ferreira-Ferreira[e], Marcos Longo[f]

[a] National Institute for Space Research - INPE, Caixa Postal 515, 12227-010 São José dos Campos, SP, Brazil
[b] College of Life and Environmental Sciences, University of Exeter, Exeter, United Kingdom
[c] National Institute for Research in Amazonia - INPA, Caixa Postal 2223, 69080-971 Manaus, AM, Brazil
[d] Lancaster Environment Centre, Lancaster University - Bailrigg, Lancaster LA1 4YW, United Kingdom
[e] Instituto de Desenvolvimento Sustentável Mamirauá, Caixa Postal 38, 69553-225 Tefé, AM, Brazil
[f] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr, Pasadena, CA 91109, United States

ARTICLE INFO

Keywords:
Hyperspectral remote sensing
Laser scanning
Data integration
Tropical forest
Carbon stock

ABSTRACT

Accurate estimates of aboveground biomass (AGB) in tropical forests are critical for supporting strategies of ecosystem functioning conservation and climate change mitigation. However, such estimates at regional and local scales are still highly uncertain. Airborne Light Detection And Ranging (LiDAR) and Hyperspectral Imaging (HSI) can characterize the structural and functional diversity of forests with high accuracy at a sub-meter resolution, and potentially improve the AGB estimations. In this study, we compared the ability of different data sources (airborne LiDAR and HSI, and their combination) and regression methods (linear model - LM, linear model with ridge regularization - LMR, Support Vector Regression - SVR, Random Forest - RF, Stochastic Gradient Boosting - SGB, and Cubist - CB) to improve AGB predictions in the Brazilian Amazon. We used georeferenced inventory data from 132 sample plots to obtain a reference field AGB and calculated 333 metrics (45 from LiDAR and 288 from HSI) that could be used as predictors for statistical AGB models. We submitted the metrics to a correlation filtering followed by a feature selection procedure (recursive feature elimination) to optimize the performance of the models and to reduce their complexity. Results showed that both LiDAR and HSI data used alone provided relatively high accurate models if adequate metrics and algorithms are chosen (RMSE = 67.6 $Mg.ha^{-1}$, RMSE% = 36%, R^2 = 0.58, for the best LiDAR model; RMSE = 68.1 $Mg.ha^{-1}$, RMSE % = 36%, R^2 = 0.58, for the best HSI model). However, HSI-only models required more metrics (5–12) than LiDAR-only models (2–5). Models combining metrics from both datasets resulted in more accurate AGB estimates, regardless of the regression method (RMSE = 57.7 $Mg.ha^{-1}$, RMSE% = 31%, R^2 = 0.70, for the best model). The most important LiDAR metrics for estimating AGB were related to the upper canopy cover and tree height percentiles, while the most important HSI metrics were associated with the near infrared and shortwave infrared spectral regions, particularly the leaf/canopy water and lignin-cellulose absorption bands. Finally, an analysis of variance (ANOVA) showed that the remote sensing data source (LiDAR, HSI, or their combination) had a greater effect size than the regression algorithms. Thus, no single algorithm outperformed the others, although the LM method was less suitable when applied to the HSI and hybrid datasets. Results show that the synergistic use of LiDAR and hyperspectral data has great potential for improving the accuracy of the biomass estimates in the Brazilian Amazon.

* Corresponding author.
E-mail addresses: catherine.almeida@inpe.br (C.T.d. Almeida), lenio.galvao@inpe.br (L.S. Galvão), luiz.aragao@inpe.br (L.E.d.O.C.e. Aragão), jean.ometto@inpe.br (J.P.H.B. Ometto), alinejacon@hotmail.com (A.D. Jacon), francisca.pereira@inpe.br (F.R.d.S. Pereira), luciane.sato@inpe.br (L.Y. Sato), aline.lopes@inpe.br (A.P. Lopes), pmlag@inpa.gov.br (P.M.L.d.A. Graça), c.silva@lancaster.ac.uk (C.V.d.J. Silva), jefferson.ferreira@mamiraua.org.br (J. Ferreira-Ferreira), mlongo@jpl.nasa.gov (M. Longo).

https://doi.org/10.1016/j.rse.2019.111323
Received 8 March 2019; Received in revised form 4 July 2019; Accepted 15 July 2019
Available online 07 August 2019

1. Introduction

Aboveground biomass (AGB) is a major component of the terrestrial carbon cycle and its accurate estimate is critical for supporting policies of ecosystem functioning conservation and climate change mitigation (Houghton et al., 2009). Amazonian forests host Earth's most extensive areas of high plant biomass (Pan et al., 2013). However, carbon stocks and balance across the Amazon are still highly uncertain (Le Quéré et al., 2018; Ometto et al., 2014).

Remote sensing has been recognized as an effective tool for quantifying carbon stocks over large areas, allowing accurate monitoring at the landscape scale (Lu et al., 2014). Several studies have estimated AGB from different sources of remotely sensed data, such as the hyperspectral imaging (HSI) (e.g., De Jong et al., 2003; Psomas et al., 2011) and Light Detection And Ranging (LiDAR) (e.g., Longo et al., 2016; Zolkos et al., 2013). Among the various types of sensors, LiDAR has been recognized as a consolidated technology to characterize complex forest structure due to its ability to capture three-dimensional information of the land surface (Koch, 2010). Moreover, LiDAR is less sensitive to signal saturation than passive optical sensors. Despite its advantages, LiDAR has restricted spectral resolution, generally covering a single spectral range in the near infrared region (Lu et al., 2014). Thus, variations in biomass due to species composition and stress may not be accurately detected by this sensor.

In contrast to LiDAR, HSI (also called imaging spectrometry/spectroscopy or hyperspectral remote sensing) sensors acquire data in a large number of narrow and contiguous spectral bands. HSI is capable of detecting absorption features useful for distinguishing functional and compositional traits (Ustin et al., 2004). For instance, hyperspectral sensors have been used to estimate land cover classes, plant functional types, tree species (Roth et al., 2015), biochemical content (Asner et al., 2015), health status (Pu et al., 2008), and biophysical properties such as Leaf Area Index (LAI) (Gong et al., 2003) and biomass (Psomas et al., 2011). On the other hand, when compared to LiDAR, the ability of the HSI instruments to detect vertical structure over dense vegetation is limited since the reflectance comes mostly from the upper canopy (Fassnacht et al., 2014).

Integrating the complementary information provided by LiDAR and HSI sensors can therefore potentially improve the accuracy of the AGB modeling (Koch, 2010). Several studies have investigated the potential of combining LiDAR and HSI data for classifying land cover or forest species (e.g., Dalponte et al., 2012; Ghosh et al., 2014; Wang and Glennie, 2015). However, few studies have evaluated this combination for estimating AGB (Anderson et al., 2008; Fassnacht et al., 2014; Latifi et al., 2012; Luo et al., 2017a, 2017b; Swatantran et al., 2011), particularly with focus on tropical regions (Clark et al., 2011; Vaglio Laurin et al., 2014). In Costa Rica, Clark et al. (2011) found that linear regression models combining a single LiDAR and hyperspectral metric were no better than the best model using two LiDAR metrics. However, they pointed out the need to analyze a wide range of LiDAR and HSI metrics, as well as other regression techniques to estimate AGB. In Sierra Leone, Vaglio Laurin et al. (2014) found improved AGB estimates using Partial Least Square Regression (PLSR) from combined LiDAR and hyperspectral data, when compared with LiDAR data alone. Thus, more research efforts are needed to explore different statistical procedures and metrics of HSI and LiDAR for AGB modeling, especially over tropical forests.

Many challenges arise from the integration of different data sources, such as the high data dimensionality, the redundancy of some metrics and the selection of the most suitable prediction model. Linear regression models (LM) have been commonly used for estimating AGB from remote sensing data, because of their simplicity and interpretability (Fassnacht et al., 2014). However, these statistical models are less flexible than non-parametric techniques, demanding large sample sizes and being affected by multicollinearity (Li et al., 2014). Nonparametric machine learning techniques, such as Support Vector Regression (SVR), Stochastic Gradient Boosting (SGB), Random Forest (RF) and Cubist (CB), are more versatile than LM in identifying complex nonlinear relationships and in dealing with high data dimensionality. Such techniques may provide more accurate AGB estimates than linear regression models, especially when multisource data are used (Lu et al., 2014).

Apart from identifying proper regression algorithms, an equally important challenge in multisource data integration is the selection of the most informative independent set of metrics for AGB estimation (Torabzadeh et al., 2014). In this context, feature selection methods, such as the recursive feature elimination (RFE), have the advantages of maximizing model performance (Guyon et al., 2002). RFE improves the generalization efficiency by avoiding overfitting while reducing the complexity of the model. The selection of a small subset of metrics generally facilitates the interpretation of the models and their inversion and applicability over large areas.

This study aims to explore optimal procedures for improving AGB modeling in the Brazilian Amazon through a comparative analysis of different data sources (airborne LiDAR and HSI, and their combination) and algorithms (linear models with (LMR) and without (LM) regularization, SVR, RF, SGB, and CB). For this purpose, we calculated a large variety of LiDAR and HSI metrics for maximizing the potential information related to vegetation biomass retrieved by each data source. By using a backward feature selection (RFE) method, we dealt with the high data dimensionality and evaluated the impact of reducing the number of input features for the models.

At the best of our knowledge, this is the first study that examines whether the use of HSI in conjunction with LiDAR data can improve AGB estimates using 12 sites regionally distributed over the Brazilian Amazon. Moreover, we addressed the synergy between airborne LiDAR and HSI data for AGB modeling from the perspectives of: (1) using both high spatial (1 m) and spectral resolution optical data; (2) detecting the metrics more related to AGB from a large set of attributes; (3) determining the optimal number of metrics required by each dataset; (4) testing the performance of different regression algorithms; and (5) examining the effect size of data source, regression algorithm and their interactions on models' performance.

2. Material and methods

2.1. Study sites and field data

This study was conducted in 12 sites in the Brazilian Amazon (Table 1; Fig. 1), representing different climate conditions (Köppen-Geiger classes Af, Am, and Aw) (Kottek et al., 2006), soil types (Ferralsols, Acrisols, and Gleysols) (Quesada et al., 2011), forest structure, species composition, and disturbance history. Forest inventory data comprised 132 sample plots collected between 2012 and 2017 (Table 1). Most plots (116) have approximately 0.25 ha and 16 plots have 0.16 ha. For the oldest plots, we assumed that potential changes in AGB due to temporal differences between forest inventories and remote sensing data acquisitions (2016–2017) had limited influence on the predictive modeling. For instance, changes in yearly AGB across the Amazon biome are, on average, 1.0 $Mg.ha^{-1}.yr^{-1}$ in old-growth forests (Baker et al., 2004); 1.3 $Mg.ha^{-1}.yr^{-1}$ in selectively logged forests (Rutishauser et al., 2015); and 6.1 $Mg.ha^{-1}.yr^{-1}$ in secondary forests (Poorter et al., 2016). This variation is within the uncertainty in field AGB estimates observed here, which will be further considered in our modeling framework.

Inventory data included species identification and measurements of DBH (Diameter at Breast Height) and total tree height. Due to differences in inventory protocols among the sites, especially with respect to the sampling of palms, lianas, and standing dead trees, we only considered the living trees in the AGB calculation. DBH measurements were obtained with metric tapes for living trees with a minimum of 10 cm DBH. For most sites, all trees that met this DBH threshold were measured along the entire plot area. A subsampling strategy for smaller

Table 1
Description of the study sites and sample plots.

Brazilian state	Site	Latitude	Longitude	Rainfall[a]	Forest type	Forest status	Field data				
		(°)	(°)	(mm.yr^{-1})			AGB mean ± sd (Mg.ha^{-1})	Plots (*n*)	Plot size[b] (m)	Year	Source
Amazonas (AM)	MAM	−2.76	−65.10	3014	SFO	IMF	232 ± 71	8	50 × 50	2016	IDSM
	ZF2	−2.60	−60.21	2603	TFO	IMF	318 ± 73	23	120 × 20	2015	LMF/INPA
	DUC	−2.95	−59.94	2433	TFO	IMF	277 ± 63	11	50 × 50(20)	2016	SL
	AUT	−3.51	−59.26	2155	TFO	IMF and DMF	166 ± 43	16	250 × 10	2017	FATE
Pará (PA)	TAP	−3.12	−54.95	2112	TFO	DMF and SS	142 ± 78	5	50 × 50	2016	SL
	SFX1	−6.43	−52.11	2164	TFO	DMF	107 ± 78	8	40 × 40	2012	SL
	SFX2	−6.56	−51.81	2194	TFO	DMF	160 ± 86	8	40 × 40	2012	SL
	PAR	−3.28	−47.52	1861	TFO	DMF and SS	101 ± 54	17	125 × 20(2)	2013	SL
Rondônia (RO)	JAM	−9.12	−63.01	2081	TFO	DMF	179 ± 72	11	50 × 50(5)	2013	SL
Mato Grosso (MT)	ALF	−9.58	−55.90	2233	TFO	DMF	174 ± 61	8	60 × 40	2017	FATE
	FN1	−12.00	−54.20	1873	TFT	DMF and SS	34 ± 38	6	50 × 50(5)	2015	SL
	FN2	−12.26	−55.10	1837	TFT	DMF	170 ± 46	11	50 × 50(5)	2015	SL
Total							189 ± 101	132			

Abbreviations: SFO, Seasonally Flooded Ombrophilous forest; TFO, *Terra Firme* (unflooded) Ombrophilous forest; TFT, *Terra Firme* (unflooded) Transitional forest (ecotone between ombrophilous and seasonal forests); IMF, Intact Mature Forest; DMF, Disturbed Mature Forest (submitted to fragmentation, fire, or logging); SS, Secondary Succession; IDSM, Instituto de Desenvolvimento Sustentável Mamirauá; LMF/INPA, Laboratório de Manejo Florestal do Instituto Nacional de Pesquisas da Amazônia; SL, Sustainable Landscapes project; FATE, Fire-Associated Transient Emissions in Amazonia.

[a] Average from the 1998–2015 TRMM 3B43 product (TRMM, 2011).

[b] The subplot size, when used, is given in parentheses.

trees (10–35 cm) was used at sites DUC, PAR, JAM, FN1, and FN2 (Table 1). For these sites, we accounted for the size-dependent sampling area when aggregating individual AGB to plot-level AGB.

The total height of trees was measured using clinometers, whenever possible. When the height was not measured for every tree, a stand-specific height-DBH (H-D) relationship based on a Weibull function (Feldpausch et al., 2012) was used. When no height data were available (ZF2, DUC, JAM, and ALF sites), the regional-specific H-D model proposed by Feldpausch et al. (2012) was used. For further uncertainty propagation of the AGB, each tree in the database was associated with a height error. When a measurement was present, we assumed an error of 12% of the total height, based on the median error found by Hunter et al. (2013). When the height was estimated by Weibull functions, we considered the residual standard error for the local or regional H–D model.

The identification of plant species was used to obtain the values of wood density (WD) with the getWoodDensity function from the R package BIOMASS (Réjou-Méchain et al., 2017). The global tree wood density database (Chave et al., 2009; Zanne et al., 2009) was used as a reference. Each tree received a wood density value based on its species- or genus-level average if at least one value in the same genus was available in the reference database. For unidentified trees, or if the genus was not determined in the reference database, the stand-level mean wood density was assigned to the tree, based on trees for which a value was attributed. The standard deviation of wood density for each tree was stored to account for uncertainty in this variable.

Based on the DBH (in cm), height (H, in m) and WD (in g.cm^{-3}), the AGB (in Mg) of individual trees was estimated using the pantropical allometric equation of Chave et al. (2014):

$$AGB_{tree}\,(Mg) = 6.73*10^{-5}*(DBH^2*H*WD)^{0.976} \quad (1)$$

To account for the uncertainty introduced by the measurements and the allometric equation, we propagated the errors using the AGBmonteCarlo function (BIOMASS package). This Monte Carlo approach simulated 1000 AGB_{tree} by adding random errors to the measurements and the allometric model parameters (Réjou-Méchain et al., 2017). The individual tree biomass was divided by its associated sampling area to convert to Mg.ha^{-1}. Then, the AGB of all trees of each plot was summed to calculate the plot-level AGB. Thus, each plot had 1000 AGB values and its respective average value (AGB_{mean}). The field plots covered a wide range of AGB_{mean}, varying from 2.7 Mg.ha^{-1} to 493.7 Mg.ha^{-1}, with a mean value of 188.5 Mg.ha^{-1} and a standard deviation of 101.05 Mg.ha^{-1}.

2.2. Remote sensing data acquisition and preprocessing

Both LiDAR and hyperspectral data were collected by manned aircraft in single transects of approximately 12.5 km × 300 m for each site. In the AUT and DUC sites, two nearly adjacent transects were needed to cover the area of the field plots. Thus, a total of 14 remote sensing transects was used. Airborne discrete-return LiDAR data were acquired between January 2016 and April 2017 using the Trimble HARRIER 68i system at an average height of 600 m above ground and a scan angle of 45°. The LiDAR sensor recorded multiple returns with a minimum point density of four points.m^{-2} and a small footprint of approximately 0.3 m. The horizontal accuracy varied among sites from 0.035 m to 0.185 m, while the vertical accuracy ranged from 0.07 m to 0.33 m. The raw point cloud of each site was preprocessed by first identifying and removing isolated noisy points with the *lasnoise* function, from the LASTools software (Isenburg, 2018). The parameters step_xy, step_z and isolated were set to 10, 5 and 5, respectively. Ground points were filtered (*GroundFilter* function with cellsize of 10, tolerance of 0.05 and 10 iterations) and then interpolated (*TINSurfaceCreate* function) into a digital terrain model (DTM) with a 1 m spatial resolution, using the FUSION/LDV software (McGaughey, 2014). To obtain the height above ground of each point, the DTM was subtracted from point elevations (function *Clipdata*, FUSION/LDV). The normalized point clouds were clipped according to the spatial extent of each field plot (function *PolyClipData*, FUSION/LDV) to further calculate the LiDAR metrics at the plot level.

Airborne hyperspectral data were collected between September and October 2017 using the AISAFenix sensor (Specim, *Spectral Imaging, Ltd.*) at an average height of 800 m above ground. To reduce variations in viewing-illumination geometry, we oriented the flight lines simultaneously close to the N-S direction. In addition, the HSI data were preferentially collected over sunny days between 10 a.m. and 1 p.m. (local time). The mean solar zenith angle (SZA) during data acquisition was 30° with a standard deviation of 7°. The at-sensor radiance was measured in 361 bands in the spectral range of 380–2500 nm, where 87 bands were located in the VNIR (visible and near infrared) region and 274 bands in the SWIR (shortwave infrared). Bandwidth ranged from 5.7 nm (SWIR) to 6.8 nm (VNIR). The spatial resolution was 1 m. Due to noise, we removed bands outside the range of 460–2330 nm and around the two major spectral intervals of atmospheric water vapor absorption

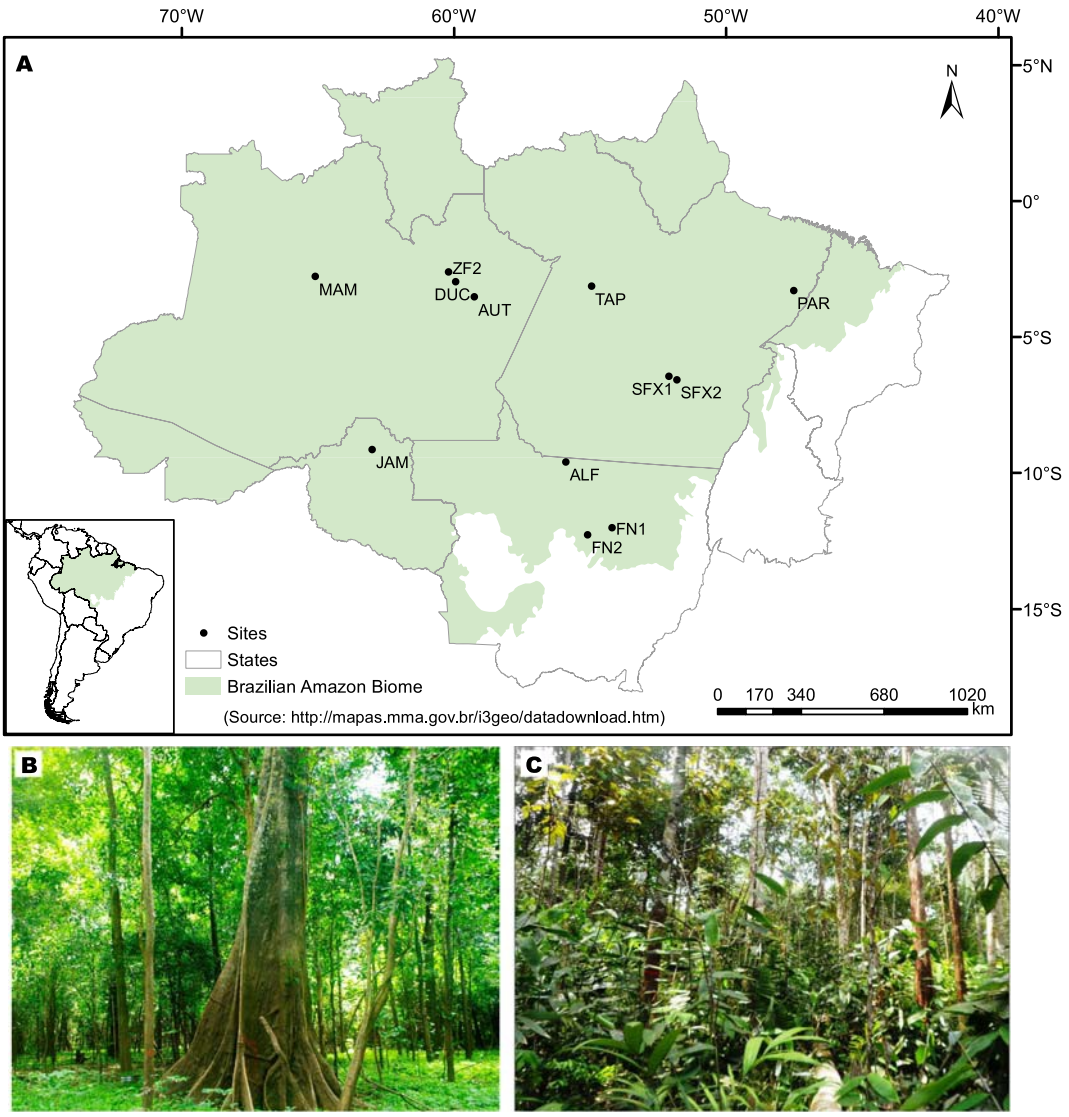

Fig. 1. (A) Distribution of the studied sites in the Brazilian Amazon Biome. Examples of sampled forests are shown in (B) for a seasonally flooded intact mature forest (MAM site) and in (C) for a *terra firme* forest degraded by understory fire (AUT site).

(1400 and 1900 nm), reducing the number of bands to 232. We used the Atmospheric/Topographic Correction for Airborne Imagery tool (ATCOR-4; version 6.3) to convert the radiance images into atmospherically-corrected surface reflectance data. Water vapor estimates were based on the 940-nm absorption feature. Data provided by a GPS onboard the aircraft were used for geometric correction of the scenes.

2.3. LiDAR metrics

Several LiDAR metrics have been proposed as potential predictors of canopy structural variables such as AGB (Lu et al., 2014; Zhang et al., 2017a). Here, we tested a variety of area-based LiDAR metrics (Table 2) related to height distribution (height statistics such as mean, standard deviation, and percentiles), canopy cover (proportion of returns and Leaf Area Density), structural complexity (Shannon and Simpson diversity indices), and topography (terrain roughness).

Height metrics were calculated from the first returns that were considered to belong to the tree canopy, i.e., points above a 2-m height (Næsset and Gobakken, 2008). We used only the first returns because they are more related to canopy surface structure (Thomas et al., 2006) and are more stable across different LiDAR acquisition settings, such as the point density (Singh et al., 2016) and flying altitude (Næsset, 2009).

Two types of canopy cover-related metrics were calculated. The first consists of point densities (PD) at different height intervals (e.g., the proportion of returns above 2 m or between 2 and 10 m) or for different return types (PD_{1st}, the proportion of first returns related to all returns). The second is based on the Leaf Area Density (LAD) profile, which corrects the LiDAR point density from occlusion effects (Bouvier et al., 2015). The LAD profile was calculated with the LAD function of the lidR package (Roussel and Auty, 2018), with a height bin of 2 m and an extinction coefficient k of 0.695. The constant k was based on the study by Stark et al. (2012) in central Amazon. Canopy cover-related metrics were also derived using just the first returns, except the PD_{1st} metric, which also considered the number of all canopy returns in its formulation.

Metrics related to canopy structural complexity are based on two indices commonly used to describe species diversity in biological systems: the Shannon (H′) and Simpson (D) indices (Magurran, 2004). These diversity indices combine richness (number of species) and evenness (species abundance distribution) into a single measure. When applied to LiDAR data, they operate as a measure of vertical structural diversity, increasing with the vertical extent of the canopy and with a more equal distribution of point density or leaf area density across the profile (Stark et al., 2012). While the Shannon index is more strongly

Table 2
Metrics calculated from LiDAR data.

Metrics	Description
Height	
H.max	Maximum height (m).
H.mean	Mean height (m) of first returns above 2 m.
H.pX	X^{th} (05, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, or 95th) percentile of height distribution of first returns above 2 m.
H.sd	Height standard deviation (m) of first returns above 2 m.
H.cv	Height coefficient of variation (%) of first returns above 2 m.
H.skew	Skewness of height distribution of first returns above 2 m.
H.kurt	Kurtosis of height distribution of first returns above 2 m.
Canopy cover	
PD_{a_b}	Number of first returns between a height interval a_b (2_10, 10_20, or 20_30) divided by the number of all first returns.
PD_h	Number of first returns above a height h (2, 6, 10, 14, 18, 22, 26, or 30) divided by the number of all first returns.
PD_{1st}	Number of first returns above 2 m divided by the number of all returns above 2 m.
LAD_{a_b}	Leaf Area Density ($m^2\,m^{-3}$) between the height interval a_b (2_10, 10_20, or 20_30).
LAD_h	Leaf Area Density ($m^2\,m^{-3}$) above the height h (2, 6, 10, 14, 18, 22, 26, or 30).
Structural complexity	
HSCI	Shannon Structural Complexity Index, calculated from the LAD profile.
DSCI	Simpson Structural Complexity Index, calculated from the LAD profile.
Topography	
Roughness	Mean terrain roughness from a 10-m DTM.

influenced by richness (in that case, canopy height), the Simpson index gives more weight to evenness (i.e. the homogeneity of canopy area profiles). Therefore, we also tested this approach for measuring structural complexity. The HSCI (Eq. 2) and DSCI (Eq. 3) indices used here are equivalent to the Shannon and Simpson indices, respectively. However, they were normalized by a fixed number of height bins to have a scale between 0 and 1:

$$HSCI = \frac{-\sum_{i=1}^{HB}[p_i * \ln(p_i)]}{\ln(HB)} \quad (2)$$

$$DSCI = \frac{1}{\sum_{i=1}^{HB}(p_i^2) * HB} \quad (3)$$

where $p_i = LAD_i / \Sigma LAD_i$, i.e. the proportion of LAD in height bin i; and *HB* is the maximum number of height bins. In this study, HB was equal to 30, because we used 2 m bins between 0 and 60 m (maximum canopy height across the field plots).

Finally, we calculated the terrain roughness for characterizing the local topographic variability. Roughness was defined as the difference between the highest and lowest altitude in a 3 × 3 moving window (Wilson et al., 2007). To avoid extreme localized roughness values, we averaged the 1-m DTM to obtain a 10-m DTM, which served as input data in the analysis.

We eliminated highly correlated metrics with the *findCorrelation* function from the R package *caret* (Kuhn, 2008). A high threshold (absolute Pearson's correlation > 0.98) was adopted to remove only the metrics with nearly perfect correlation since the RFE algorithm later selects the most important variables to estimate AGB. We also checked for linear dependencies, removing them with the function *findLinearCombos* from the same package. A total of 45 LiDAR metrics were calculated. After filtering by predictor's correlation and linear dependency, 34 metrics remained for modeling. The LiDAR metrics removed in this process consisted of six height percentiles (H.p25, H.p30, H.p50, H.p60, H.p70, and H.p75) and five canopy cover metrics (PD_{10}, PD_{26}, PD_{30}, LAD_{10}, and LAD_{30}).

2.4. HSI metrics

In addition to the 232 reflectance bands (R_λ, which λ is the

Table 3
Vegetation indices calculated from the AISAFenix reflectance data.

Abbr.	Vegetation index	Equation	Reference
ARI1	Anthocyanin Reflectance Index 1	$(1/R_{549}) - (1/R_{701})$	Gitelson et al. (2006)
ARI2	Anthocyanin Reflectance Index 2	$[(1/R_{549}) - (1/R_{701})] * R_{797}$	Gitelson et al. (2006)
CAI	Cellulose Absorption Index	$0.5\,(R_{2039} + R_{2199}) - R_{2100}$	Nagler et al. (2000)
CRI1	Carotenoid Reflectance Index 1	$(1/R_{515}) - (1/R_{549})$	Gitelson et al. (2006)
CRI2	Carotenoid Reflectance Index 2	$(1/R_{515}) - (1/R_{701})$	Gitelson et al. (2006)
D_{LAI}	Difference for Leaf Area Index	$R_{1724} - R_{969}$	le Maire et al. (2008)
DWSI1	Disease Water Stress Index 1	R_{797}/R_{1662}	Apan et al. (2004)
DWSI2	Disease Water Stress Index 2	R_{1662}/R_{549}	Apan et al. (2004)
DWSI3	Disease Water Stress Index 3	R_{1662}/R_{680}	Apan et al. (2004)
DWSI4	Disease Water Stress Index 4	R_{549}/R_{680}	Apan et al. (2004)
DWSI5	Disease Water Stress Index 5	$(R_{797} + R_{549})/(R_{1662} + R_{680})$	Apan et al. (2004)
EVI	Enhanced Vegetation Index	$2.5\,(R_{797} - R_{673})/(R_{797} + 6\,R_{673} - 7.5\,R_{474} + 1)$	Huete et al. (2002)
GNDVI	Green Normalized Difference Vegetation Index	$(R_{797} - R_{549})/(R_{797} + R_{549})$	Gitelson et al. (1996)
LWVI1	Leaf Water Vegetation Index 1	$(R_{1096} - R_{983})/(R_{1096} + R_{983})$	Galvão et al. (2005)
LWVI2	Leaf Water Vegetation Index 2	$(R_{1096} - R_{1204})/(R_{1096} + R_{1204})$	Galvão et al. (2005)
ND_{Bleaf}	Normalized Difference for Leaf Biomass	$(R_{2160} - R_{1540})/(R_{2160} + R_{1540})$	le Maire et al. (2008)
ND_{chl}	Normalized Difference for Leaf Chlorophyll	$(R_{927} - R_{708})/(R_{927} + R_{708})$	le Maire et al. (2008)
NDLI	Normalized Difference Lignin Index	$[\log(1/R_{1751}) - \log(1/R_{1679})]/[\log(1/R_{1751}) + \log(1/R_{1679})]$	Serrano et al. (2002)
NDNI	Normalized Difference Nitrogen Index	$[\log(1/R_{1512}) - \log(1/R_{1679})]/[\log(1/R_{1512}) + \log(1/R_{1679})]$	Serrano et al. (2002)
NDVI	Normalized Difference Vegetation Index	$(R_{797} - R_{680})/(R_{797} + R_{680})$	Rouse et al. (1973)
NDWI	Normalized Difference Water Index	$(R_{859} - R_{1237})/(R_{859} + R_{1237})$	Gao (1996)
PRI	Photochemical Reflectance Index	$(R_{529} - R_{570})/(R_{529} + R_{570})$	Gamon et al. (1992)
PSRI	Plant Senescence Reflectance Index	$(R_{680} - R_{502})/R_{749}$	Merzlyak et al. (1999)
PWI	Plant Water Index	R_{900}/R_{969}	Peñuelas et al. (1997)
REP	Red-Edge Position	$700 + 40\,[(R_{re} - R_{701})/(R_{742} - R_{701})]$ $R_{re} = (R_{673} + R_{783})/2$	Guyot and Baret (1988)
RVSI	Red-Edge Vegetation Stress Index	$[(R_{714} + R_{749})/2] - R_{735}$	Merton (1998)
SR	Simple Ratio	R_{797}/R_{680}	Jordan (1969)
VI_{green}	Vegetation Index green	$(R_{549} - R_{680})/(R_{549} + R_{680})$	Gitelson et al. (2002)
VOG1	Vogelmann Index 1	R_{742}/R_{721}	Vogelmann et al. (1993)
VOG2	Vogelmann Index 2	$(R_{735} - R_{749})/(R_{714} + R_{728})$	Vogelmann et al. (1993)

wavelength band center in nm), we calculated several metrics from the HSI data: 30 vegetation indices (Table 3), 20 continuum-removal absorption parameters, and 6 sub-pixel metrics based on the linear spectral mixture analysis (SMA). These metrics explored the potential information associated with vegetation properties at the main spectral regions: visible region (460–690 nm), mainly associated with pigments; red-edge interval (690–760 nm), sensitive to changes in chlorophyll; near infrared (NIR: 760–1300 nm), expressing scattering of radiation by canopy constituents and having absorption bands due to leaf water at selected wavelengths (980 and 1200 nm); and SWIR (1500–2330 nm), having absorption bands due to lignin-cellulose and nitrogen.

Five continuum-removal absorption bands were defined from fixed wavelength edges: 461–536 nm (495-nm band), 556–749 nm (670-nm band), 893–1074 nm (980-nm band), 1097–1265 nm (1200-nm band), and 2039–2199 nm (2100-nm band). The continuum-removed spectrum was calculated by dividing the reflectance values within the absorption band by the corresponding values of a continuum line established between the edges (Clark and Roush, 1984). To reduce noise in the original reflectance, the spectra were firstly smoothed using a Savitzky-Golay filter with a window size of five bands and a first polynomial order. The continuum-removed absorption bands were characterized by the depth (D_c) at the absorption center (c), the width at half depth (W_c), the band area (A_c, the sum of depths along the band), and the asymmetry (As_c, the ratio of the area left to area right of the band center) (Kokaly et al., 2009).

The fractional abundance of the green vegetation (GV), shade, and nonphotosynthetic vegetation/soil (NP) endmembers were calculated using the *unmix* function from the R package *hsdar* (Lehnert et al., 2018). To select endmembers for GV and NP, we applied sequentially the minimum noise fraction (MNF) and the pixel purity index (PPI) techniques using the Environment for Visualizing Images (ENVI; Harris Geospatial Solutions, Inc.). Candidate endmembers detected by the PPI were projected over an *n*-dimensional scatterplot for finding the purest pixels at each site. The final GV and NP endmembers were then obtained by averaging the purest pixels of all sites. For the shade endmember, we considered a photometric shade with a uniform reflectance of zero (Clark et al., 2011).

All HSI metrics were first 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 shade fraction below 30% (S_{0_30}), between 30 and 60% (S_{30_60}), and above 60% (S_{60}). Highly correlated HSI metrics were eliminated before the RFE, as previously described in Section 2.3. The number of HSI metrics was thus reduced from 288 to 60. The removed HSI metrics consisted of 216 reflectance bands, four vegetation indices (CRI2, DWSI1, DWSI4, and VOG1), five absorption features (depth at 670 nm and area of 495, 670, 980, and 2100 nm bands), and the mean shade fraction.

2.5. Regression models

Six regression algorithms were considered in this study (Table 4), encompassing three main approaches: (i) linear models (LM and LMR), (ii) kernel-based models (SVR), and (iii) tree-based models (RF, SGB, and CB). All algorithms were implemented in the R package *caret* (Kuhn, 2008), which requires other packages listed in Table 4. An overview of each model is introduced below.

LM and LMR models are parametric methods, which account for linear relationships between response and predictors. LM (multivariable regression with ordinary least squares), associated with some technique of feature selection (e.g., stepwise), is the most common method applied to AGB estimation (Lu et al., 2014). As a parametric technique, it requires assumptions such as linearity, residual normality, homoscedasticity, and independence (Osborne and Waters, 2002). Furthermore, conventional LM may generate spurious results due to multicollinearity. Regularization methods such as ridge regression are valuable for addressing this issue, reducing the impact of redundant variables by shrinking their coefficients (Duzan and Shariff, 2015).

Support vector machine is a non-parametric machine learning technique widely used for classification purposes (Mountrakis et al., 2011). This method is also effective for regression tasks and is commonly referred to as SVR (Basak et al., 2017). The main idea behind SVR is to transform a nonlinear regression into a linear regression by mapping the input data into a high-dimensional feature space, using a kernel function. In this study, we tested three kernels: linear, polynomial, and radial basis function (RBF). We further report the results of the RBF, which generally performed better than linear and polynomial kernels with a lower number of metrics (Fig. S1). Besides the kernel function, the required parameters are the cost, which controls the complexity of the boundary between support vectors, and the sigma, which is a smoothing parameter. The range of values for the sigma parameter was estimated with the *sigest* function from the R package *kernlab* (Karatzoglou et al., 2004). The SVR method has proven its robustness to dimensionality, outliers in the training data, and the generalization ability (Monnet et al., 2011).

RF is an ensemble learning method that combines predictions (by averaging) of multiple Classification and Regression Trees (CART) (Breiman, 2001). Each tree is independently created from a bootstrap sample of the original data (a bagging approach). Moreover, each node of the tree is split using a specified number of randomly selected features (*mtry*). In this study, the mtry was defined as one-third of the total number of features and the RF was computed with a number of trees (*ntree*) of 1000. RF has become popular in remote sensing applications due to its promising predictive capabilities for high-dimensional datasets. Furthermore, RF is insensitive to multicollinearity, data noise, outliers, and overfitting (Belgiu and Dragut, 2016).

SGB uses a boosting ensemble method for combining predictions of several regression trees. Simple trees are fitted sequentially using the loss function gradient from the prior tree to increase emphasis on observations modeled poorly. At each iteration, a random subsample of the training dataset (without replacement) is used as input (Friedman, 2002). Instead of developing single complex trees, relatively small trees are combined by averaging their weighted predictions. The SGB involves parameters for controlling the learning process: (i) the number of boosting iterations (*n.trees*); (ii) the number of nodes per tree (*interaction.depth*); (iii) the learning rate (*shrinkage*), which penalizes the importance of each consecutive iteration; and (iv) the minimum terminal node size (*n.minobsinnode*) (Elith et al., 2008). Several advantages of the SGB algorithm have been highlighted, including its low sensitivity to outliers, great ability to deal with unbalanced training datasets, and its robustness in dealing with interaction among predictors (Friedman, 2002).

CB is a rule-based tree model, which produces linear regression models instead of simple values in the terminal nodes of trees, based on the M5 model tree (RuleQuest, 2018). In contrast to RF and SGB, CB does not retrieve one final model but a set of rules associated with sets of multivariable models. CB can also use a boosting–like scheme called *committees*, in which subsequent trees are created using adjusted versions to the training set outcome. Predictions from all the committees are averaged to produce the final prediction (John et al., 2018). In addition, the predictions generated by the model rules can be adjusted using nearby points from the training set data (defined by the parameter *neighbors*). CB has been shown to be a viable method for AGB estimation across different sites and scales (Blackard et al., 2008; Li et al., 2014; John et al., 2018).

2.6. Modeling framework: feature selection and model validation

We tested three datasets for AGB modeling: (1) 34 LiDAR metrics; (2) 60 HSI metrics; and (3) their combination (94 predictors). All remote sensing metrics were normalized (centered by mean and scaled by the standard deviation). The six regression algorithms (LM, LMR, SVR,

Table 4
Description of the regression models used in this study, including the parameters considered and the criteria used to rank the feature importance for AGB estimation.

Type	Abbr.	Model	Parameters	Feature rank criteria	R package
Linear	LM	Linear Model	–	Absolute value of t-statistic	stats
	LMR	Linear Model with Ridge Regularization	*alpha* = 0 *lambda* = 0.01, 0.5, 1	Absolute value of coefficients	glmnet, Matrix
Kernel-based	SVR	Support Vector Regression with Radial Basis Function Kernel	*cost* = 0.5, 1, 2, 4 *sigma* = e^i (i = −5, …,1)	Squared weights*	kernlab
Tree-based	RF	Random Forest	*ntree* = 1000 *mtry* = k/3	Increase in mean squared error by permuting a variable	randomForest
	SGB	Stochastic Gradient Boosting	*n.trees* = 50, 100, 150, 200, 250, 300 *interaction.depth* = 2 *shrinkage* = 0.1 *n.minobsinnode* = 5	Sum of the empirical improvement in squared error over all trees	gbm, plyr
	CB	Cubist	*committees* = 10, 20, 30 *neighbors* = 9	Usage (Linear combination of the rule conditions and terminal model)	Cubist

k is the number of predictors. * Guyon et al., 2002

RF, SGB, and CB) were applied to the three datasets in a modeling framework composed by two main steps: (1) selection of the most relevant metrics; and (2) validation of the selected models considering the field AGB uncertainty.

For feature selection, we applied the RFE algorithm (*rfe* routine of the *caret* package), using the AGB_{mean} of each plot as the response variable. The RFE was used to assess the effect of the number of input features over the model performance. The performance was evaluated by the Root Mean Squared Error (referred to $RMSE_{rfe}$), quantified in a 5-fold cross-validation scheme, repeated 10 times. Model parameters were optimized by using an internal 4-fold cross-validation and selecting the parameters with the lowest RMSE. The RFE procedure started with all available predictors of each dataset. The predictors were ranked according to a criterion of importance, specific for each regression method (Table 4). Less important features were sequentially removed prior to modeling until the two most important variables remained. At the end of the process, the optimal feature subset size was selected, defined as the lowest number of predictors whose mean $RMSE_{rfe}$ was within the 95% confidence interval of the lowest $RMSE_{rfe}$. This approach selects the most parsimonious yet informative model.

The best set of metrics and parameters selected for each dataset and regression method (Tables S1, S2, and S3) was used to train 1000 models, each with a Monte Carlo field-AGB simulation as the response variable. This yields a probability distribution of model performance, which accounts for variations due to uncertainties in the field data. We applied a 5-fold cross-validation scheme with 10 repetitions to quantify the performance of each model, in terms of coefficient of determination and RMSE (hereafter termed as CV-R^2 and CV-RMSE, respectively). The CV-RMSE was expressed both in AGB units ($Mg.ha^{-1}$) and as a percentage relative to the mean AGB of all sample plots (CV-RMSE%).

A two-way analysis of variance (ANOVA) was applied to examine the influence of the data source, regression method, and their corresponding interactions on model performance (CV-R^2 and CV-RMSE). Subsequently, a Tukey's test was considered for pairwise comparison of mean CV-R^2 and CV-RMSE calculated from the 1000 model runs of each 18 combinations of data sources and regression methods. Since the statistical significance (p-value) is affected by large samples, we also calculated the effect size as a measure of practical significance. For the ANOVA, we calculated the eta-squared (η^2), the ratio of the sum of the squares of the factor by the total sum of squares. For multiple comparisons between models, we calculated the Cohen's d (Cohen, 1988), the absolute difference between groups, standardized by the residual standard error from the ANOVA. We considered that a difference in mean CV-R^2 or CV-RMSE between models is practically significant when $d \geq 1$, that is, two groups differ by 1 standard error or more.

Finally, for analyzing the spatial variability of biomass, the dataset and regression method that produced the highest CV-R^2 and the lowest CV-RMSE were used to predict AGB on a regular 50 × 50 m grid (corresponding to the field plots area). It resulted in 1000 AGB estimations per pixel from which we calculated the AGB mean and standard deviation.

3. Results

3.1. Selection of LiDAR and HSI metrics to estimate AGB

LM was the method whose accuracy was mostly affected by the number of input variables, showing an increase in $RMSE_{rfe}$ after reaching the best accuracy (Fig. 2). The more variables were used in the LM, the greater the increase in $RMSE_{rfe}$, particularly when the two data sources were combined. This pattern was expected given the limitations of this parametric method in relation to the high dimensionality. The use of regularization (LMR models) solved well this problem. The methods LMR, SVR, and CB, when used only with LiDAR data, were less affected by the number of input metrics. Thus, adding variables into these models did not greatly improve their performance (reduction of the $RMSE_{rfe}$ in up to 4.6%).

LiDAR-only models required fewer metrics (from 2 for LM, LMR, and CB to 5 for RF) than HSI-only models (from 5 for LM to 12 for CB) to achieve optimal performance. The number of metrics selected for the multisensor models varied between 6 for LM to 38 for SVR. The contribution of each data source to the combined models, both in number of selected metrics and in their importance for the model performance, depended on the regression method considered (Fig. 3). The linear models (LM and LMR) selected more HSI than LiDAR variables when using the combined dataset. SVR prioritized the selection of LiDAR metrics (24 against 14 HSI metrics), which had greater relative importance for the model performance. RF and SGB also selected more LiDAR than HSI variables (14 vs. 9 in RF and 17 vs. 12 in SGB), but the most influential variable was derived from the HSI data (W_{2100}). The CB method had a more even contribution from LiDAR and HSI variables (10 LiDAR metrics vs. 11 HSI metrics). However, the ranking of the variables showed that the CB had a greater influence of a LiDAR metric (LAD_{20_30}).

The most important LiDAR and HSI metrics for estimating AGB were generally consistent among the different models (Tables S1, S2, and S3). The most informative LiDAR metrics were related to canopy cover of the upper layers (LAD_{20_30}, LAD_{22}, PD_{22}, LAD_{26}, LAD_{18}, and PD_{18}), height percentiles (e.g. H.p95, H.p40, and H.p05), and mean height (H.mean), showing a positive association with AGB (Fig. 4A). Structural complexity metrics (DSCI or HSCI) were selected for the methods SVR and SGB based on combined data. Some LiDAR metrics related to topography (roughness), height distribution variability (H.sd, H.cv, H.skew, and H.kurt) and canopy cover (LAD_{2_10} and LAD_{10_20}) were not

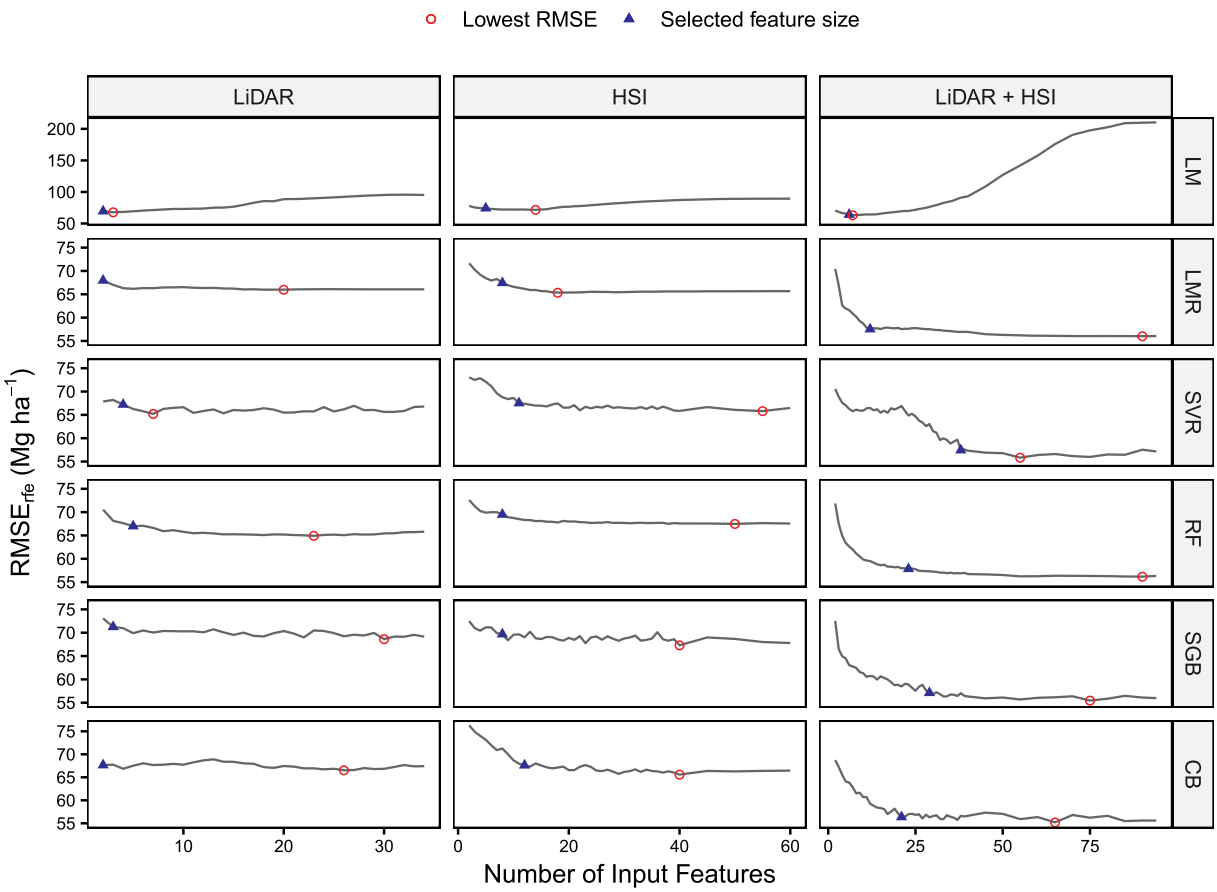

Fig. 2. Effect of the subset feature size on the cross-validated $RMSE_{rfe}$ for the regression methods and data sources used. The selected feature size was the smallest possible whose $RMSE_{rfe}$ was within the 95% confidence interval of the lowest $RMSE_{rfe}$. Note that the $RMSE_{rfe}$ scale for the LM method is different from the others.

selected by any model.

For the HSI data, the NIR and SWIR spectral regions were the most sensitive to AGB variations, including the absorption bands at 980 nm (leaf water) and 2100 nm (lignin-cellulose) (Fig. 4B and 5). Thus, metrics from these absorption bands (D_{980}, W_{980}, As_{980}, W_{2100}, and D_{2100}) were ranked as very informative for AGB estimation. Some vegetation indices and reflectance bands from the NIR (LWVI1, PWI, and R_{1091}) and SWIR (NDNI, CAI, ND_{Bleaf}, and R_{1646}) regions were also highly ranked. The proportion of shaded pixels (S_{0_30} and S_{30_60}) was also important for the AGB estimation, either directly, being selected by the methods SVR, SGB, and CB, or indirectly, being associated with the reflectance. Few metrics from the visible region (W_{495}, PRI, As_{670}, and R_{461}) were selected by the RFE. From that, the most informative was the width of the 495 nm chlorophyll absorption band (W_{495}), selected by five models (all except LM) with only HSI data and three models (RF, SGB, and CB) with the combined data. Vegetation indices resulting from a combination of visible and NIR reflectance were selected a few times (SR, PSRI, and VI_{green}) or not selected (e.g., NDVI, EVI, and GNDVI).

3.2. Performance of the data sources and regression methods for AGB modeling

The ANOVA results (Table 5) showed that the data source had the greatest effect on models' performance, explaining 65% of the variation in CV-R^2 and 55% of the variation in CV-RMSE. The regression method and its interaction with data source had a smaller contribution to the CV-R^2 (η^2 of 0.14 and 0.09, respectively) and CV-RMSE (η^2 of 0.10 and 0.07, respectively) variation. Therefore, there was no single best regression method. However, the LM method was less suitable for HSI and hybrid data, while the LMR presented high performance for all analyzed data sources (Fig. 6).

The combination of LiDAR and HSI data improved the performance of the models for all regression methods by reducing the CV-RMSE and increasing the CV-R^2 (Fig. 6 and Table 6). The improvements in CV-RMSE (reduction of 4.05–13.83 Mg.ha^{-1}) and CV-R^2 (increase of 0.05–0.18) achieved by the multisource models relative to models with single data were both statistically and practically significant (Cohen's $d \geq 1$) for all regression algorithms (Fig. 7). Relative to the best single-model of each method, the improvements in the combined models reached up to 15% reduction in CV-RMSE and 21% increase in CV-R^2.

Overall, models based on single-LiDAR data performed similarly to models based on single-HSI data, with no practical difference (Fig. 7). Only the LM method presented significantly superior performance with LiDAR when compared to the HSI data. SGB performed slightly better with LiDAR data than HSI data with a significant practical difference for CV-R^2 (increase in 0.03), but with no practical difference for CV-RMSE. All models underestimated the AGB for values > 300 Mg.ha^{-1}. However, models with multisensor data showed slightly lower underestimation (Fig. 8). We also found an overestimation for low AGB values (< 50 Mg.ha^{-1}), mainly with the HSI data.

3.3. Spatial variability of the predicted AGB

The spatial distribution of the HSI and LiDAR data and the AGB map (mean and standard deviation) derived from their combination are exemplified for the SFX1 (Fig. 9) and DUC (Fig. 10) sites. The AGB predictions covered both the variability within and between sites. In the SFX1 site, the predicted AGB ranged from zero Mg.ha^{-1}, due to intensively degraded areas, to 223 Mg.ha^{-1}, due to the presence of tall trees. The DUC site is an old-growth forest accounting for greater AGB.

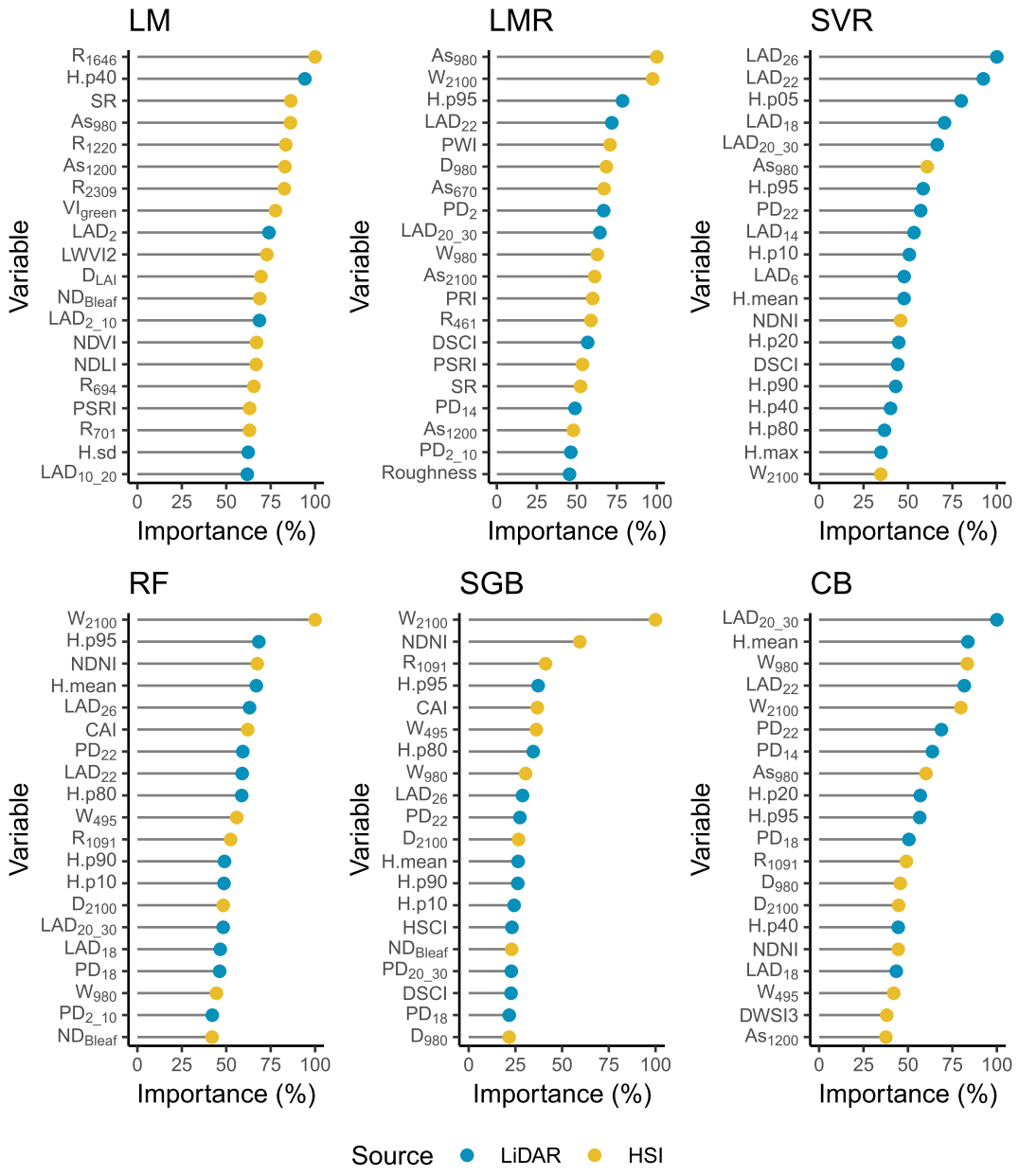

Fig. 3. Relative importance of the 20 highest ranked variables for each regression method with the combined dataset. The abbreviations of LiDAR and HSI metrics are given in Table 2 and Section 2.4, respectively.

In this site, the predicted AGB varied from 193 $Mg.ha^{-1}$ to 454 $Mg.ha^{-1}$, due to variations in canopy density and height. The AGB uncertainty (standard deviation) was greater (~ 15 $Mg.ha^{-1}$) at the tails of the predicted interval, i.e., at the locations with very low or very high predicted AGB.

4. Discussion

4.1. Single-LiDAR versus single-HSI AGB predictions

Our study confirmed the reliability of LiDAR-based AGB predictions in tropical ecosystems, consistent with previous tropical studies using small-footprint airborne LiDAR (d'Oliveira et al., 2012; Hansen et al., 2015; Kronseder et al., 2012; Longo et al., 2016; Mauya et al., 2015; Vaglio Laurin et al., 2014; Zolkos et al., 2013). The LiDAR metrics selected here as important for estimating AGB were also comparable with metrics identified in other studies, such as the mean height (Latifi et al., 2012; Longo et al., 2016), height percentiles (Vaglio Laurin et al., 2014; Li et al., 2014; Longo et al., 2016), and canopy-cover attributes (Latifi et al., 2012).

Previous studies that compared LiDAR with hyperspectral sensors have shown that LiDAR was more powerful for biomass prediction (Clark et al., 2011; Fassnacht et al., 2014; Vaglio Laurin et al., 2014). Koch (2010) states that a direct AGB estimation based only on HSI data is not likely, especially in high biomass stands. However, our results suggest that single-HSI models can provide good AGB predictions even in dense tropical forests with an accuracy equivalent to LiDAR models. The wide range of HSI metrics calculated in this study, exploring the information of different vegetation properties (e.g., canopy structure, biochemistry, leaf/canopy water content, and plant physiology or stress), contributed to the good performance of the HSI models. On the other hand, while few LiDAR variables generally contained most of the information needed to estimate AGB, a larger set of HSI metrics was necessary to achieve a similar performance of the models.

The absorption bands from the SWIR and NIR regions, as well some vegetation indices from the same spectral regions, were the most

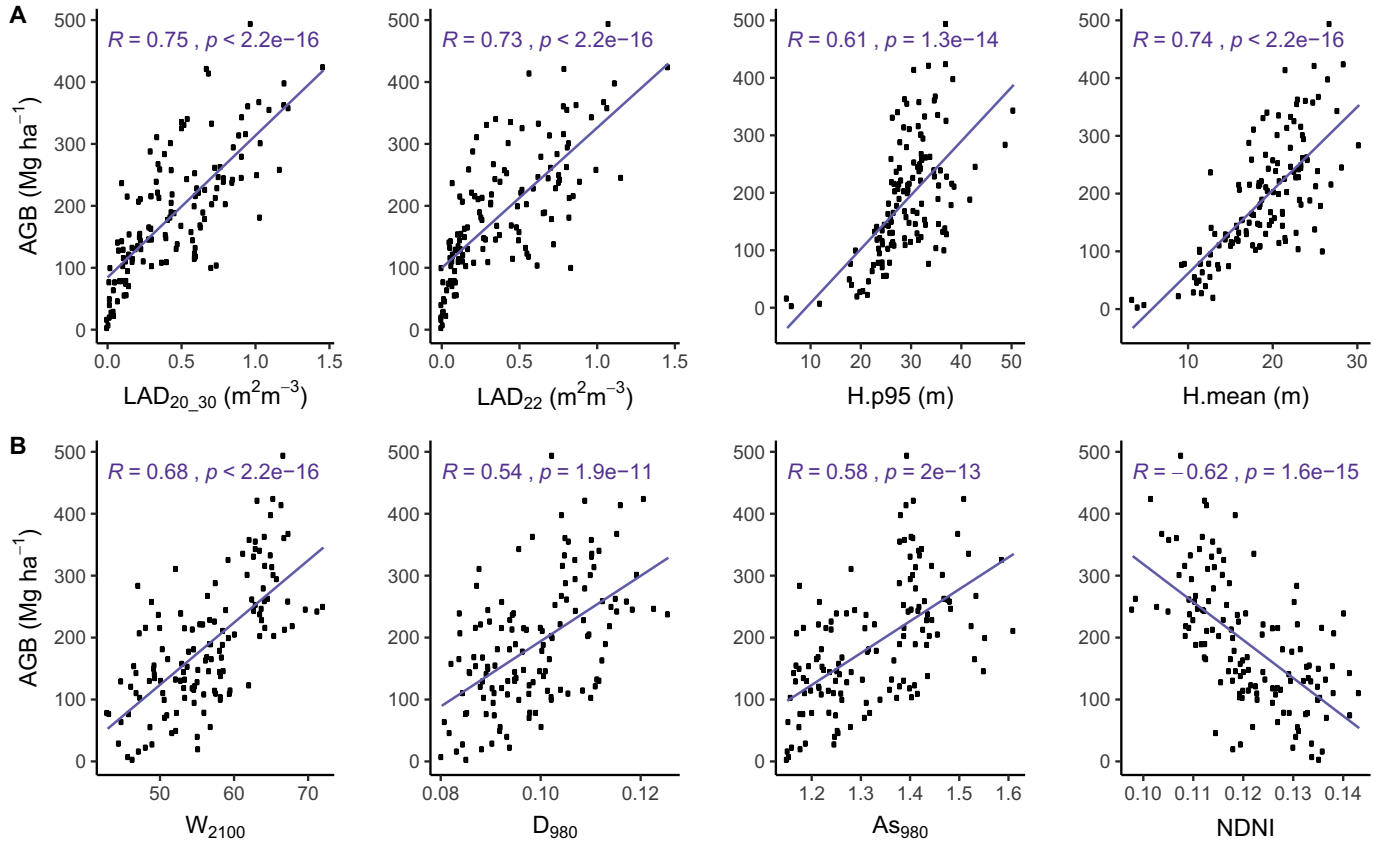

Fig. 4. Scatterplots of the four most important LiDAR (A) and HSI (B) metrics for aboveground biomass (AGB) estimation. The abbreviations of LiDAR and HSI metrics are given in Table 2 and Section 2.4, respectively. The blue line represents a linear fit. The correlation coefficient (R) with *p*-value is showed upward in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

influential metrics for estimating AGB with HSI data. For instance, the leaf/canopy water absorption bands, centered at 980 nm and 1200 nm, were indicated as important in the analysis. The same was verified for the 2100 nm SWIR absorption band, related to nitrogen, lignin, and cellulose (Kokaly et al., 2009). Previous studies suggested that such biochemical traits co-varied with canopy structure (Serrano et al., 2002; Kokaly et al., 2009). Therefore, optical metrics from the SWIR and NIR spectral regions have been recommended to estimate canopy structural attributes such as LAI (Gong et al., 2003; le Maire et al., 2008) and AGB (Psomas et al., 2011; Swatantran et al., 2011). They have been used also to estimate aboveground forest productivity (Smith et al., 2002)

In addition to the high spectral resolution, the high spatial resolution of the hyperspectral images used in this study contributed positively to the AGB models. The spatial resolution of 1 m provides information on the distribution of crowns and gaps. This resolution can be more directly related to the forest inventory information used to establish the models. Sub-pixel-based metrics, such as the proportion of shaded pixels, served as a measure of the canopy spatial arrangement, improving the AGB models. The proportion of shade increased with increasing amounts of AGB, reducing the overall reflectance. These results are consistent with those found by Barbier and Couteron (2015), who observed a negative linear relationship between the mean reflectance and the maximum DBH, a measure of forest structure, due to the shade proportion. Moreover, studies based on texture metrics from high spatial resolution optical data have shown good potential to provide non-saturating proxies for stand parameters, including AGB (Barbier and Couteron, 2015; Ploton et al., 2017). As a result, Barbier and Couteron (2015) state that LiDAR is not the only option for monitoring canopy structure and carbon stocks in tropical forests.

Our findings showed that some HSI indices commonly used for biomass estimation (e.g., SR and NDVI) saturated for AGB above 100 $Mg.ha^{-1}$. Thus, these metrics may be more useful for estimating AGB in simpler stand structures than in dense forests. In contrast, the most relevant HSI metrics for AGB estimation found here were almost unaffected by saturation at high AGB values.

4.2. Single-models versus combined-models

The improvements in AGB models based on the integration of LiDAR and hyperspectral data were consistent with the studies performed in temperate mixed forests from the USA (Anderson et al., 2008), tropical forests from Africa (Vaglio Laurin et al., 2014) and wetland vegetation from China (Luo et al., 2017b). The gain in explained variance (R^2) by the use of the hybrid approach reported in these studies, when compared with the best single-model, was within the range found in our investigation (absolute increase of 6–9%). In contrast, some studies performed in tropical (Clark et al., 2011) and temperate forests (Fassnacht et al., 2014; Latifi et al., 2012; Luo et al., 2017a) have shown only slight (around 2% absolute increase of R^2) or no improvements in AGB estimation after combining LiDAR and HSI for AGB modeling.

The differences in results from the literature can be explained by several factors that influence the performance of the AGB models. Examples of these factors include the regression technique chosen for analysis; the number and type of metrics selected as potential input data; the type of vegetation under study; and the quality of the field and remote sensing data used to obtain the models. For instance, some studies suggest that LiDAR data provide a more straightforward connection with vegetation structure, being able to produce satisfactory predictions with relatively simple techniques, such as linear regression approaches (Li et al., 2014; Longo et al., 2016). On the other hand,

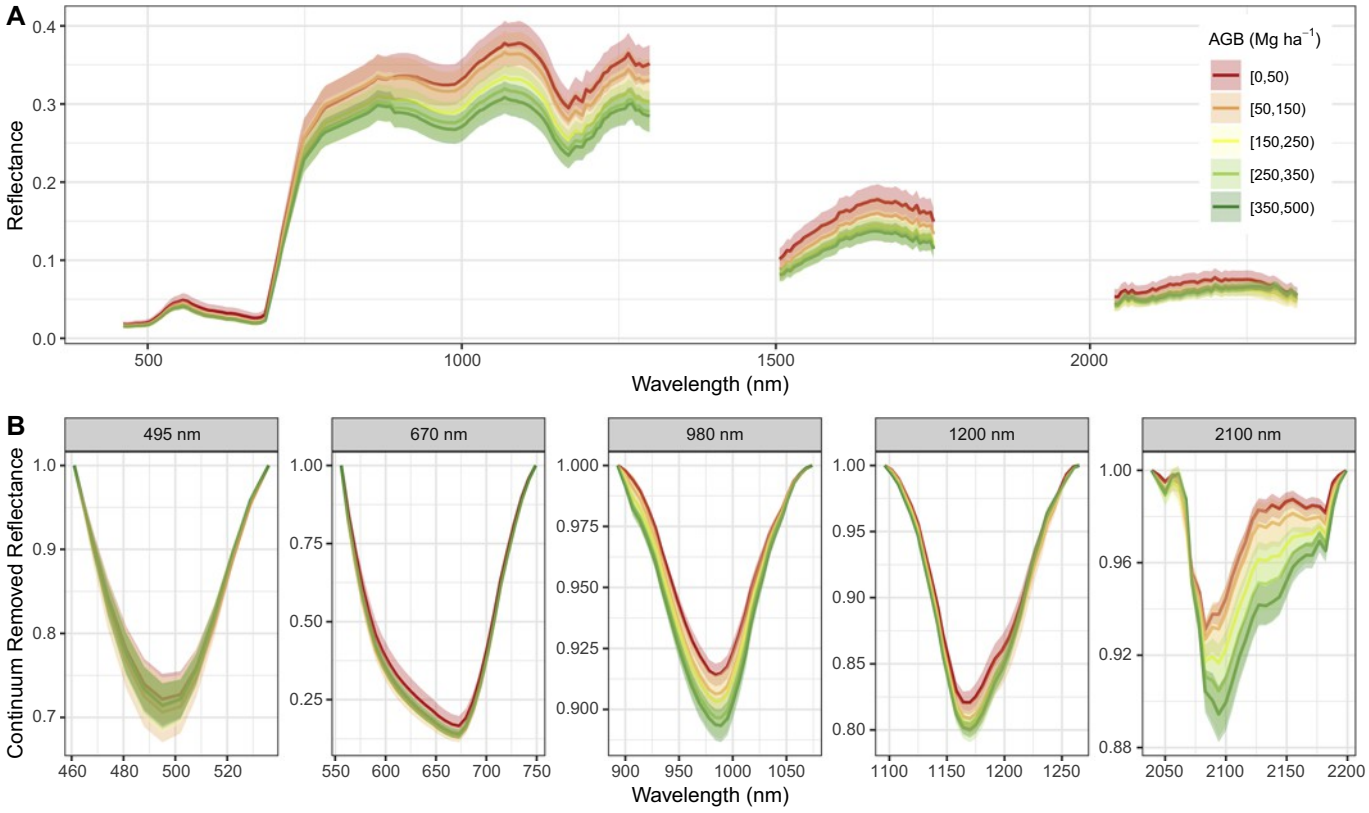

Fig. 5. Reflectance spectra (A) and continuum-removed reflectance spectra (B) across five aboveground biomass (AGB) ranges, indicated by the different colors. Spectral values are shown as mean ± standard deviation.

Table 5
Analysis of variance of the cross-validated R^2 and RMSE respective the data source, regression method, and their interaction.

Factor	Degree of freedom	Sum of squares	Mean square	F value	p-value	η^2
Response variable: *CV-R^2*						
Data	2	50.0	25.0	51,079.6	< 2e-16	0.65
Method	5	11.1	2.2	4515.9	< 2e-16	0.14
Data:method	10	7.3	0.7	1495.5	< 2e-16	0.09
Residuals	17,982	8.8	0.0			
Response variable: *CV-RMSE* (Mg.ha^{-1})						
Data	2	336,205.1	168,102.6	18,120.1	< 2e-16	0.55
Method	5	62,476.6	12,495.3	1346.9	< 2e-16	0.10
Data:method	10	45,849.6	4585.0	494.2	< 2e-16	0.07
Residuals	17,982	166,821.0	9.3			

hyperspectral measurements relate indirectly with biophysical properties, and thus, may need more complex models (Torabzadeh et al., 2014). In this study, the conventional linear regression model was not very suited for high dimensional datasets (based on HSI and multisensor source). However, the linear model with regularization showed superior performance by solving the issue of multicollinearity between the metrics. Nevertheless, this good performance was only possible because several metrics used in this study were non-saturating with large amounts of AGB, showing a consistent linear relationship with it. Studies based on metrics that saturate over dense vegetation may not find the same results.

Few studies have applied machine learning techniques for estimating AGB from multisource remote sensing data. Fassnacht et al. (2014) verified that the RF method outperformed other approaches (stepwise linear regression, SVR, Gaussian processes, and k-nearest neighbor) when using combined LiDAR and HSI data. Feng et al. (2017) compared different data sources and modeling approaches (linear, nonlinear, RF, and SVR) under stratification and non-stratification conditions of vegetation types. For the combination of LiDAR and RapidEye data, RF had the best performance under stratification. RF emerged also as the best algorithm for different data sources in the study by Cao et al. (2018) when compared to SVR, neural networks, k-nearest neighbor, and generalized linear mixed model. In our study, the regression method had little effect on the models' performance. With the exception of the LM with HSI and hybrid data, all the evaluated algorithms were useful for estimating AGB from remote sensing data.

Selecting the most appropriate metrics to estimate AGB is another factor that can affect model performance. We showed that it was possible to reduce considerably the number of metrics used as input data without losing much accuracy in AGB estimates. Even regression methods not entirely affected by the high data dimensionality can benefit from the reduction in the number of features. More elaborated feature selection procedures can produce parsimonious models for practical applications. Therefore, models based on multisource datasets require strategies to overcome the trade-off between the high data dimensionality and the loss of information for achieving a proper number of features.

The characteristics of the vegetation are also relevant to AGB prediction. In our study, we considered a wide variety of vegetation types, from intact old-growth forests to secondary forests, also including areas under different levels of degradation by fire, logging, or fragmentation. The information gain in AGB modeling provided by the HSI may be related to the discrimination of different vegetation types or conditions, such as canopy stress (Swatantran et al., 2011). Other studies based on multispectral data (Vieira et al., 2003; Zhang et al., 2017b) have demonstrated the potential of metrics related to the NIR and SWIR spectral regions and shade fractions for differentiating vegetation at different regrowth stages. Nevertheless, it is important to note that variations in the remote sensing data acquisitions, especially the

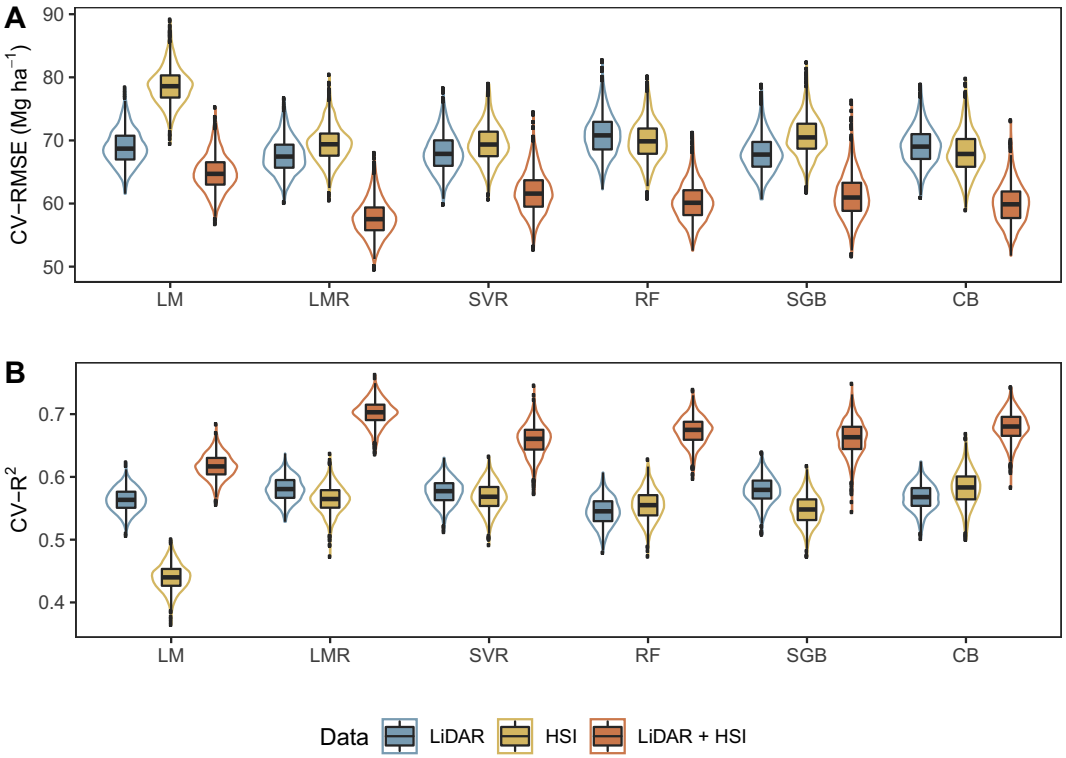

Fig. 6. Distribution of the 1000 cross-validated RMSE (A) and R^2 (B) for each regression method (abbreviations in Table 4) and data source (LiDAR, HSI, and their combination).

Table 6
Average cross-validated performance (for the 1000 model runs) for each regression method and data source.

Model	Data	#Features	Mean CV-RMSE		Mean CV-R^2
			Mg.ha^{-1}	%	
LM	LiDAR	2	68.90	36.54	0.56
	HSI	5	78.69	41.73	0.44
	Combined	6	64.85	34.39	0.62
LMR	LiDAR	2	67.60	35.85	0.58
	HSI	8	69.50	36.86	0.56
	Combined	12	57.69	30.59	0.70
SVR	LiDAR	4	68.10	36.11	0.58
	HSI	11	69.54	36.88	0.57
	Combined	38	61.78	32.77	0.66
RF	LiDAR	5	70.91	37.61	0.54
	HSI	8	69.96	37.10	0.55
	Combined	23	60.26	31.96	0.67
SGB	LiDAR	3	67.90	36.01	0.58
	HSI	8	70.68	37.48	0.55
	Combined	29	61.26	32.49	0.66
CB	LiDAR	2	69.12	36.66	0.57
	HSI	12	68.11	36.12	0.58
	Combined	21	59.98	31.81	0.68

varying view-illumination geometry, may affect some metrics (Galvão et al., 2013). In our study, the remote sensing data acquisition was designed to reduce such effects by orienting simultaneously most of the flight lines in the same direction (N-S). Although some variations in the average SZA remained (SZA = 30° ± 7° across sites), since the data were collected at different locations, we observed that such variations did not produce a systematic effect on the residuals of our best model (Fig. S2).

Modeling AGB in the extensive and highly diverse Amazon region has some obstacles such as the acquisition of high quality and standardized field data. For instance, the variable size of the field plots may be a source of uncertainty in the data analysis. Small plots are more susceptible to spatial heterogeneity, GPS location errors, and boundary effects (i.e., confusion in the inclusion/exclusion of trees at the edges of the plot). Moreover, the shape of the plots may also favor the edge effect, in cases of large perimeter-area ratio (Mauya et al., 2015). In our study, most plots were larger than 0.24 ha, the minimum area required to achieve model errors lower than 20% of field biomass (Zolkos et al., 2013). The few plots smaller than this size or the plots with greater perimeter-area ratio did not influence the residuals of our best model (Fig. S3). Another issue is the scarcity of field data in some under-represented regions and the considerable uncertainties related to field measurements and allometric equations. Terrestrial LiDAR offers a possible alternative to address this issue by improving field estimates of AGB, and therefore, the calibration and validation of models based on remote sensing data (Stovall and Shugart, 2018).

5. Conclusions

In this study, we explored the potential of combining LiDAR and HSI data for estimating AGB in the Brazilian Amazon, using six regression methods and a great number and type of metrics. We concluded that:

(1) Both LiDAR and HSI data used alone can effectively estimate AGB in tropical forests of the Amazon if proper metrics and regression methods are considered. However, HSI models required more input variables (5–12) than LiDAR models (2–5) for estimating AGB.
(2) The accuracy of the AGB estimates was improved in up to 15% in RMSE and 21% in R^2 after using the hybrid dataset relative to the single model of best performance.
(3) The most informative LiDAR metrics for estimating AGB were related to the upper canopy cover and tree height percentiles.
(4) The most important HSI metrics were associated with the NIR and SWIR spectral regions, mainly the water and lignin-cellulose absorption bands.
(5) From ANOVA, results showed that the source of remote sensing

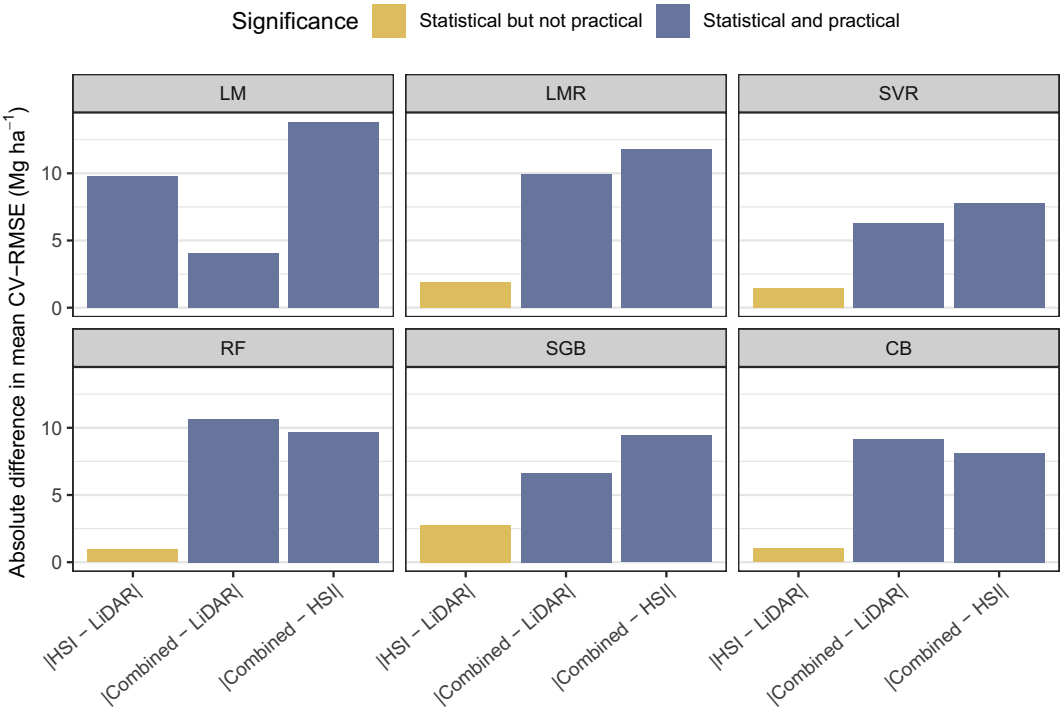

Fig. 7. Difference in mean cross-validated RMSE between models based on different data sources for each regression method.

data (HSI, LiDAR, or their combination) had a more important effect than the regression algorithms to estimate AGB. Thus, there was no single best regression method.

This study contributes to the investigation of the potential of LiDAR and hyperspectral remote sensing to estimate the AGB of tropical forests. More accurate estimates of forest carbon are highly required considering the current scenario of global environmental changes.

Acknowledgments

This work was supported by Project Environmental Satellite

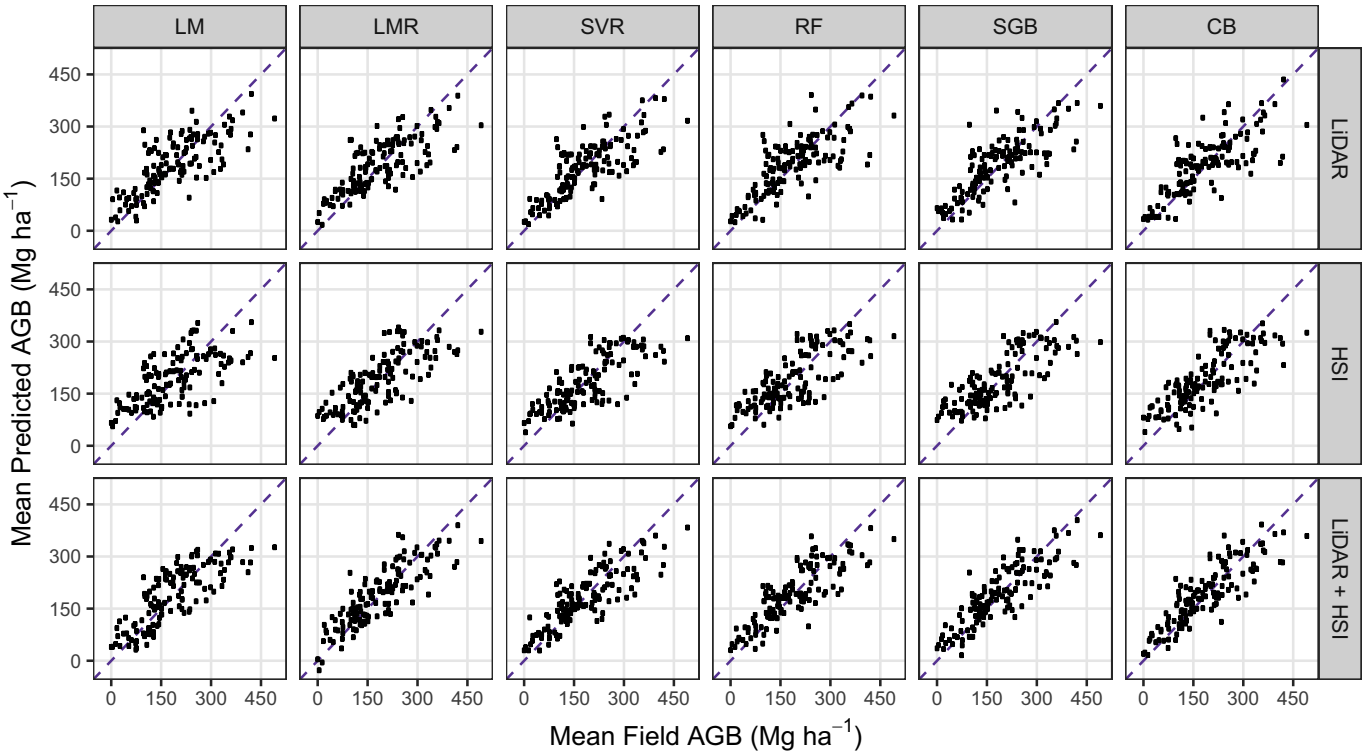

Fig. 8. Field AGB_{mean} versus predicted AGB (mean of cross-validated predictions from the 1000 model runs) from the different methods and data sources. The blue dashed 1:1 line is provided for reference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

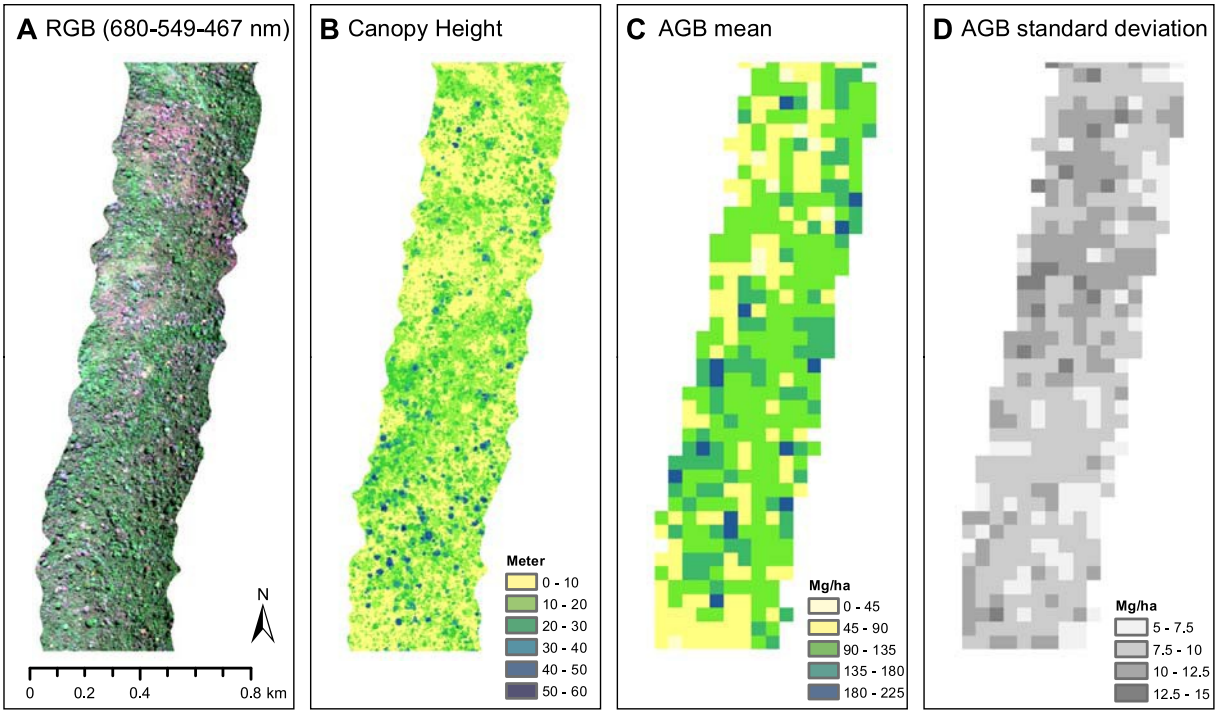

Fig. 9. (A) Spatial variability of HSI data (AISAFenix true color composite). (B) LiDAR data (Canopy Height Model). (C) Mean and (D) standard deviation of AGB predictions from the LMR method with multisensor data. Figs. A and B are in 1 m resolution, while Figs. C and D are in 50 m resolution. Results refer to the SFX1 site.

Monitoring in the Amazon Biome (MSA-BNDES) - Activity 7 "Improvement of biomass estimation methods and emission estimation models for change of land use", funded by Amazon Fund (process #14209291). This study was partially funded by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) [grant numbers 140502/2016-5 and 305054/2016-3] and FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) [grant number 2016/21043-8]. The research carried out at the Jet Propulsion Laboratory, California Institute of Technology, was under a contract with the National Aeronautics and Space Administration. ML was supported by the NASA Postdoctoral Program, administered by Universities Space Research Association under contract with NASA. We thank the FATE (Fire-Associated Transient Emissions in Amazonia) program, the IDSM (Instituto de Desenvolvimento Sustentável Mamirauá), and the LMF/

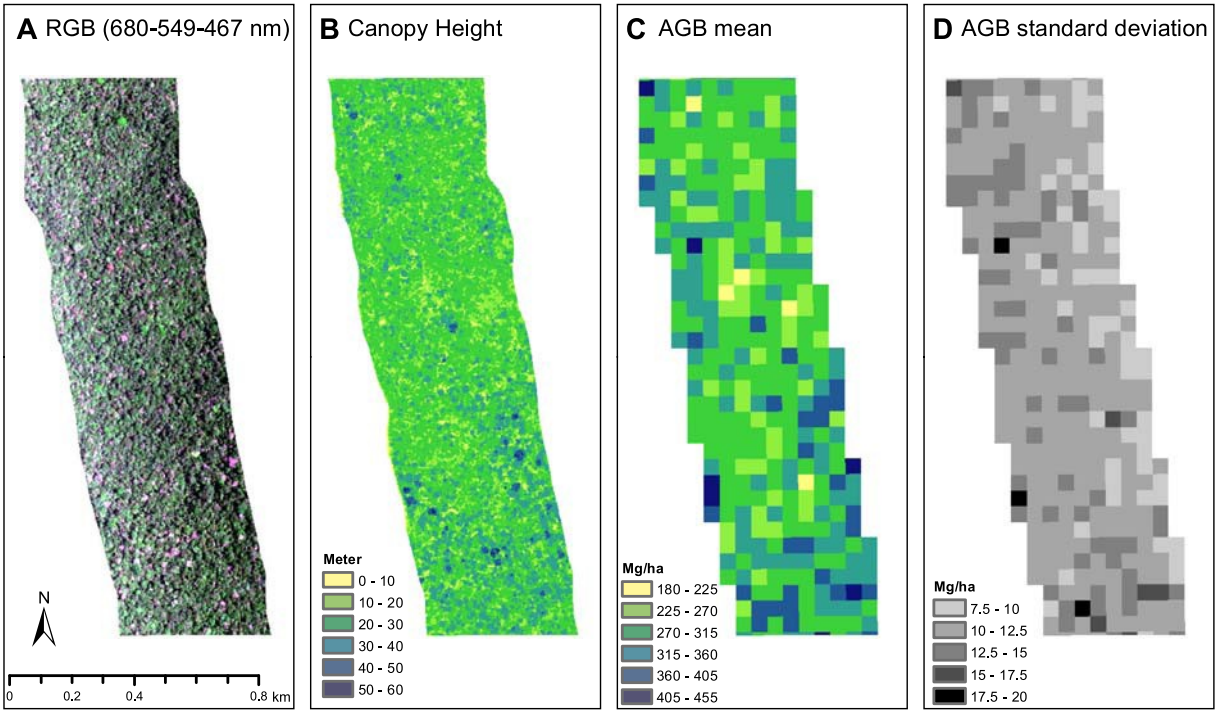

Fig. 10. (A) Spatial variability of HSI data (AISAFenix true color composite). (B) LiDAR data (Canopy Height Model). (C) Mean and (D) standard deviation of AGB predictions from the LMR method with multisensor data. Figs. A and B are in 1 m resolution, while Figs. C and D are in 50 m resolution. Results refer to the DUC site.

INPA (Laboratório de Manejo Florestal do Instituto Nacional de Pesquisas da Amazônia) for providing part of the field data used in the analysis. Data from 77 field plots were acquired by the Sustainable Landscapes Brazil project, supported by the EMBRAPA (Brazilian Agricultural Research Corporation), the US Forest Service, and USAID, and the US Department of State. We also thank the editors and three reviewers for providing constructive suggestions that helped to improve the quality of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2019.111323.

References

Anderson, J.E., Plourde, L.C., Martin, M.E., Braswell, B.H., Smith, M.L., Dubayah, R.O., Hofton, M.A., Blair, J.B., 2008. Integrating waveform lidar with hyperspectral imagery for inventory of a northern temperate forest. Remote Sens. Environ. 112, 1856–1870. https://doi.org/10.1016/j.rse.2007.09.009.

Apan, A., Held, A., Phinn, S., Markley, J., 2004. Detecting sugarcane "orange rust" disease using EO-1 Hyperion hyperspectral imagery. Int. J. Remote Sens. 25, 489–498. https://doi.org/10.1080/01431160310001618031.

Asner, G.P., Martin, R.E., Anderson, C.B., Knapp, D.E., 2015. Quantifying forest canopy traits: imaging spectroscopy versus field survey. Remote Sens. Environ. 158, 15–27. https://doi.org/10.1016/j.rse.2014.11.011.

Baker, T.R., Phillips, O.L., Malhi, Y., Almeida, S., Arroyo, L., Di Fiore, A., Erwin, T., et al., 2004. Increasing biomass in Amazonian forest plots. Philos. Trans. R. Soc. B Biol. Sci. 359, 353–365. https://doi.org/10.1098/rstb.2003.1422.

Barbier, N., Couteron, P., 2015. Attenuating the bidirectional texture variation of satellite images of tropical forest canopies. Remote Sens. Environ. 171, 245–260. https://doi.org/10.1016/j.rse.2015.10.007.

Basak, D., Pal, S., Patranabis, D.C., 2017. Support vector regression. Neural Inf. Process.–Lett. Rev. 11, 203–224. https://doi.org/10.1007/978-3-319-70087-8_72.

Belgiu, M., Dragut, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Blackard, J.A., Finco, M.V., Helmer, E.H., Holden, G.R., Hoppus, M.L., Jacobs, D.M., Lister, A.J., et al., 2008. Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. Remote Sens. Environ. 112, 1658–1677. https://doi.org/10.1016/j.rse.2007.08.021.

Bouvier, M., Durrieu, S., Fournier, R.A., Renaud, J.P., 2015. Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. Remote Sens. Environ. 156, 322–334. https://doi.org/10.1016/j.rse.2014.10.004.

Breiman, L.E.O., 2001. Random forest. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cao, L., Pan, J., Li, R., Li, J., Li, Z., 2018. Integrating airborne LiDAR and optical data to estimate forest aboveground biomass in arid and semi-arid regions of China. Remote Sens. 10. https://doi.org/10.3390/rs10040532.

Chave, J., Coomes, D., Jansen, S., Lewis, S.L., Swenson, N.G., Zanne, A.E., 2009. Towards a worldwide wood economics spectrum. Ecol. Lett. 12, 351–366. https://doi.org/10.1111/j.1461-0248.2009.01285.x.

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

Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J. Geophys. Res. 89, 6329–6340. https://doi.org/10.1029/JB089iB07p06329.

Clark, M.L., Roberts, D.A., Ewel, J.J., Clark, D.B., 2011. Estimation of tropical rain forest aboveground biomass with small-footprint lidar and hyperspectral sensors. Remote Sens. Environ. 115, 2931–2942. https://doi.org/10.1016/j.rse.2010.08.029.

Cohen, J., 1988. Statistical Power Analysis for Behavioural Sciences, 2nd edition. Erlbaum, Hillsdale, NJ.

Dalponte, M., Bruzzone, L., Gianelle, D., 2012. 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. 123, 258–270. https://doi.org/10.1016/j.rse.2012.03.013.

De Jong, S.M., Pebesma, E.J., Lacaze, B., 2003. Above-ground biomass assessment of Mediterranean forests using airborne imaging spectrometry: the DAIS Peyne experiment. Int. J. Remote Sens. 24, 1505–1520. https://doi.org/10.1080/01431160210145560.

d'Oliveira, M.V.N., Reutebuch, S.E., McGaughey, R.J., Andersen, H.E., 2012. Estimating forest biomass and identifying low-intensity logging areas using airborne scanning lidar in Antimary State Forest, Acre State, Western Brazilian Amazon. Remote Sens. Environ. 124, 479–491. https://doi.org/10.1016/j.rse.2012.05.014.

Duzan, H., Shariff, N.S.B.M., 2015. Ridge regression for solving the multicollinearity problem: review of methods and models. J. Appl. Sci. 15, 393–404. https://doi.org/10.3923/jas.2015.392.404.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., Koch, B., 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. Remote Sens. Environ. 154, 102–114. https://doi.org/10.1016/j.rse.2014.07.028.

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

Feng, Y., Lu, D., Chen, Q., Keller, M., Moran, E., Dos-Santos, M.N., Bolfe, E.L., Batistella, M., 2017. Examining effective use of data sources and modeling algorithms for improving biomass estimation in a moist tropical forest of the Brazilian Amazon. Int. J. Digit. Earth 10, 996–1016. https://doi.org/10.1080/17538947.2017.1301581.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.

Galvão, L.S., Formaggio, A.R., Tisot, D.A., 2005. Discrimination of sugarcane varieties in southeastern Brazil with EO-1 Hyperion data. Remote Sens. Environ. 94, 523–534. https://doi.org/10.1016/j.rse.2004.11.012.

Galvão, L.S., Breuning, F.M., dos Santos, J.R., Moura, Y.M., 2013. View-illumination effects on hyperspectral vegetation indices in the Amazonian tropical forest. Int. J. Appl. Earth Obs. Geoinf. 21, 291–300. https://doi.org/10.1016/j.jag.2012.07.005.

Gamon, J.A., Peñuelas, J., Field, C.B., 1992. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sens. Environ. 41, 35–44. https://doi.org/10.1016/0034-4257(92)90059-S.

Gao, B.C., 1996. NDWI-A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3.

Ghosh, A., Fassnacht, F.E., Joshi, P.K., Koch, B., 2014. A framework for mapping tree species combining hyperspectral and LiDAR data: role of selected classifiers and sensor across three spatial scales. Int. J. Appl. Earth Obs. Geoinf. 26, 49–63. https://doi.org/10.1016/j.jag.2013.05.017.

Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58, 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7.

Gitelson, A.A., Kaufman, Y.J., Stark, R., Rundquist, D., 2002. Novel algorithms for remote estimation of vegetation fraction. Remote Sens. Environ. 80, 76–87. https://doi.org/10.1016/S0034-4257(01)00289-9.

Gitelson, A.A., Keydan, G.P., Merzlyak, M.N., 2006. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. Geophys. Res. Lett. 33, 2–6. https://doi.org/10.1029/2006GL026457.

Gong, P., Pu, R., Biging, G.S., Larrieu, M.R., 2003. Estimation of forest leaf area index using vegetation indices derived from Hyperion hyperspectral data. IEEE Trans. Geosci. Remote Sens. 41, 1355–1362. https://doi.org/10.1109/TGRS.2003.812910.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

Guyot, G., Baret, F., 1988. Utilisation de la haute resolution spectrale pour suivre l'etat des couverts vegetaux. In: Guyenne, T.D., Hunt, J.J. (Eds.), Spectral Signatures of Objects in Remote Sensing. 287. pp. 279–286 of ESA Special Publication.

Hansen, E.H., Gobakken, T., Bollandsås, O.M., Zahabu, E., Næsset, E., 2015. Modeling aboveground biomass in dense tropical submontane rainforest using airborne laser scanner data. Remote Sens. 7, 788–807. https://doi.org/10.3390/rs70100788.

Houghton, R.A., Hall, F., Goetz, S.J., 2009. Importance of biomass in the global carbon cycle. J. Geophys. Res. Biogeosci. 114, 1–13. https://doi.org/10.1029/2009JG000935.

Huete, A., Didan, K., Miura, T., Rodriguez, E., Gao, X., Ferreira, L., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2.

Hunter, M.O., Keller, M., Victoria, D., Morton, D.C., 2013. Tree height and tropical forest biomass estimation. Biogeosciences 10, 8385–8399. https://doi.org/10.5194/bg-10-8385-2013.

Isenburg, M., 2018. LAStools - efficient LiDAR processing software (version 171030, unlicensed). obtained from. http://rapidlasso.com/LAStools.

John, R., Chen, J., Giannico, V., Park, H., Xiao, J., Shirkey, G., Ouyang, Z., Shao, C., Lafortezza, R., Qi, J., 2018. Grassland canopy cover and aboveground biomass in Mongolia and Inner Mongolia: spatiotemporal estimates and controlling factors. Remote Sens. Environ. 213, 34–48. https://doi.org/10.1016/j.rse.2018.05.002.

Jordan, C.F., 1969. Derivation of leaf-area index from quality of light on the forest floor. Ecology 50, 663–666. https://doi.org/10.2307/1936256.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kernlab – an S4 package for Kernel methods in R. J. Stat. Softw. 11, 1–20.

Koch, B., 2010. Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. ISPRS J. Photogramm. Remote Sens. 65, 581–590. https://doi.org/10.1016/j.isprsjprs.2010.09.001.

Kokaly, R.F., Asner, G.P., Ollinger, S.V., Martin, M.E., Wessman, C.A., 2009. Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. Remote Sens. Environ. 113, S78–S91. https://doi.org/10.1016/j.rse.2008.10.018.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Zeitschrift 15, 259–263. https://doi.org/10.1127/0941-2948/2006/0130.

Kronseder, K., Ballhorn, U., Böhmc, V., Siegert, F., 2012. Above ground biomass estimation across forest types at different degradation levels in central Kalimantan using lidar data. Int. J. Appl. Earth Obs. Geoinf. 18, 37–48. https://doi.org/10.1016/j.jag.2012.01.010.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26.

Latifi, H., Fassnacht, F., Koch, B., 2012. Forest structure modeling with combined airborne hyperspectral and LiDAR data. Remote Sens. Environ. 121, 10–25. https://doi.org/10.1016/j.rse.2012.01.015.

le Maire, G., François, C., Soudani, K., Berveiller, D., Pontailler, J.Y., Bréda, N., Genet, H., Davi, H., Dufrêne, E., 2008. Calibration and validation of hyperspectral indices for the estimation of broadleaved forest leaf chlorophyll content, leaf mass per area, leaf area index and leaf canopy biomass. Remote Sens. Environ. 112, 3846–3864. https://

doi.org/10.1016/j.rse.2008.06.005.
Le Quéré, C., Andrew, R.M., Friedlingstein, P., Sitch, S., Hauck, J., Pongratz, J., Pickers, P.A., et al., 2018. Global carbon budget 2018. Earth Syst. Sci. Data Discuss. 10, 2141–2194.
Lehnert, L.W., Meyer, H., Bendix, J., 2018. Hsdar: Manage, Analyse and Simulate Hyperspectral Data in R. R Package Version 0.7.1.
Li, M., Im, J., Quackenbush, L.J., Liu, T., 2014. Forest biomass and carbon stock quantification using airborne LiDAR data: a case study over Huntington wildlife forest in the Adirondack Park. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 3143–3156. https://doi.org/10.1109/JSTARS.2014.2304642.
Longo, M., Keller, M., dos-Santos, M.N., Leitold, V., Pinagé, E.R., Baccini, A., Saatchi, S., Nogueira, E.M., Batistella, M., Morton, D.C., 2016. Aboveground biomass variability across intact and degraded forests in the Brazilian Amazon. Glob. Biogeochem. Cycles 30, 1639–1660. https://doi.org/10.1002/2016GB005465.
Lu, D., Chen, Q., Wang, G., Liu, L., Li, G., Moran, E., 2014. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. Int. J. Digit. Earth 9, 63–105. https://doi.org/10.1080/17538947.2014.990526.
Luo, S., Wang, C., Xi, X., Pan, F., Peng, D., Zou, J., Nie, S., Qin, H., 2017a. Fusion of airborne LiDAR data and hyperspectral imagery for aboveground and belowground forest biomass estimation. Ecol. Indic. 73, 378–387. https://doi.org/10.1016/j.ecolind.2016.10.001.
Luo, S., Wang, C., Xi, X., Pan, F., Qian, M., Peng, D., Nie, S., Qin, H., Lin, Y., 2017b. Retrieving aboveground biomass of wetland Phragmites australis (common reed) using a combination of airborne discrete-return LiDAR and hyperspectral data. Int. J. Appl. Earth Obs. Geoinf. 58, 107–117. https://doi.org/10.1016/j.jag.2017.01.016.
Magurran, A.E., 2004. Measuring Biological Diversity. Blackwell Science, Ltd.
Mauya, E.W., Hansen, E.H., Gobakken, T., Bollandsås, O.M., Malimbwi, R.E., Næsset, E., 2015. Effects of field plot size on prediction accuracy of aboveground biomass in airborne laser scanning-assisted inventories in tropical rain forests of Tanzania. Carbon Balance Manag 10, 1–14. https://doi.org/10.1186/s13021-015-0021-x.
McGaughey, R.J., 2014. FUSION/LDV: Software for LIDAR Data Analysis and Visualization, Manual. USFS Pacific Northwest Research Station, Seattle, Wash.
Merton, R.N., 1998. Monitoring community hysteresis using spectral shift analysis and the red-edge vegetation stress index. In: Proceedings of the Seventh Annual JPL Airborne Earth Science Workshop. NASA, jet Propulsion Laboratory, Pasadena, California, USA. 12–16 January 1998.
Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B., Rakitin, V.Y., 1999. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. Physiol. Plant. 106, 135–141. https://doi.org/10.1034/j.1399-3054.1999.106119.x.
Monnet, J.M., Chanussot, J., Berger, F., 2011. Support vector regression for the estimation of forest stand parameters using airborne laser scanning. IEEE Geosci. Remote Sens. Lett. 8, 580–584. https://doi.org/10.1109/LGRS.2010.2094179.
Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. ISPRS J. Photogramm. Remote Sens. 66, 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001.
Næsset, E., 2009. Effects of different sensors, flying altitudes, and pulse repetition frequencies on forest canopy metrics and biophysical stand properties derived from small-footprint airborne laser data. Remote Sens. Environ. 113, 148–159. https://doi.org/10.1016/j.rse.2008.09.001.
Næsset, E., Gobakken, T., 2008. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. Remote Sens. Environ. 112, 3079–3090. https://doi.org/10.1016/j.rse.2008.03.004.
Nagler, P.L., Daughtry, C.S.T., Goward, S.N., 2000. Plant litter and soil reflectance. Remote Sens. Environ. 71, 207–215. https://doi.org/10.1016/S0034-4257(99)00082-6.
Ometto, J.P., Aguiar, A.P., Assis, T., Soler, L., Valle, P., Tejada, G., Lapola, D.M., Meir, P., 2014. Amazon forest biomass density maps: tackling the uncertainty in carbon emission estimates. Clim. Chang. 124, 545–560. https://doi.org/10.1007/s10584-014-1058-7.
Osborne, J., Waters, E., 2002. Four assumptions of multiple regression that researchers should always test. Pract. Assessment, Res. Eval. 8, 1. http://pareonline.net/getvn.asp?v=8&n=2.
Pan, Y., Birdsey, R.A., Phillips, O.L., Jackson, R.B., 2013. The structure, distribution, and biomass of the world's forests. Annu. Rev. Ecol. Evol. Syst. 44, 593–622. https://doi.org/10.1146/annurev-ecolsys-110512-135914.
Peñuelas, J., Piñol, J., Ogaya, R., Filella, I., 1997. Estimation of plant water concentration by the reflectance Water Index WI (R900/R970). Int. J. Remote Sens. 18, 2869–2875. https://doi.org/10.1080/014311697217396.
Ploton, P., Barbier, N., Couteron, P., Antin, C.M., Ayyappan, N., Balachandran, N., 2017. Toward a general tropical forest biomass prediction model from very high resolution optical satellite images. Remote Sens. Environ. 200, 140–153. https://doi.org/10.1016/j.rse.2017.08.001.
Poorter, L., Bongers, F., Aide, T.M., Almeyda Zambrano, A.M., Balvanera, P., Becknell, J.M., Boukili, V., et al., 2016. Biomass resilience of Neotropical secondary forests. Nature 530, 211–214. https://doi.org/10.1038/nature16512.
Psomas, A., Kneubühler, M., Huber, S., Itten, K., 2011. Hyperspectral remote sensing for estimating aboveground biomass and for exploring species richness patterns of grassland habitats. Int. J. Remote Sens. 32, 9007–9031. https://doi.org/10.1080/01431161.2010.532172.
Pu, R., Kelly, M., Anderson, G.L., Gong, P., 2008. Using CASI hyperspectral imagery to detect mortality and vegetation stress associated with a new hardwood forest disease. Photogramm. Eng. Remote. Sens. 74, 65–75. https://doi.org/10.14358/PERS.74.1.65.
Quesada, C.A., Lloyd, J., Anderson, L.O., Fyllas, N.M., Schwarz, M., Czimczik, C.I., 2011. Soils of Amazonia with particular reference to the RAINFOR sites. Biogeosciences 8, 1415–1440. https://doi.org/10.5194/bg-8-1415-2011.
Réjou-Méchain, M., Tanguy, A., Piponiot, C., Chave, J., Hérault, B., 2017. Biomass: an R package for estimating above-ground biomass and its uncertainty in tropical forests. Methods Ecol. Evol. 8, 1163–1167. https://doi.org/10.1111/2041-210X.12753.
Roth, K.L., Roberts, D.A., Dennison, P.E., Peterson, S.H., Alonzo, M., 2015. The impact of spatial resolution on the classification of plant species and functional types within imaging spectrometer data. Remote Sens. Environ. 171, 45–57. https://doi.org/10.1016/j.rse.2015.10.004.
Rouse, J.W., et al., 1973. Monitoring vegetation systems in the great plains with ERTS. In: ERTS-1 SYMPOSIUM, n. 3, Washington, DC. Proceedings…. NASA, Washington, pp. 309–317.
Roussel, J.R., Auty, D., 2018. lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications. R Package Version 1.6.1. https://CRAN.R-project.org/package=lidR.
RuleQuest, 2018. Data Mining with Cubist. https://www.rulequest.com/cubist-info.html.
Rutishauser, E., Hérault, B., Baraloto, C., Blanc, L., Descroix, L., Sotta, E.D., Ferreira, J., et al., 2015. Rapid tree carbon stock recovery in managed Amazonian forests. Curr. Biol. 25, R787–R788. https://doi.org/10.1016/j.cub.2015.07.034.
Serrano, L., Peñuelas, J., Ustin, S.L., 2002. Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: decomposing biochemical from structural signals. Remote Sens. Environ. 81, 355.
Singh, K.K., Chen, G., Vogler, J.B., Meentemeyer, R.K., 2016. When big data are too much: effects of LiDAR returns and point density on estimation of forest biomass. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 3210–3218. https://doi.org/10.1109/JSTARS.2016.2522960.
Smith, M.L., Ollinger, S.V., Martin, M.E., Aber, J.D., Hallett, R.A., Goodale, C.L., 2002. Direct estimation of aboveground forest productivity through hyperspectral remote sensing of canopy nitrogen. Ecol. Appl. 12, 1286–1302.
Stark, S.C., Leitold, V., Wu, J.L., Hunter, M.O., de Castilho, C.V., Costa, F.R.C., Mcmahon, S.M., et al., 2012. Amazon forest carbon dynamics predicted by profiles of canopy leaf area and light environment. Ecol. Lett. 15, 1406–1414. https://doi.org/10.1111/j.1461-0248.2012.01864.x.
Stovall, A.E.L., Shugart, H.H., 2018. Improved biomass calibration and validation with terrestrial lidar: implications for future LiDAR and SAR missions. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 3527–3537. https://doi.org/10.1109/JSTARS.2018.2803110.
Swatantran, A., Dubayah, R., Roberts, D., Hofton, M., Blair, J.B., 2011. Mapping biomass and stress in the Sierra Nevada using lidar and hyperspectral data fusion. Remote Sens. Environ. 115, 2917–2930. https://doi.org/10.1016/j.rse.2010.08.027.
Thomas, V., Treitz, P., McCaughey, J.H., Morrison, I., 2006. Mapping stand-level forest biophysical variables for a mixedwood boreal forest using lidar: an examination of scanning density. Can. J. For. Res. 36, 34–47. https://doi.org/10.1139/x05-230.
Torabzadeh, H., Morsdorf, F., Schaepman, M.E., 2014. Fusion of imaging spectroscopy and airborne laser scanning data for characterization of forest ecosystems - a review. ISPRS J. Photogramm. Remote Sens. 97, 25–35. https://doi.org/10.1016/j.isprsjprs.2014.08.001.
Tropical Rainfall Measuring Mission (TRMM), 2011. TRMM (TMPA/3B43) Rainfall Estimate L3 1 Month 0.25 Degree x 0.25 Degree V7. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD. https://doi.org/10.5067/TRMM/TMPA/MONTH/7.
Ustin, S.L., Roberts, D.A., Gamon, J.A., Asner, G.P., Green, R.O., 2004. Using imaging spectroscopy to study ecosystem processes and properties. Bioscience 54, 523. https://doi.org/10.1641/0006-3568(2004)054[0523:UISTSE]2.0.CO;2.
Vaglio Laurin, G., Chen, Q., Lindsell, J.A., Coomes, D.A., Frate, F. Del, Guerriero, L., Pirotti, F., et al., 2014. Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. ISPRS J. Photogramm. Remote Sens. 89, 49–58. https://doi.org/10.1016/j.isprsjprs.2014.01.001.
Vieira, I.C.G., De Almeida, A.S., Davidson, E.A., Stone, T.A., Reis De Carvalho, C.J., Guerrero, J.B., 2003. Classifying successional forests using Landsat spectral properties and ecological characteristics in eastern Amazônia. Remote Sens. Environ. 87, 470–481. https://doi.org/10.1016/j.rse.2002.09.002.
Vogelmann, J.E., Rock, B.N., Moss, D.M., 1993. Red edge spectral measurements from sugar maple leaves. Int. J. Remote Sens. 14, 1563–1575. https://doi.org/10.1080/01431169308953986.
Wang, H., Glennie, C., 2015. Fusion of waveform LiDAR data and hyperspectral imagery for land cover classification. ISPRS J. Photogramm. Remote Sens. 108, 1–11. https://doi.org/10.1016/j.isprsjprs.2015.05.012.
Wilson, M.F.J., O'Connell, B., Brown, C., Guinan, J.C., Grehan, A.J., 2007. Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. Mar. Geod. https://doi.org/10.1080/01490410701295962.
Zanne, A.E., Lopez-Gonzalez, G., Coomes, D.A., Ilic, J., Jansen, S., Lewis, S.L., Miller, R.B., Swenson, N.G., Wiemann, M.C., Chave, J., 2009. Data from: towards a worldwide wood economics spectrum. Dryad Data Repository. https://doi.org/10.5061/dryad.234.
Zhang, Z., Cao, L., She, G., 2017a. Estimating forest structural parameters using canopy metrics derived from airborne LiDAR data in subtropical forests. Remote Sens. 9. https://doi.org/10.3390/rs9090940.
Zhang, W., Hu, B., Woods, M., Brown, G., 2017b. Characterizing forest succession stages for wildlife habitat assessment using multispectral airborne imagery. Forests 8. https://doi.org/10.3390/f8070234.
Zolkos, S.G., Goetz, S.J., Dubayah, R., 2013. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sens. Environ. 128, 289–298. https://doi.org/10.1016/j.rse.2012.10.017.