INTERNATIONAL JOURNAL OF DIGITAL EARTH, 2017
http://dx.doi.org/10.1080/17538947.2017.1301581

Examining effective use of data sources and modeling algorithms for improving biomass estimation in a moist tropical forest of the Brazilian Amazon

Yunyun Feng[a,b], Dengsheng Lu [a,b], Qi Chen[a,c], Michael Keller[d,e], Emilio Moran[b], Maiza Nara dos-Santos[d], Edson Luis Bolfe[f] and Mateus Batistella[g]

[a]The Nurturing Station for the State Key Laboratory of Subtropical Silviculture, Key Laboratory of Carbon Cycling in Forest Ecosystems and Carbon Sequestration of Zhejiang Province, School of Environmental & Resource Sciences, Zhejiang Agriculture and Forestry University, Lin An, People's Republic of China; [b]Center for Global Change and Earth Observations, Michigan State University, East Lansing, MI, USA; [c]Department of Geography, University of Hawaii at Manoa, Honolulu, HI, USA; [d]Brazilian Agricultural Research Corporation – Embrapa Informática Agropecuária, Campinas, Brazil; [e]USDA Forest Service, International Institute of Tropical Forestry, Rio Piedras, PR, USA; [f]Brazilian Agricultural Research Corporation – Embrapa, Secretariat of Intelligence and Macro Strategy, Brasília, Brazil; [g]Brazilian Agricultural Research Corporation (Embrapa), Brasília, Brazil

ABSTRACT

Previous research has explored the potential to integrate lidar and optical data in aboveground biomass (AGB) estimation, but how different data sources, vegetation types, and modeling algorithms influence AGB estimation is poorly understood. This research conducts a comparative analysis of different data sources and modeling approaches in improving AGB estimation. RapidEye-based spectral responses and textures, lidar-derived metrics, and their combination were used to develop AGB estimation models. The results indicated that (1) overall, RapidEye data are not suitable for AGB estimation, but when AGB falls within 50–150 Mg/ha, support vector regression based on stratification of vegetation types provided good AGB estimation; (2) Lidar data provided stable and better estimations than RapidEye data; and stratification of vegetation types cannot improve estimation; (3) The combination of lidar and RapidEye data cannot provide better performance than lidar data alone; (4) AGB ranges affect the selection of the best AGB models, and a combination of different estimation results from the best model for each AGB range can improve AGB estimation; (5) This research implies that an optimal procedure for AGB estimation for a specific study exists, depending on the careful selection of data sources, modeling algorithms, forest types, and AGB ranges.

ARTICLE HISTORY
Received 2 January 2017
Accepted 22 February 2017

KEYWORDS
Lidar; RapidEye; aboveground biomass; moist tropical forest; support vector regression; random forest; linear regression; stratification

1. Introduction

Aboveground biomass (AGB) estimation has gained substantial attention in the past two decades due to its importance in studies related to climate change and ecosystem services, thus much research has been conducted to explore approaches to accurately estimate AGB using remotely sensed data (Lu 2006; Barbosa et al. 2014; Kumar et al. 2015; Lu et al. 2016; Timothy et al. 2016). In particular, the satellite data – optical sensor multispectral data (e.g. Landsat) and long wavelength radar data (e.g. ALOS PALSAR L-band) – may be the most commonly used sources for AGB estimation at

CONTACT Dengsheng Lu luds@zafu.edu.cn School of Environmental & Resource Sciences, Zhejiang Agriculture and Forestry University, Lin An 311300, People's Republic of China

the regional and local scales (Foody, Boyd, and Cutler 2003; Lu 2005; Avitabile et al. 2012; Kelsey and Neff 2014; Dube and Mutanga 2015; Zhao et al. 2016a). However, the use of optical and radar data for AGB estimation becomes difficult when AGB reaches the saturation values such as 150 Mg/ha (Lu, Batistella, and Moran 2005; Zhao et al. 2016a, 2016b). Recently, airborne lidar became another important data source for AGB estimation at the local scale (Lim et al. 2003; Chen 2013; Lu et al. 2016). Because lidar-estimated canopy height is strongly related to forest AGB even when AGB is high (Chen 2013; Lu et al. 2016), lidar has been regarded as the most accurate remote sensing approach for AGB estimation (Lim et al. 2003; Koch 2010; Gleason and Im 2011; Chen 2013).

Different sensor data have their own characteristics, for example, optical sensor data capture land surface features that are suitable for land cover classification and radar can penetrate forest canopy to a certain depth to capture branch and stem features, depending on the wavelength (Lu et al. 2016). However, both optical and radar data cannot effectively obtain forest height features that are critical for AGB estimation, especially when AGB is relatively high (Zhao et al. 2016a). Lidar complements the shortcomings of existing optical and radar systems by providing canopy height information, thus, lidar data do not have the data saturation problem and are often used in AGB estimation at the local to landscape scales (Chen 2013; Lu et al. 2016). A long-standing research question in lidar remote sensing for AGB estimation is to understand whether the addition of passive optical imagery to airborne lidar can further improve AGB modeling performance (Lu et al. 2016).

Generally, different approaches may be used to integrate lidar and other data sources such as optical data (Chen 2013; Zhang and Lin 2016). For example, optical sensor images can be used to conduct vegetation classification, and lidar-based AGB estimation models can be established based on different vegetation types (Chen et al. 2016). By doing so, some previous studies have shown that the integration of both types of data can improve the model performance (Chen et al. 2012). However, such an approach depends on the accuracy of vegetation classification. Another option is that both lidar-based metrics and optical sensor-based variables are directly used as predictors to establish the AGB estimation models (Clark et al. 2011; Vaglio Laurin et al. 2014). Nevertheless, with such a method, previous studies indicated that passive optical sensor data did (Vaglio Laurin et al. 2014) or did not (Clark et al. 2011) improve the AGB model performance when combined with airborne lidar. For example, Fassnacht et al. (2014) found that the combination of lidar and passive optical sensor data did not result in better performance. Hyde et al. (2006) found that combination of lidar and QuickBird image did not improve AGB estimation in mixed coniferous forests in California; lidar data alone provided a better performance. Clark et al. (2011) found that lidar and hyperspectral data fusion has lower estimation accuracy than lidar data alone in the moist tropical forest in Costa Rica. A similar conclusion was also obtained by Latifi, Fassnacht, and Koch (2012) in temperate coniferous forests in Germany using lidar and hyperspectral data. However, other research showed that the combination of multispectral (Popescu, Randolph, and John 2004; Xu et al. 2015) or hyperspectral (Vaglio Laurin et al. 2014) imagery with airborne lidar improved the estimation of forest volume and/or AGB. Thus, more research is needed along this direction, especially over different forest types and biomes.

Many algorithms have been used for AGB estimation as summarized in Lu et al. (2016). Linear regression models have often been used, but the relationships between AGB and remote sensing variables may be not linear, thus it has problems of overestimation or underestimation when AGB is small or very high (Zhao et al. 2016a). Therefore, much research has shifted to explore the use of nonparametric algorithms such as K-nearest neighbor, artificial neural network, support vector regression (SVR), and random forest (RF) (Breidenbach et al. 2010; Vauhkonen et al. 2010; Mitchard et al. 2011; Lu et al. 2016). Li et al. (2014) conducted a comparative analysis of different modeling approaches (e.g. ordinary least squares, generalized additive model, Cubist, bagging, boosted regression trees, RF, and SVR) for AGB and carbon estimation, and found that SVR provided the best performance. Gleason and Im (2012) compared four modeling approaches – linear mixed-effects regression, RF, SVR, and Cubist for AGB estimation based on single tree species and canopy scales, and found that SVR provided the best performance at the canopy scale, but different modeling

approaches had similar performance at individual tree scale. Although previous literature has summarized the major approaches for modeling AGB (e.g. Sileshi 2014; Lu et al. 2016), it is unclear how different data sources, vegetation types, and modeling approaches affect AGB estimation results, especially in the secondary forests and agroforestry systems in the moist tropical regions.

Relatively less research has been done for AGB estimation in moist tropical regions yet. The main reasons may be (1) the difficulty in collection of ground truth data; (2) the frequent cloud cover problem resulting in difficulty of capturing good-quality optical sensor data, or lack of radar and lidar data; and (3) complex tree species composition and forest stand structures. In recent years, some high temporal resolution satellite images such as RapidEye with high spatial resolution offer new opportunities for conducting AGB estimation in the moist tropical regions. Also, recent availability of airborne lidar data for typical sites in the Amazon basin provides new chances for exploring the approaches to improve AGB estimation through a comparative analysis of different data sources and modeling algorithms. However, many questions remain to be answered, for example, can the combination of lidar and optical sensor data improve AGB estimation in the moist tropical region? Do different AGB ranges affect AGB estimation? Do vegetation types affect the selection of AGB modeling algorithm? Do nonparametric algorithms such as SVR have better performance in AGB estimation than conventional linear regression model? Therefore, the primary goal of this research is to explore the optimal procedure for AGB modeling suitable for the moist tropical region through a comparative analysis of different algorithms (i.e. linear, nonlinear, RF, and SVR) and data sources (lidar, RapidEye, and their combination) under stratification and non-stratification conditions. Specifically, this research will examine (1) the performance of RapidEye data alone in AGB estimation, (2) the incorporation of RapidEye and lidar in improving AGB estimation, (3) the roles of nonparametric algorithms in AGB modeling, (4) the role of stratification of vegetation types in improving AGB estimation, and (5) the impacts of AGB ranges on the selection of AGB modeling algorithms. This study is novel, in that little research has been done to address the synergy of airborne lidar and passive optical sensor data for AGB modeling from the perspectives of (1) using relatively high spatial resolution multispectral satellite data such as RapidEye; (2) focusing on Amazonian secondary forests and agroforestry systems; and (3) identifying a proper procedure for AGB estimation.

2. Study area

The study area (about 10 km^2) is located in the municipality of Tomé-Açu, approximately 240 km to the south of Belém, the capital of the state of Pará in Brazil (Figure 1). According to the Köppen classification, Tomé-Açu has a humid mesothermal climate (Ami), with average annual relative air humidity rate of about 85% and temperatures of 26°C. The average annual rainfall is about 2300 mm. The topography varying from 14 to 96 m is characterized by low flat plateaus, terraces, and lowlands (Rodrigues et al. 2011). The region was originally covered by lowland dense evergreen forest (Batistella, Bolfe, and Moran 2013), but now it is mainly a mosaic of pasture lands, secondary forests and agroforestry systems.

Tomé-Açu was settled by Japanese immigrants who implanted horticulture in the 1920s and, later, black pepper (*Piper nigrum* L.). In 1931, the local farmers created the Agricultural Cooperative of Tomé-Açu (Camta) and the region became an important producer of black pepper in the world. Beginning in the 1980s, the local farmers developed a variety of agroforestry systems. Recently, these systems include about 70 crop species, such as fruit crops (acerola, orange, papaya, melon, cupuaçu, and passion fruit), oil palm and other native and exotic trees including teak in hundreds of polycultural combinations (Bolfe and Batistella 2011). The agroforestry systems include various life forms, allowing the permanent use of farm fields (Yamada and Gholz 2002). The farmers also use the land for cattle ranching and eventually pasture areas are followed by secondary forests, which can be important for carbon sequestration and other ecosystems services.

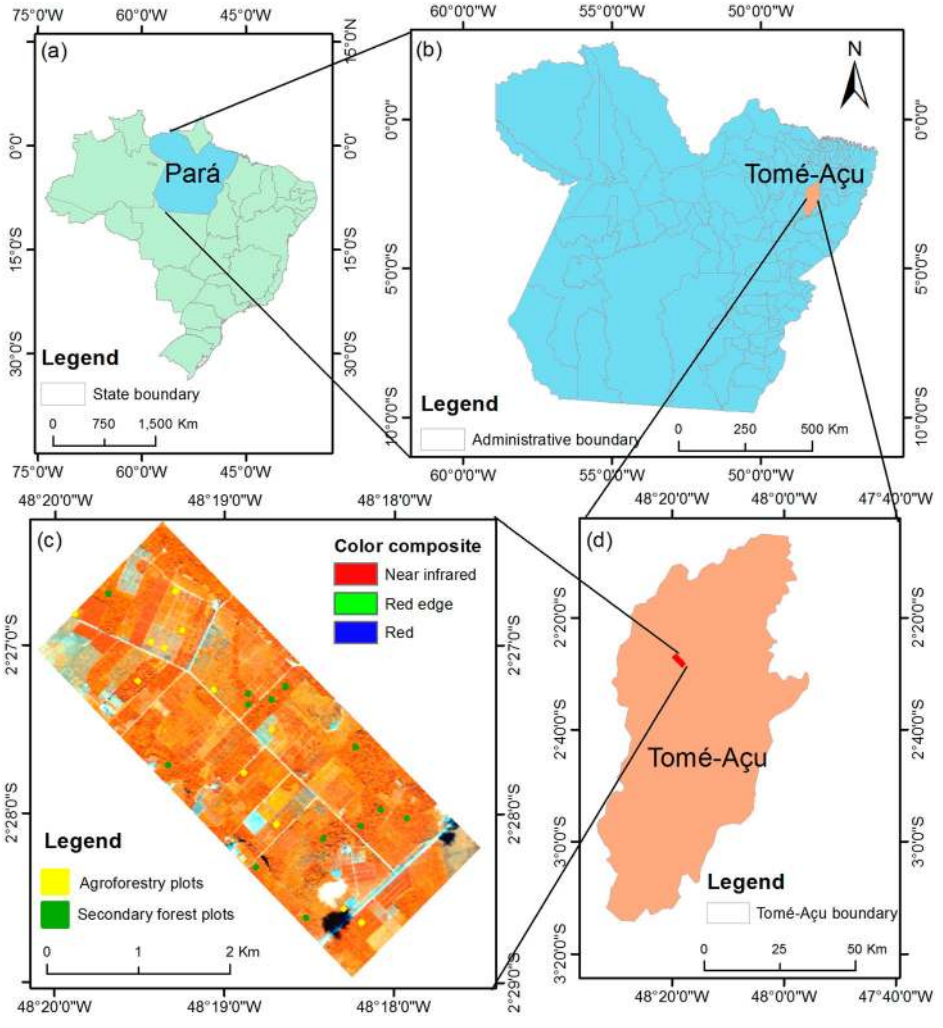

Figure 1. The location of study area – Tomé-Açu (d), para state (b), Brazil (a) (Note: (c) illustrates the study area using the color composite with near-infrared, red edge, and red spectral bands as RGB).

3. Methods

Figure 2 illustrates the strategy of modeling AGB using different scenarios. The major steps include (1) collection of different data sources such as sample plots, lidar, and RapidEye; preprocessing of these datasets; (2) extraction of variables from lidar and RapidEye; (3) selection of variables suitable for AGB estimation using different modeling approaches; and (4) comparative analysis of different models and evaluation of AGB estimation results.

3.1. Data collection and preprocessing

3.1.1. Collection and calculation of AGB at sample plots

Forest inventory for the selected sample plots was conducted in 2014 and 2015, based on the methods indicated in the standardized protocols for tropical forest (Walker et al. 2011). As shown in Figure 1, a total of 25 samples with each field plot size of 30 m × 30 m were established

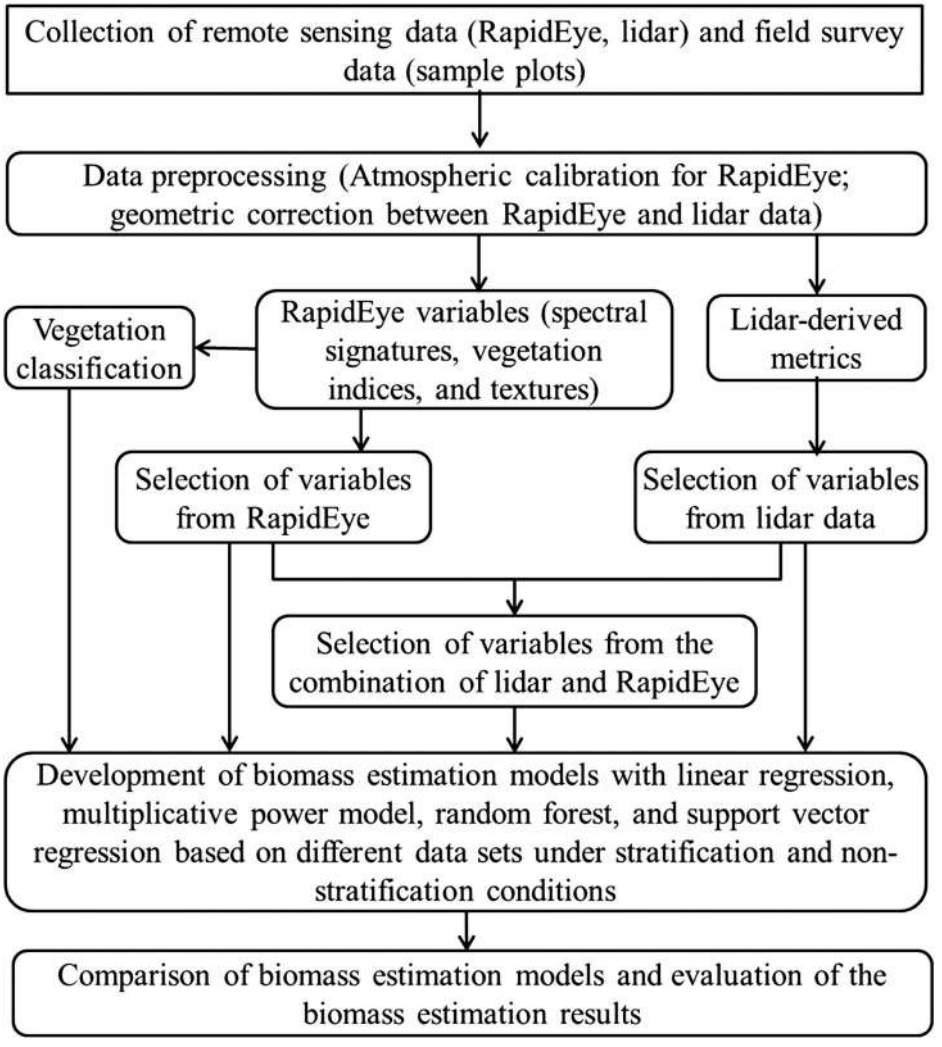

Figure 2. Strategy of biomass estimation using different modeling approaches (linear regression, multiplicative power model, random forest, and support vector regression) based on various data sources (lidar, RapidEye, and their combination) under non-stratification and stratification of vegetation types.

using a differential GNSS (Trimble GeoXH-6000) and the sample plots were distributed in different locations within the study area of about 10 km^2. The plot locations were identified to cover different ages and floristic compositions (Bolfe, Batistella, and Ferreira 2012). Within each field plot, diameter at breast height (DBH) was measured with metric tapes. Tree height was measured with tape and clinometer with the estimated accuracy of 10% of the tree height (Hunter et al. 2013). For the trees within agroforestry sample plots, the location, tree species, tree height, DBH, and crown size were measured. For secondary forest sample plots, the same variables, except tree height, were measured.

Based on field measurements for each plot, we estimated AGB using allometric equations for three categories of species: palm, liana, and dicotyledonous trees. The wood densities of different tree species were obtained from the global tree wood density database (Chave et al. 2009; Zanne et al. 2009). The following allometric equations were used to calculate AGB for individual trees:

For palm species (Nascimento and Laurance 2002):

$$\mathrm{AGB(kg)} = 0.001(\exp(0.9285\ln(D^2) + 5.7236) \times 1.05001), \quad (1)$$

For liana species (Schnitzer, DeWalt, and Chave 2006):

$$\mathrm{AGB(kg)} = \exp(-1.484 + 2.657\ln(D)), \quad (2)$$

For dicotyledonous trees in agroforestry plots (Chave et al. 2005):

$$\mathrm{AGB(kg)} = 0.0509\rho D^2 H, \quad (3)$$

For dicotyledonous trees in secondary forest plots (Chave et al. 2005):

$$\mathrm{AGB(kg)} = \rho \times \exp(-1.499 + 2.148\ln(D) + 0.207(\ln(D))^2 - 0.028(\ln(D))^3), \quad (4)$$

where ρ is wood density (g/cm^3), D and H are DBH (cm) and total tree height (m). The estimated AGB for individual trees within the sample plot were summed and converted to total AGB at plot scale (Mg/ha). A summary of basic statistics of the sample plots is provided in Table 1. Overall, the collected samples have wide AGB ranges from 10.6 to 506.2 Mg/ha. Agroforestry has smaller mean and standard deviation than secondary forest.

3.1.2. Collection and preprocessing of remote sensing data

GEOID Laser Mapping collected airborne lidar data on 2 September 2013 using an Optech ALTM Orion M200 sensor with an integrated GPS, IMU system. The average flight altitude was 853 m above ground and the instrument total field of view was 11°. GEOID simultaneously collected global positioning system data at a ground station to permit post-processing for estimated horizontal and vertical accuracies (1 σ) of 0.3 and 0.15 m, respectively. Over the area of interest, GEOID acquired an average of 24 returns m^{-2} and at least 99.5% of the area had a minimum of 4 returns m^{-2}.

RapidEye data with 5 m spatial resolution were acquired on 3 August 2012. The atmospheric calibration was conducted using FLAASH, converting digital number to surface reflectance (Nascimento and Laurance 2002; Schnitzer, DeWalt, and Chave 2006). This imagery was geometrically registered into lidar using the second order polynomial algorithm based on 15 ground control points. A root mean squared error (RMSE) of 0.7 pixels was obtained. Using the RapidEye data (near infrared image for Figure 3(a)), we visually classified three types of vegetation covers – agroforestry, secondary forest, and others (e.g. oil palm tree) (Figure 3(b)). The palm tree plantation was not included in the agroforestry class because of its considerably different stand structure resulting in the use of different approaches for AGB estimation (Chen et al. 2016).

3.2. Extraction of variables from lidar and RapidEye data

For airborne lidar data, we first generated a digital terrain model (DTM) from the ground returns that were classified by the data provider. The height of each return in the vegetation point cloud was calculated by subtracting their DTM elevations from Z coordinates. Based on the vegetation point cloud heights, a total of 15 variables (Table 2) at the plot scale of 30 m by 30 m were extracted

Table 1. Summary of sample plots used in research.

	Year of field measurements	Number of samples	Mean (Mg/ha)	Standard deviation	AGB range (Mg/ha)
Agroforestry	2014	12	95.6	74.96	10.6–255.5
Secondary Forest	2015	13	204.89	140.49	73.9–506.2
Total		25	152.46	124.63	10.6–506.2

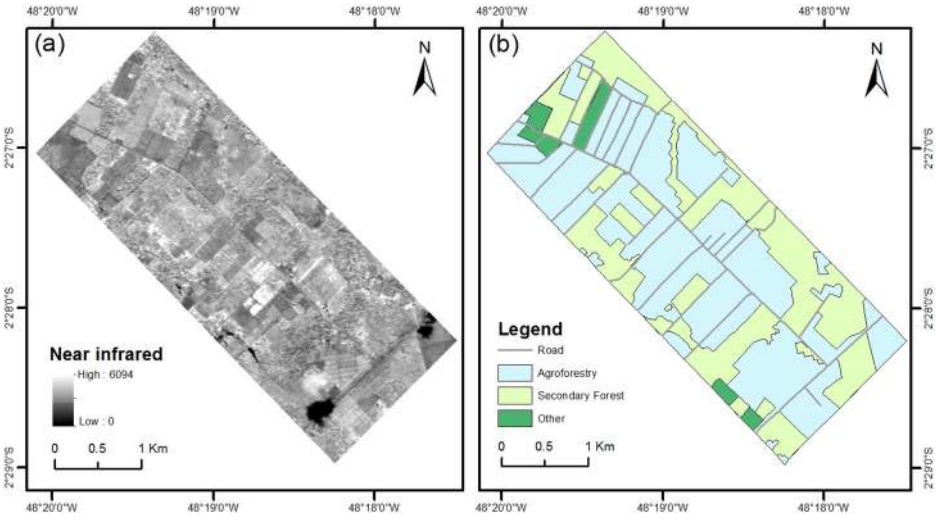

Figure 3. A comparison of near-infrared spectral image (a) from RapidEye and vegetation distribution (b) using the visual interpretation based on RapidEye data.

(Chen et al. 2016). For RapidEye imagery, 10 vegetation indices (Table 2) were extracted. Meanwhile, the textural images using gray-level co-occurrence matrix were extracted from the RapidEye imagery based on each spectral band with a window size of 7 × 7 pixels (Lu and Batistella 2005).

3.3. Development of AGB estimation models

Different modeling approaches as summarized in Lu et al. (2016) may be used for AGB estimation and these approaches have their own characteristics in data requirement and in selection of proper variables. In this research, four modeling approaches – stepwise regression, multiplicative model, RF and SVR were used to develop AGB estimation models based on different data sources – lidar, RapidEye, and their combination.

(1) Statistic-based modeling approaches for AGB estimation

Table 2. A summary of variables used in research.

Variable category		Variables	Key references
Lidar	Metrics	Mean, standard deviation, skewness, kurtosis, quadratic mean height, and percentile height (10th, 20th, … , 90th)	Chen et al. (2016)
RapidEye	Vegetation indices	DVI – Difference Vegetation Index EVI – Enhanced Vegetation Index MNLI – Modified Nonlinear Vegetation Index MSAVI – Modified Soil Adjusted Vegetation Index MSR – Modified Simple Ratio NLI – Nonlinear Vegetation Index OSAVI – Optimization of Soil Adjusted Vegetation Index RDVI – Renormalized Difference Vegetation Index RVI – Ratio Vegetation Index SAVI – Soil Adjusted Vegetation Index	Dube et al. (2014); Li et al. (2015); Zhou et al. (2013, 2015)
	Textures	GLCM-based measures: mean, variance, second moment, dissimilarity, homogeneity, contrast, entropy, and correlation	Lu and Batistella (2005)
Combination	Combination of lidar and RapidEye	All variables from lidar and RapidEye, and vegetation classification data	

As summarized in Table 2, different variables from lidar, RapidEye and their combination can be used for AGB modeling. However, the number of remote sensing-derived variables may be larger than the number of sample plots. This situation will generate difficulty for developing AGB estimation models using regression-based approaches. In order to reduce the number of variables for AGB modeling, the Pearson's product moment correlation coefficients were used to analyze the relationships between AGB and remote sensing variables to identify potential variables that had statistically significant correlation with AGB but had weak correlation between these variables. The stepwise regression analysis was then used to identify variables for AGB estimation models. The AGB at sample plots was used as a dependent variable, and the lidar- and RapidEye-derived variables and their combination were used as independent variables. Linear and multiplicative power models were employed, as illustrated in following equations (Chen et al. 2016):

$$\widehat{AGB} = f(a, x) = a_0 + a_1x_1 + a_2x_2 \cdots + a_ix_i, \tag{5}$$

$$\widehat{AGB} = f(\varphi, z) = \varphi_0 z_1^{\varphi_1} z_2^{\varphi_2} \ldots z_i^{\varphi_3}, \tag{6}$$

where a_i and x_i are parameters and variables used in the linear regression model; φ_i and z_i are parameters and variables used in the multiplicative power model. The models above were developed using the variables from RapidEye, lidar, and the combination of both datasets, respectively. In addition to non-stratification, these modeling approaches were also examined by stratification of vegetation types – agroforestry and secondary forests.

(2) Machine learning-based modeling approaches for AGB estimation

The machine learning approaches, such as neural network, support vector machine, and RF are often used for land cover classification based on remote sensing data (Lu and Weng 2007). These approaches can also be used for estimation of forest attributes such as AGB (Cortes and Vapnik 1995; Yu et al. 2011; Gleason and Im 2012; Li et al. 2014). Of the machine learning-based modeling approaches, RF and SVR are used in this research. RF – a nonparametric ensemble modeling approach is regarded as a robust approach to overcome the overfitting problem and provides a potential solution to better classification or AGB estimation (Breiman 2001). RF can use discrete or continuous datasets, can deal with noise and large datasets (Ismail, Mutanga, and Kumar 2010; Vincenzi et al. 2011), and has been widely used for AGB estimation in recent years (Baccini et al. 2008; Eskelson, Barrett, and Temesgen 2009; Vauhkonen et al. 2010; Avitabile et al. 2012; Hudak et al. 2012; Pflugmacher et al. 2014; Tanase et al. 2014; Chen 2015). In this research, the RF based on the R software was used to develop AGB estimation models for the secondary forests and agroforestry systems in the moist tropical region of the Brazilian Amazon.

Selection of suitable variables is an important part in AGB modeling procedure. In this research, RF was used to identify potential variables because RF can provide ranking of variable importance for AGB estimation. There are three parameters in the RF algorithm: the number of trees (NumTre), minimum number of observations per tree leaf (MinNum), and number of variables randomly selected at each split (NumVar). The MinNum and NumVar are often assigned as default values in RF software package. The analyst will find the optimal NumTre value through examining the trend of errors. When a tentative number such as 1000 was given, an error trend map was produced to show the error distribution. Iterating this procedure until the error trend became stable, this number was then selected as the optimal NumTre value. According to the scenarios of data sources – lidar-derived metrics, RapidEye-derived variables, and their combination, RF was separately used to identify the best variables based on the optimal NumTre and importance ranking for each scenario.

Pearson's product moment correlation analysis was used to examine their linear relationships between the selected potential variables for each scenario. The variables having less importance ranking but having high correlation coefficients with another variable will be removed, and RF model fitting procedure was conducted again. This process repeated until identifying the optimal

combination of the variables with highest and stable R^2 value for each scenario. Because there are many potential variables (spectral bands, texture, and lidar metrics) and SVR did not provide proper approaches for selection of variables, also because both RF and SVR belong to nonparametric algorithms, the variables which were identified using RF were also used in the SVR for AGB modeling.

SVR is another commonly used machine learning algorithm for AGB estimation using remote sensing data (Mountrakis, Im, and Ogole 2011; Marabel and Alvarez-Taboada 2013). This approach can use small training sample data to produce relatively high estimation accuracy (Lu et al. 2016). Therefore, this approach is often used to solve small-sample, nonlinear, and high-dimensional problems (Mountrakis, Im, and Ogole 2011). This feature is especially valuable for the Brazilian Amazon where collection of sample plots is a challenge and the number of collected sample plots is often small (Lu et al. 2012). In the SVR, one critical step is to optimize three parameters: the kernel, precision, and penalty parameters (Cherkassky and Ma 2004). The Grid-search approach is often used to identify the optimal parameters. The kernel option can be linear, polynomial, and radial basis function (RBF). In this research, RBF is used. By setting up the kernel, penalty and precision values repeatedly, cross-validation is used to determine the optimal parameters when the prediction accuracy becomes the highest. The SVR was separately used to estimate AGB for each scenario under stratification and non-stratification conditions.

3.4. Evaluation of AGB estimation results

The coefficient of determination (R^2) is often used to evaluate the goodness of fit of a developed model, while RMSE and relative RMSE (RMSEr) are often used to assess the prediction performance using the developed models (Zolkos, Goetz, and Dubayah 2013). The samples used for validation of predicted AGB should be different from the samples for AGB modeling calibration (Chen et al. 2012; Lu et al. 2016). A common approach is to divide the ground-truth sample plots into k folds and then use cross-validation for model assessment: $k-1$ folds are used for model calibration and the remaining one fold is used for model validation; this process is iterated for k times (here k is the number of sample plots). Note that AGB regression model calibration includes two aspects: (1) selection of variables as model predictors and (2) determination of the model coefficients based on criteria such as minimizing the sum of the squares of model prediction errors (Hastie, Tibshirani, and Friedman 2009). This research used the cross-validation – re-select variables in each iteration and refit model coefficients for $k-1$ folds of plots.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}, \tag{7}$$

$$\mathrm{RMSEr} = \frac{\mathrm{RMSE}}{\bar{y}} \times 100, \tag{8}$$

where $\hat{y}_i$ and y_i are the predicted AGB and corresponding AGB at the sample plot i; $\bar{y}$ is the mean AGB of all sample plots (total number of n).

4. Results

4.1. Comparative analysis of AGB estimation models and spatial distributions

Through comparative analysis of R^2 values for AGB estimation models using different algorithms under various data scenarios, Table 3 summarizes the best AGB estimation model for each scenario. Note that models were developed using all field plots under stratification of vegetation types (i.e. agroforestry ($n = 12$), secondary forests ($n = 13$)) and non-stratification ($n = 25$) and the model R^2 were calculated using the same plots for model calibration and prediction (see the next section

Table 3. Summary of AGB estimation models using different algorithms based on various data sources under stratification and non-stratification conditions.

		Stratification					
		Agroforestry		Secondary forest		Non-stratification	
Methods	Data sets	Models	R^2	Models	R^2	Models	R^2
Linear models	RapidEye	$Y = 117.063T_{b4di} + 3.468$	0.53	$Y = -1746.756T_{b2ho} + 1537.148$	0.43	$Y = -28.988T_{b4me} + 686.798$	0.21
	Lidar	$Y = 29.716H_{std} - 29.511$	0.79	$Y = 25.066H_{qm} - 195.895$	0.86	$Y = 14.489H_{70th} - 52.787$	0.79
	Combination	$Y = 27.448H_{std} + 4.11T_{b4va} - 133.564T_{b2di} + 39.53$	0.94	$Y = 21.266H_{qm} - 667.26T_{b2ho} + 373.784$	0.90	$Y = 17.187H_{70th} + 0.397S_{b2} - 317.751$	0.82
Multi- plicative models	RapidEye	$Y = 584198.323S_{EVI}^{4.988}T_{b5me}^{-3.549}$	0.58	$Y = 48.955T_{b2ho}^{-5.032}$	0.42	$Y = 5.163E18T_{b1me}^{-11.812}T_{b2me}^{-3.097}T_{b1sm}^{-0.868}T_{b5en}^{4.313}T_{b4me}^{8.973}S_{MSAVI}^{-4.811}$	0.70
	Lidar	$Y = 8.284H_{80th}^{1.057}$	0.74	$Y = 1.424H_{qm}^{1.772}$	0.87	$Y = 2.285H_{70th}^{1.536}$	0.84
	Combination	$Y = 8.284H_{80th}^{1.057}$	0.74	$Y = 1.861H_{qm}^{1.432}T_{b2ho}^{-2.398}$	0.91	$Y = 2.659H_{70th}^{1.515}T_{b2va}^{0.151}$	0.86
Random forest	RapidEye	S_{b3}, T_{b4va}, T_{b5di}, T_{b1me}	0.93	S_{b1},T_{b2ho}, T_{b4en}, T_{b3di}	0.82	S_{b1}, T_{b2di}, T_{b4se}, T_{b5en}	0.82
	Lidar	H_{std}, H_{60th}	0.91	H_{me}, H_{40th}	0.98	H_{qm}, H_{70th}	0.97
	Combination	H_{me}, H_{sk}, H_{60th}, T_{b4di}	0.92	H_{70th}, H_{me}, H_{60th}, T_{b2ho}	0.98	H_{std}, H_{qm}, H_{70th}, T_{b2di}	0.96
Support	RapidEye	S_{b3}, T_{b4va}, T_{b5di}, T_{b1me}	0.72	S_{b1}, T_{b2ho}, T_{b4en}, T_{b3di}	0.47	S_{b1}, T_{b2di}, T_{b4se}, T_{b5en}	0.80
vector	Lidar	H_{std}, H_{60th}	0.93	H_{me}, H_{40th}	0.90	H_{qm}, H_{70th}	0.89
regression	Combination	H_{me}, H_{sk}, H_{60th}, T_{b4di}	0.82	H_{70th}, H_{me}, H_{60th}, T_{b2ho}	0.97	H_{std}, H_{qm}, H_{70th}, T_{b2di}	0.93

Note: S_{bi} represents spectral band *i*, T_{bixx} represents a texture image which was developed using the texture measure xx (xx can be such texture measures as me – mean, va – variance, di – dissimilarity, ho – homogeneity, en – entropy, se – second moment) on spectral band *i*, for example, T_{b2ho} represents the texture images developed using the homogeneity based on spectral band 2. H_{xx} represents the lidar-derived metrics, for example, H_{80th} represents the 80th percentile height, H_{me}, H_{std}, H_{sk}, and H_{qm} represent mean, standard deviation, skewness, and quadratic mean height. S_{EVI} and S_{MSAVI} represent Enhanced Vegetation Index and Modified Soil Adjusted Vegetation Index from RapidEye image.

for results from cross-validation). Overall, the machine learning-based algorithms have higher R^2 values than linear regression and multiplicative power model, but different algorithms have their own capabilities in modeling AGB using different data sources. For example, for linear regression approach, the lidar-based variables have higher R^2 values than RapidEye-based variables and the combination of lidar and RapidEye produces higher R^2 values than lidar or RapidEye data alone. For the multiplicative power model, the conclusion is similar to the linear regression, except the combination of lidar and RapidEye for agroforestry. For RF, the R^2 values can be very high for different data sources with either stratification or non-stratification conditions. However, the combination of lidar and RapidEye data cannot improve the modeling performance. In contrast, for SVR, combination of lidar and RapidEye data improved R^2 values for non-stratification or for secondary forests, but not for agroforestry.

The predicted AGB spatial distributions using these developed models are illustrated in Figure 4. For RapidEye data under non-stratification (Figure 4(a–d)), the linear-based model and RF produced much more pixels with high AGB values than the multiplicative-based model and SVR. The multiplicative-based model produced many pixels with very high AGB amounts, but for lidar or the combination of lidar and RapidEye data, the spatial patterns of AGB distribution looked similar using different modeling algorithms. For the stratification of vegetation types (Figure 4(A–D)), the RapidEye-based AGB estimation results using different modeling algorithms have more pixels having high AGB values than lidar or combination of lidar and RapidEye data. The different spatial patterns of AGB distributions using different modeling approaches and data sources can be better explained by the statistical results, as summarized in Table 4.

Table 4 indicates that the RapidEye-based models provided much higher mean values than lidar and data combination regardless of which modeling approaches were used. Compared with the mean values based on the non-stratification scenarios, the stratification increased mean values for all results except using linear model based on lidar and using multiplicative model and RF based on RapidEye data; this is especially obvious for the SVR approach based on lidar and combination of both datasets. Overall, lidar-based results have relatively stable mean and standard deviation amounts among different modeling approaches. Table 4 indicates that different modeling algorithms and data sources produce highly different estimation results, implying that an improved estimation result for a specific study area may be developed using the combination of the results from different algorithms and/or data sources.

This situation can be better explained using the scatterplots between residues and AGB reference data (Figure 5). These scatterplots show that underestimation and overestimation are obvious, especially for the AGB estimation using RapidEye data. If we calculate the mean values of overestimation and underestimation for the test samples (see Table 5), we can clearly find that RapidEye data have much higher overestimation and underestimation values than lidar data; and combination of

Table 4. Summary of statistical values for the predicted AGB images.

		Non-stratification				Stratification			
Model	Data sources	Mean	Std	Min	Max	Mean	Std	Min	Max
Linear models	RapidEye	132.35	69.03	1.45	423.41	151.95	120.75	1.46	612.40
	Lidar	97.39	70.21	0.84	280.02	95.98	72.19	2.18	654.41
	Combination	94.90	73.23	1.32	307.08	104.57	84.69	0.16	594.60
Multiplicative models	RapidEye	146.08	174.34	3.08	783.96	144.87	142.20	2.86	729.20
	Lidar	79.14	72.44	1.42	361.88	80.61	65.99	1.33	339.50
	Combination	82.74	76.54	1.43	365.53	93.18	85.08	2.16	550.20
Random forest	RapidEye	152.02	64.65	30.64	412.43	136.02	83.99	26.60	356.94
	Lidar	79.72	74.26	23.14	421.53	86.21	78.61	23.18	422.26
	Combination	93.40	58.40	29.11	389.14	102.12	71.57	25.94	413.39
Support vector regression	RapidEye	97.69	44.24	9.79	146.80	118.26	80.92	0.35	464.62
	Lidar	68.14	75.81	0.39	860.00	101.88	73.94	10.34	527.54
	Combination	60.06	48.74	20.71	478.32	103.91	71.52	0.52	504.26

Note: Std, Min, and Max represent standard deviation, minimum, and maximum.

Figure 4. The predicted AGB distributions with 30 m spatial resolution using different models and data sources under stratification and non-stratification conditions (Note: (a), (b), (c), and (d) represent linear regression, multiplicative model, random forest, and support vector regression based on non-stratification; (A), (B), (C), and (D) represent linear regression, multiplicative model, random forest, and support vector regression based on stratification of vegetation types; 1, 2, and 3 represent RapidEye, lidar, and combination of both data; others represent other vegetation, cloud, and shadow).

lidar and RapidEye data cannot improve the overestimation and underestimation problems under non-stratification condition. Considering the modeling algorithms, the nonparametric algorithms (i.e. RF and SVR here) have much smaller overestimation problem than linear regression and

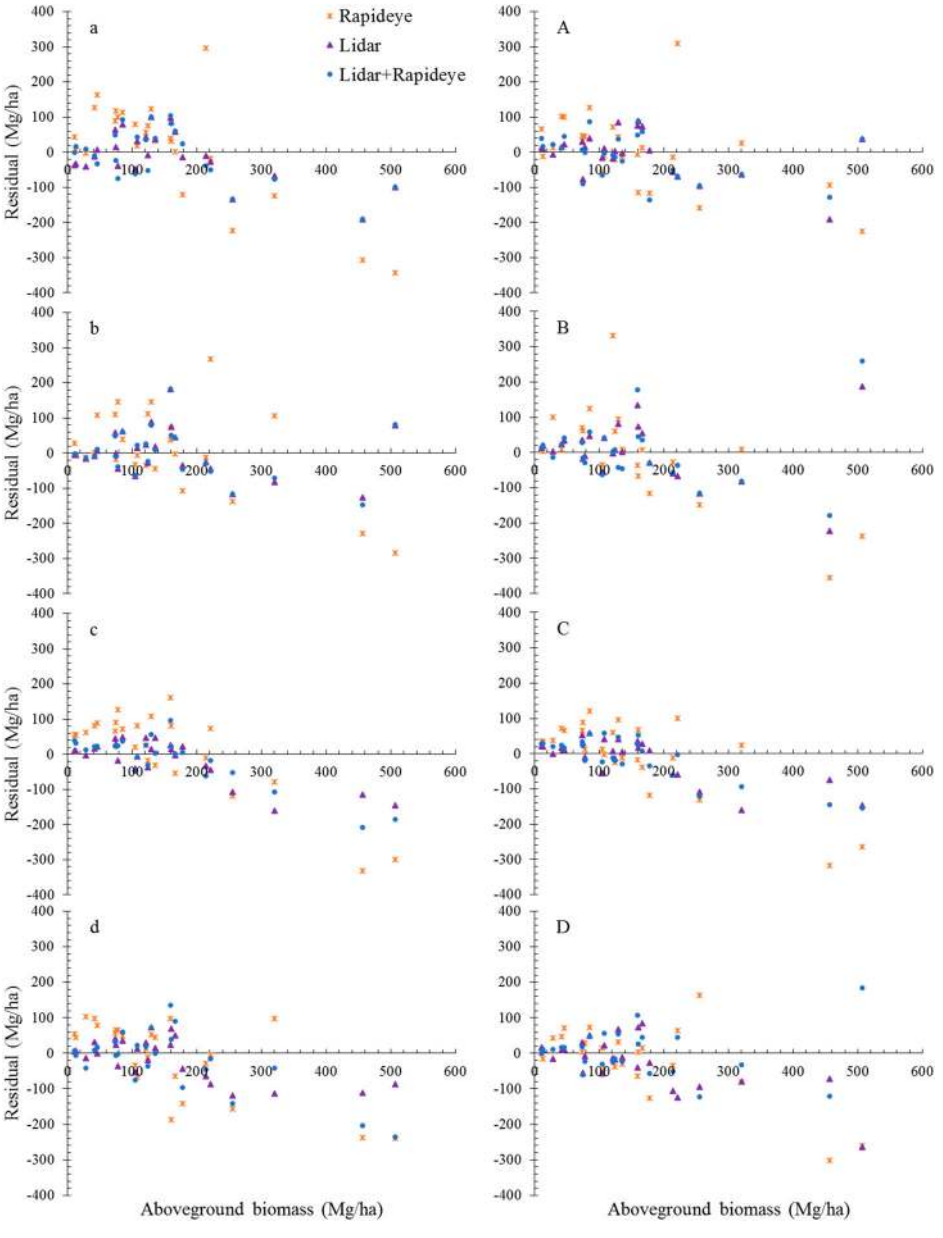

Figure 5. Residual analysis of different modeling results (Note: (a), (b), (c), and (d) represent linear regression, multiplicative model, random forest, and support vector regression based on non-stratification; (A), (B), (C), and (D) represent linear regression, multiplicative model, random forest, and support vector regression based on stratification of vegetation types).

multiplicative model, but underestimation problem is not reduced, and the multiplicative model has the best performance in reducing underestimation problem. Overall, for the non-stratification condition, the RF based on lidar or combination of lidar and RapidEye has the best performance in reducing overestimation problem, and multiplicative model based on lidar or combination of lidar and RapidEye has the best performance in reducing the underestimation problem. Stratification is mainly valuable for the RapidEye data in reducing overestimation or underestimation when linear regression, RF or SVR are used; and for the combination of lidar and RapidEye in reducing underestimation problem when linear regression was used. For lidar data, stratification is not needed,

Table 5. A comparison of mean values of overestimation or underestimation from different datasets and algorithms under stratification and non-stratification.

		Non-stratification		Stratification	
Model	Data	Overestimation	Underestimation	Overestimation	Underestimation
Linear regression	RapidEye	84.16	162.68	76.70	69.98
	Lidar	56.56	54.09	35.01	59.17
	Combination	54.61	64.93	42.17	57.38
Multiplicative model	RapidEye	98.10	71.14	278.17	127.67
	Lidar	54.99	45.75	50.74	65.91
	Combination	56.10	44.35	57.91	59.38
Random forest	RapidEye	74.57	116.98	59.93	93.90
	Lidar	27.37	59.33	26.67	68.37
	Combination	28.10	79.71	32.60	58.98
Support vector regression	RapidEye	66.03	106.31	42.21	92.28
	Lidar	32.40	62.39	37.46	65.43
	Combination	45.64	67.65	49.18	46.56

because it cannot considerably reduce overestimation or underestimation problem. Overall, RF based on lidar data under non-stratification provides the best performance with the lowest overestimation problem and the multiplicative model based on lidar and combination of lidar and RapidEye under non-stratification has the best performance with the lowest underestimation value.

4.2. Comparative analysis of AGB prediction performance

The accuracy assessment results based on different scenarios (Table 6) indicate that overall, RapidEye-based prediction results have the highest RMSE and RMSEr values compared with lidar or the combination of lidar and RapidEye data with stratification or non-stratification conditions. The lidar-based prediction approaches have relatively stable performance. The combination of lidar and RapidEye data cannot improve AGB estimation using different approaches, except SVR under stratification condition. Specifically, for RapidEye data, stratification of vegetation types is helpful in improving AGB estimation, except using multiplicative model; and the SVR algorithm based on stratification provides the best performance with the lowest RMSEr of 64.2%. For lidar data, stratification is only valuable for linear regression model; the SVR algorithm based on non-stratification provides the best estimation with the lowest RMSEr of 38.2%. For the combination of lidar and RapidEye data, the RF algorithm based on stratification provides the best performance with RMSEr of 39.6%. Table 5 also indicates that considering individual vegetation types – agroforestry and secondary forest, agroforestry has lower RMSE but higher RMSEr than secondary

Table 6. Summary of accuracy assessment results.

		Non-stratification		Stratification		Agroforestry		Secondary forests	
Model	Data source	RMSE	RMSEr	RMSE	RMSEr	RMSE	RMSEr	RMSE	RMSEr
Linear models	RapidEye	141.96	93.12	103.19	67.68	76.45	79.92	122.80	59.93
	Lidar	70.25	46.08	61.79	40.53	37.26	38.95	77.85	38.00
	Combination	73.62	48.29	62.47	40.97	57.06	59.65	67.06	32.73
Multiplicative models	RapidEye	100.37	65.83	463.14	303.78	634.40	663.19	202.38	98.77
	Lidar	66.78	43.80	81.73	53.61	57.70	60.32	95.17	46.45
	Combination	66.56	43.66	84.42	55.37	68.14	71.23	97.04	47.36
Random forest	RapidEye	116.42	76.36	104.49	68.53	61.87	64.68	132.14	64.49
	Lidar	60.12	39.43	59.61	39.10	40.28	42.11	73.05	35.65
	Combination	69.12	45.34	60.44	39.64	44.50	46.52	72.09	35.18
Support vector regression	RapidEye	104.61	68.61	97.92	64.23	70.97	74.20	117.43	57.32
	Lidar	58.26	38.21	76.73	50.33	36.69	38.35	104.50	51.00
	Combination	84.46	55.40	64.01	41.99	54.50	56.98	71.68	34.98

Note: The RMSE unit is Mg/ha, and RMSEr is non-unit variable.

forest for all data sources and algorithms, except for lidar data with SVR. The high RMSEr in the agroforestry is due to its low mean value, as shown in Table 1. In summary, the SVR approach based on lidar data with non-stratification is recommended for AGB estimation in this research.

The overall RMSE and RMSEr did not provide much information about the overestimation or underestimation, and about which AGB ranges have high estimation uncertainty. On the other hand, the RMSE and RMSEr analysis results at three AGB ranges (see Table 7) – less than 50 Mg/ha, between 50 and 150 Mg/ha, and greater than 150 Mg/ha – provide much more useful information of the AGB estimation errors. Considering the best AGB estimation at three AGB ranges, the results based on non-stratification have better accuracy than those from stratification. For example, when AGB is less than 50 Mg/ha, the multiplicative models based on lidar or combination of lidar and RapidEye data provided much better performance than any other approaches; when AGB is within 50–150 Mg/ha, RF based on combination of lidar and RapidEye provides the best estimation; and when AGB is greater than 150 Mg/ha, SVR based on lidar data has the best estimation. This situation implies that different modeling approaches have various performances depending on the size of AGB, providing new challenges in collection of sample plots, and selection of suitable variables and algorithms for AGB modeling.

5. Discussion

Since optical, radar and lidar with different spatial resolutions are available, selection of suitable data sources for AGB modeling for a specific study area is one of the important steps in the AGB estimation procedure to produce a good estimation (Lu et al. 2016). Because Landsat imagery has a long-term history of data availability and suitable spatial and spectral resolutions, previous research has extensively explored its application for AGB estimation and has proven that reasonable good results can be obtained (e.g. Lu 2005; 2006) when AGB has not reached the saturation value such as 150 Mg/ha (Zhao et al. 2016a). However, use of high spatial resolution images such as RapidEye,

Table 7. Summary of RMSE and RMSEr results at different AGB ranges.

			AGB Range (Mg/ha)					
			<50		50–150		>150	
	Model	Datasets	RMSE	RMSEr	RMSE	RMSEr	RMSE	RMSEr
Non-stratification	Linear models	RapidEye	94.41	403.91	87.63	84.66	195.57	74.17
		Lidar	28.87	123.49	53.90	52.08	94.95	36.01
		Combination	17.31	74.06	62.08	59.98	97.70	37.05
	Multiplicative models	RapidEye	50.73	217.03	85.73	82.83	159.08	60.33
		Lidar	8.52	36.46	48.42	46.78	93.64	35.51
		Combination	8.45	36.15	44.31	42.81	95.28	36.13
	Random forest	RapidEye	70.56	301.84	73.94	71.44	161.05	61.07
		Lidar	13.67	58.49	36.80	35.56	87.12	33.04
		Combination	27.59	118.03	32.29	31.20	102.57	38.90
	Support vector regression	RapidEye	79.10	338.39	50.24	48.54	147.33	55.87
		Lidar	15.99	68.43	38.30	37.01	83.01	31.48
		Integration	21.23	90.81	41.89	40.48	125.91	47.75
Stratification	Linear models	RapidEye	71.17	304.47	57.23	55.30	144.25	54.70
		Lidar	14.25	60.95	40.63	39.25	88.28	33.48
		Combination	30.60	130.89	47.18	45.58	84.04	31.87
	Multiplicative models	RapidEye	252.42	1079.88	679.53	656.54	206.44	78.29
		Lidar	22.04	94.30	38.95	37.63	118.36	44.88
		Combination	24.85	106.30	38.67	37.37	118.76	71.58
	Random forest	RapidEye	51.64	220.92	64.04	61.87	147.85	56.07
		Lidar	16.60	71.02	37.45	36.18	85.70	32.50
		Combination	24.15	103.30	35.53	34.33	87.05	33.01
	Support vector regression	RapidEye	43.23	184.96	36.23	35.01	147.40	55.90
		Lidar	12.43	53.19	39.38	38.05	114.42	43.39
		Combination	12.90	55.17	38.96	37.65	92.96	35.25

Note: The RMSE unit is Mg/ha, and RMSEr is non-unit variable.

IKONOS, and QuickBird for AGB estimation has not been fully examined (Thenkabail et al. 2004; Leboeuf et al. 2007). The possible reasons may be the cost for image purchase, the lack of shortwave infrared spectral bands, and the relatively small swath size.

This research explored the use of RapidEye image for modeling AGB distribution in the moist tropical region and indicated that RapidEye, overall, cannot produce satisfactory AGB estimation. However, the high spatial resolution indeed provides rich spatial information that can be used for AGB estimation. Proper selection of textures from RapidEye image is valuable in reducing the spectral heterogeneity effects. In particular, the SVR algorithm based on the combination of RapidEye spectral and textural images provide the best estimation with RMSEr of only 35.0% when AGB was within 50 and 150 Mg/ha. This research also indicated that RapidEye image is not suitable for AGB estimation when AGB is too small or too high, similar conclusion to the previous research using Landsat imagery (Zhao et al. 2016a). The possible reason is that when AGB is small, forest canopy is not dense enough to cover the ground, thus the soils and/or grass have important effects on the spectral signature; and when AGB is too high, the spectral signature cannot reflect the AGB change due to the data saturation problem (Lu et al. 2016; Zhao et al. 2016a). In this situation, lidar data can overcome the problems in RapidEye data, as shown in this study. In previous research, lidar has been proven the most accurate data source for AGB estimation (Chen 2013), and this research also confirmed that lidar provided reliable AGB predictions in the moist tropical region using different modeling approaches such as linear regression or nonparametric algorithms such as SVR.

Because lidar and high spatial resolution optical sensor data have different capability in representing surface features, researchers have explored the approaches to combine both types of data for improving AGB estimation. Previous studies have obtained different conclusions, depending on the selection of different data sources, algorithms, and characteristics of the study areas (Clark et al. 2011; Latifi, Fassnacht, and Koch 2012; Fassnacht et al. 2014; Vaglio Laurin et al. 2014; Xu et al. 2015). This study demonstrated that combination of lidar and RapidEye cannot improve AGB estimation, either using traditional linear regression or using nonparametric algorithms such as RF and SVR. However, this research indeed shows that when AGB is within 50–150 Mg/ha, combination of lidar and RapidEye improved AGB estimation, the RMSEr amount decreased from 35.6% for lidar data to 31.2% for combination of lidar and RapidEye data using RF. When AGB is high, such as higher than 150 Mg/ha, combination of lidar and RapidEye data cannot improve AGB estimation due to the data saturation problem in RapidEye data. This research implies that if we can reduce the impacts of bare soils and grass on forest spectral signatures when forest cover is not dense enough (i.e. when AGB is small), or we can reduce the data saturation problem when AGB is very high, for example, use of ALOS PALSAR data (Zhao et al. 2016b), combination of lidar and RapidEye image may be valuable.

Stratification has been regarded as an effective approach to improve AGB estimation, especially for Landsat imagery (Zhao et al. 2016a). This research also confirms that when RapidEye is used for AGB modeling, stratification of vegetation types is an effective way to improve AGB estimation based on either linear regression or machine learning algorithms. This research also indicates that when lidar is used for AGB modeling, stratification is not necessary in the moist tropical region. This implies that lidar data are less influenced by data saturation problem when AGB is high and by bare soils and grass when AGB is small. Another option to improve AGB estimation is to conduct the stratification based on AGB ranges, as shown in Table 6 in this research. This research implies that if AGB modeling is based on different AGB ranges, the modeling results can be improved, but this requires much higher number of sample plots, which is often difficult, especially in the moist tropical regions.

As Table 7 indicates that when AGB is less than 50 Mg/ha and within 50–150 Mg/ha, multiplicative model and RF based on combination of lidar and RapidEye provided the best estimation, and when AGB is greater than 150 Mg/ha, the SVR based on lidar data without stratification provides the best performance. In other words, we can develop a new product based on the combination of these three models, as illustrated in Figure 6. This new product further improved AGB estimation

with RSME and RMSEr of 56.5 Mg/ha and 37.0%. This implies that an optimal procedure for AGB estimation for a specific study exists, depending on the careful selection of data sources, modeling algorithms, forest types, and AGB ranges.

Linear regression is commonly used for AGB estimation modeling in previous studies (Lu et al. 2016; Zhao et al. 2016a, 2016b). However, the requirement in data distribution and constraints in selection of variables make the linear regression approach perform poorly. The flexible selection of different types of variables and free data distribution requirement in machine learning approaches provide a better choice for AGB modeling, as shown in this research, that is, the RF and SVR provide better AGB estimation than linear regression. This research also indicates that the AGB estimation depends on different factors such as data sources and modeling algorithms. This requires us to carefully identify a modeling approach for different data sources.

Previous research often focused on the exploration of individual factors such as identification of the best variables or best algorithms for AGB estimation. In reality, AGB estimation is a complex procedure that requires careful design of each step, such as the calculation of AGB from sample plots, extraction and selection of remote sensing variables, selection of AGB modeling approaches and evaluation of prediction results (Lu et al. 2016). It is critical to carefully deal with each step to minimize errors or uncertainties. For example, the calculation of AGB from sample plots using allometric models is one of the critical steps resulting in uncertainty of AGB estimation (Chen, Laurin, and Valentini 2015). The complex species composition in the moist tropical region and the difficulty in developing or selecting an allometric equation for specific tree species result in high uncertainty in the AGB reference data themselves.

One of the key issues in AGB modeling and estimation is how to determine the sample size for model calibration. Given the relatively small extent of our study area (i.e. the population), the sampling fraction is actually higher than most of the previous studies. Nevertheless, the number of sample plots seems relatively small. Although we have selected field plots to span different forest ages and composition, are our sample sizes large enough to model AGB in this area? A comparison between the final AGB map (Figure 6(b)) and field measurement (Table 1) indicates that the field plots approximately cover the AGB range of whole regions, for both agroforestry and secondary forest. However, if resources exist to improve our sampling in the future, we would consider increasing our sample size for agroforestry. As shown in Table 6, secondary forests have smaller RMSEr than agroforestry. We expect that the relationships between remotely sensed data and AGB are species-

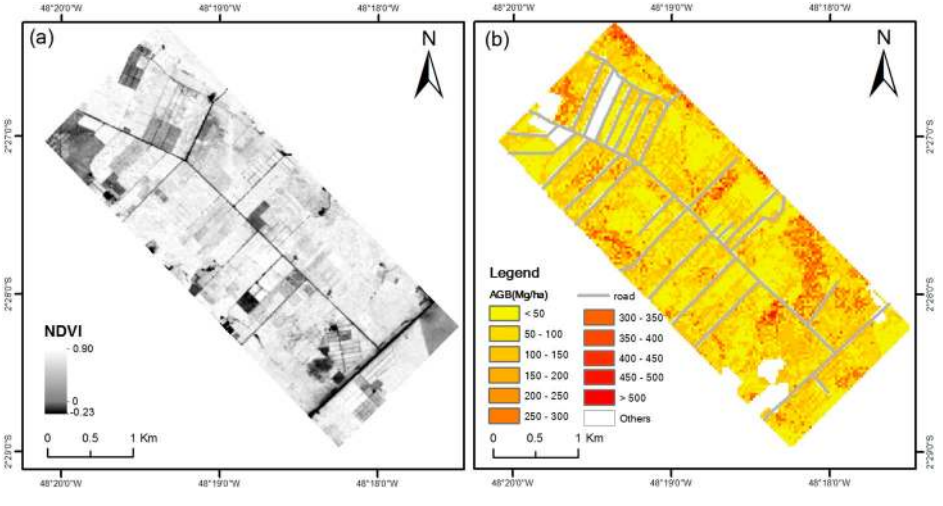

Figure 6. A comparison of NDVI image (a) from RapidEye and the distribution of the predicted AGB with 30 m spatial resolution (b) using the combination of three best AGB estimation models.

dependent because different tree species, for example, have different DBH-tree height relationships and wood densities (Chen 2015; Chen et al. 2016). However, when the models are developed at the plot level, we are essentially modeling the average relationship for tree species within plots. Therefore, the variations in these relationships are substantially reduced for secondary forest, but not necessarily for agroforestry because only one or a few tree species exist in an agroforestry field. To capture such variations, the further improvements in our efforts are to improve sample sizes for agroforestry and/or use statistical models (e.g. mixed-effects models) that are amenable to relatively small sample sizes (Chen et al. 2012, 2016).

6. Conclusions

This research conducted a comparative analysis of different data sources and modeling algorithms for AGB estimation under different stratification conditions in the secondary forests and agroforestry systems of the Brazilian Amazon. The following conclusions were obtained:

(1) RapidEye data alone are not suitable for AGB estimation, but the AGB modeling based on stratification of vegetation types is valuable to improve AGB estimation. In particular, if AGB falls within 50–150 Mg/ha, SVR based on stratification provides the best performance.
(2) Lidar data provide stable and better performance than RapidEye data no matter which modeling approaches are used. Overall, the SVR algorithm with non-stratification provides the best performance.
(3) For the combination of lidar and RapidEye data, RF under stratification of vegetation types provides the best performance but the performance is not as good as the SVR based on lidar data alone under non-stratification condition.
(4) Stratification is not needed for lidar data, but it indeed improves AGB estimation if RapidEye data alone are used for AGB modeling.
(5) AGB ranges affect the selection of the best AGB modeling. A combination of different estimation results from the best models can further improve AGB estimation.
(6) An optimal procedure for AGB modeling should take the selection of data sources, modeling algorithms, forest types, and AGB ranges into account.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was financially supported by the National Natural Science Foundation of China (No# 41571411) and the Zhejiang A&F University's Research and Development Fund for the talent startup project (No# 2013FR052). Keller, dos-Santos, Bolfe, and Batistella acknowledge the support from the Brazilian National Council for Scientific and Technological Development – CNPq (No#457927/2013-5). Data were acquired by the Sustainable Landscapes Brazil project supported by the Brazilian Agricultural Research Corporation (EMBRAPA), the US Forest Service, the USAID, and the US Department of State.

ORCID


Dengsheng Lu http://orcid.org/0000-0003-4767-5710

References

Avitabile, V., A. Baccini, M. A. Friedl, and C. Schmullius. 2012. "Capabilities and Limitations of Landsat and Land Cover Data for Aboveground Woody Biomass Estimation of Uganda." *Remote Sensing of Environment* 117: 366–380.

Baccini, A., N. Laporte, S. J. Goetz, M. Sun, and H. Dong. 2008. "A First Map of Tropical Africa's Above-ground Biomass Derived from Satellite Imagery." *Environmental Research Letters* 3 (4), 045011. doi:10.1088/1748-9326/3/4/045011.

Barbosa, J. M., I. Melendez-Pastor, J. Navarro-Pedreño, and M. D. Bitencourt. 2014. "Remotely Sensed Biomass Over Steep Slopes: An Evaluation among Successional Stands of the Atlantic Forest, Brazil." *ISPRS Journal of Photogrammetry and Remote Sensing* 88: 91–100.

Batistella, M., E. L. Bolfe, and E. F. Moran. 2013. "Agroforestry in Tomé-Açu: An Alternative to Pasture in the Amazon." In *Human-environment Interactions*, edited by E. S. Brondízio and E. F. Moran, 321–342. Dordrecht: Springer-Verlag .

Bolfe, E. L., and M. Batistella. 2011. "Análise florística e estrutural de sistemas silviagrícolas em Tomé-Açu, Pará." *Pesquisa Agropecuária Brasileira* 46: 1139–1147.

Bolfe, É. L., M. Batistella, and M. C. Ferreira. 2012. "Correlation of Spectral Variables and Aboveground Carbon Stock of Agroforestry Systems." *Pesqui. Agropecu. Bras* 47: 1261–1269.

Breidenbach, J., E. Næsset, V. Lien, T. Gobakken, and S. Solberg. 2010. "Prediction of Species Specific Forest Inventory Attributes Using a Nonparametric Semi-Individual Tree Crown Approach Based on Fused Airborne Laser Scanning and Multispectral Data." *Remote Sensing of Environment* 114: 911–924.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Chave, J., C. Andalo, S. Brown, M. A. Cairns, J. Q. Chambers, D. Eamus, H. Fölster, et al. 2005. "Tree Allometry and Improved Estimation of Carbon Stocks and Balance in Tropical Forests." *Oecologia* 145 (1): 87–99.

Chave, J., D. Coomes, S. Jansen, S. L. Lewis, N. G. Swenson, and A. E. Zanne. 2009. "Towards a Worldwide Wood Economics Spectrum." *Ecology Letters* 12 (4): 351–366.

Chen, Q. 2013. "Lidar Remote Sensing of Vegetation Biomass." In *Remote Sensing of Natural Resources*, edited by Q. Weng and G. Wang, 399–420. Taylor & Francis: CRC Press.

Chen, Q. 2015. "Modeling Aboveground Tree Woody Biomass Using National-scale Allometric Methods and Airborne Lidar." *ISPRS Journal of Photogrammetry and Remote Sensing* 106: 95–106.

Chen, Q., G. V. Laurin, and R. Valentini. 2015. "Uncertainty of Remotely Sensed Aboveground Biomass Over an African Tropical Forest: Propagating Errors from Trees to Plots to Pixels." *Remote Sensing of Environment* 160: 134–143.

Chen, Q., D. Lu, M. Keller, M. N. dos-Santos, E. L. Bolfe, Y. Feng, and C. Wang. 2016. "Modeling and Mapping Agroforestry Aboveground Biomass in the Brazilian Amazon Using Airborne Lidar Data." *Remote Sensing* 8 (1): 21. doi:10.3390/rs8010021.

Chen, Q., G. Vaglio Laurin, J. J. Battles, and D. Saah. 2012. "Integration of Airborne Lidar and Vegetation Types Derived from Aerial Photography for Mapping Aboveground Live Biomass." *Remote Sensing of Environment* 121: 108–117.

Cherkassky, V., and Y. Ma. 2004. "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression." *Neural Networks* 17 (1): 113–126.

Clark, M. L., D. A. Roberts, J. J. Ewel, and D. B. Clark. 2011. "Estimation of Tropical Rain Forest Aboveground Biomass with Small-footprint Lidar and Hyperspectral Sensors." *Remote Sensing of Environment* 115: 2931–2942.

Cortes, C., and V. N. Vapnik. 1995. "Support Vector Networks." *Machine Learning* 20: 273–297.

Dube, T., and O. Mutanga. 2015. "Investigating the Robustness of the New Landsat-8 Operational Land Imager Derived Texture Metrics in Estimating Plantation Forest Aboveground Biomass in Resource Constrained Areas." *ISPRS Journal of Photogrammetry and Remote Sensing* 108: 12–32.

Dube, T., O. Mutanga, A. Elhadi, and R. Ismail. 2014. "Intra-and-inter Species Biomass Prediction in a Plantation Forest: Testing the Utility of High Spatial Resolution Spaceborne Multispectral RapidEye Sensor and Advanced Machine Learning Algorithms." *Sensors* 14 (8): 15348–15370.

Eskelson, B. N. I., T. M. Barrett, and H. Temesgen. 2009. "Imputing Mean Annual Change to Estimate Current Forest Attributes." *Silva Fennica* 43: 649–658.

Fassnacht, F. E., F. Hartig, H. Latifi, C. Berger, J. Hernández, P. Corvalán, and B. Koch. 2014. "Importance of Sample Size, Data Type and Prediction Method for Remote Sensing-based Estimations of Aboveground Forest Biomass." *Remote Sensing of Environment* 154: 102–114.

Foody, G. M., D. S. Boyd, and M. E. J. Cutler. 2003. "Predictive Relations of Tropical Forest Biomass from Landsat TM Data and Their Transferability Between Regions." *Remote Sensing of Environment* 85: 463–474.

Gleason, C. J., and J. Im. 2011. "A Review of Remote Sensing of Forest Biomass and Biofuel: Options for Small-area Applications." *GIScience & Remote Sensing* 48: 141–170.

Gleason, C. J., and J. Im. 2012. "Forest Biomass Estimation from Airborne LiDAR Data Using Machine Learning Approaches." *Remote Sensing of Environment* 125: 80–91.

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. 2nd ed. New York: Springer Verlag.

Hudak, A. T., E. K. Strand, L. A. Vierling, J. C. Byrne, J. U. H. Eitel, S. Martinuzzi, and M. J. Falkowski. 2012. "Quantifying Aboveground Forest Carbon Pools and Fluxes from Repeat LiDAR Surveys." *Remote Sensing of Environment* 123: 25–40.

Hunter, M. O., M. Keller, D. Victoria, and D. C. Morton. 2013. "Tree Height and Tropical Forest Biomass Estimation." *Biogeosciences* 10: 8385–8399. doi:10.5194/bg-10-8385-2013.

Hyde, P., R. Dubayah, W. Walker, J. B. Blair, M. Hofton, and C. Hunsaker. 2006. "Mapping Forest Structure for Wildlife Habitat Analysis Using Multi-sensor (LiDAR, SAR/InSAR, ETM+, QuickBird) Synergy." *Remote Sensing of Environment* 102: 63–73.

Ismail, R., O. Mutanga, and L. Kumar. 2010. "Modeling the Potential Distribution of Pine Forests Susceptible to Sirex Noctilio Infestations in Mpumalanga, South Africa." *Transactions in GIS* 14 (5): 709–726.

Kelsey, K. C., and J. C. Neff. 2014. "Estimates of Aboveground Biomass from Texture Analysis of Landsat Imagery." *Remote Sensing* 6: 6407–6422.

Koch, B. 2010. "Status and Future of Laser Scanning, Synthetic Aperture Radar and Hyperspectral Remote Sensing Data for Forest Biomass Assessment." *ISPRS Journal of Photogrammetry and Remote Sensing* 65: 581–590.

Kumar, L., P. Sinha, S. Taylor, and A. F. Alqurashi. 2015. "Review of the use of Remote Sensing for Biomass Estimation to Support Renewable Energy Generation." *Journal of Applied Remote Sensing* 9: 097696. doi:10.1117/1.JRS.9.097696.

Latifi, H., F. Fassnacht, and B. Koch. 2012. "Forest Structure Modeling with Combined Airborne Hyperspectral and Lidar Data." *Remote Sensing of Environment* 121: 10–25.

Leboeuf, A., A. Beaudoin, R. A. Fournier, L. Guindon, J. E. Luther, and M. C. Lambert. 2007. "A Shadow Fraction Method for Mapping Biomass of Northern Boreal Black Spruce Forests Using QuickBird Imagery." *Remote Sensing of Environment* 110: 488–500.

Li, M., J. Im, L. J. Quackenbush, and T. Liu. 2014. "Forest Biomass and Carbon Stock Quantification Using Airborne LiDAR Data: A Case Study Over Huntington Wildlife Forest in the Adirondack Park." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (7): 3143–3156.

Li, W., Z. Niu, X. Liang, Z. Li, N. Huang, S. Gao, C. Wang, and S. Muhammad. 2015. "Geostatistical Modeling Using LiDAR-Derived Prior Knowledge with SPOT-6 Data to Estimate Temperate Forest Canopy Cover and Above-Ground Biomass via Stratified Random Sampling." *International Journal of Applied Earth Observation and Geoinformation* 41: 88–98.

Lim, K., P. Treitz, K. Baldwin, I. Morrison, and J. Green. 2003. "Lidar Remote Sensing of Biophysical Properties of Tolerant Northern Hardwood Forests." *Canadian Journal of Remote Sensing* 29 (5): 658–678.

Lu, D. 2005. "Aboveground Biomass Estimation Using Landsat TM Data in the Brazilian Amazon." *International Journal of Remote Sensing* 26: 2509–2525.

Lu, D. 2006. "The Potential and Challenge of Remote Sensing-based Biomass Estimation." *International Journal of Remote Sensing* 27: 1297–1328.

Lu, D., and M. Batistella. 2005. "Exploring TM Image Texture and its Relationships with Biomass Estimation in Rondônia, Brazilian Amazon." *Acta Amazonica* 35: 249–257.

Lu, D., M. Batistella, and E. Moran. 2005. "Satellite Estimation of Aboveground Biomass and Impacts of Forest Stand Structure." *Photogrammetric Engineering and Remote Sensing* 71: 967–974.

Lu, D., Q. Chen, G. Wang, L. Liu, G. Li, and E. Moran. 2016. "A Survey of Remote Sensing-based Aboveground Biomass Estimation Methods in Forest Ecosystems." *International Journal of Digital Earth* 9 (1): 63–105.

Lu, D., Q. Chen, G. Wang, E. Moran, M. Batistella, M. Zhang, G. V. Laurin, and D. Saah. 2012. "Aboveground Forest Biomass Estimation with Landsat and LiDAR Data and Uncertainty Analysis of the Estimates." *International Journal of Forestry Research* 16. doi:10.1155/2012/436537.

Lu, D., and Q. Weng. 2007. "A Survey of Image Classification Methods and Techniques for Improving Classification Performance." *International Journal of Remote Sensing* 28 (5): 823–870.

Marabel, M., and F. Alvarez-Taboada. 2013. "Spectroscopic Determination of Aboveground Biomass in Grasslands Using Spectral Transformations, Support Vector Machine and Partial Least Squares Regression." *Sensors* 13 (8): 10027–10051.

Mitchard, E. T. A., S. S. Saatchi, S. L. Lewis, T. R. Feldpausch, I. H. Woodhouse, B. Sonké, C. Rowland, and P. Meir. 2011. "Measuring Biomass Changes due to Woody Encroachment and Deforestation/Degradation in a Forest-savanna Boundary Region of Central Africa Using Multi-temporal L-band Radar Backscatter." *Remote Sensing of Environment* 115 (11): 2861–2873.

Mountrakis, G., J. Im, and C. Ogole. 2011. "Support Vector Machines in Remote Sensing: A Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (3): 247–259.

Nascimento, H. E. M., and F. W. Laurance. 2002. "Total Aboveground Biomass in Central Amazonian Rainforests: a Landscape-scale Study." *Forest Ecology and Management* 168: 311–321.

Pflugmacher, D., W. B. Cohen, R. E. Kennedy, and Z. Yang. 2014. "Using Landsat-derived Disturbance and Recovery History and Lidar to Map Forest Biomass Dynamics." *Remote Sensing of Environment* 151: 124–137.

Popescu, S. C., H. W. Randolph, and A. S. John. 2004. "Fusion of Small-footprint Lidar and Multispectral Data to Estimate Plot-level Volume and Biomass in Deciduous and Pine Forests in Virginia, USA." *Forest Science* 50 (4): 551–565.

Rodrigues, T. E., P. L. dos Santos, P. A. M. Rolim, E. Santos, R. S. Rego, J. M. L. da Silva, M. A. Valente, and J. R. N. F. Gama. 2011. *Caracterização e Classificação dos Solos do Município de Tomé-Açu, Pará*. Belém: Embrapa Amazônia Oriental. pp. 49.

Schnitzer, S. A., S. J. DeWalt, and J. Chave. 2006. "Censusing and Measuring Lianas: A Quantitative Comparison of the Common Methods." *Biotropica* 38 (5): 581–591.

Sileshi, G. W. 2014. "A Critical Review of Forest Biomass Estimation Models, Common Mistakes and Corrective Measures." *Forest Ecology and Management* 329: 237–254.

Tanase, M. A., R. Panciera, K. Lowell, S. Tian, J. M. Hacker, and J. P. Walker. 2014. "Airborne Multi-Temporal L-band Polarimetric SAR Data for Biomass Estimation in Semi-arid Forests." *Remote Sensing of Environment* 145: 93–104.

Thenkabail, P. S., N. Stucky, B. W. Griscom, M. S. Ashton, J. Diels, B. Van Der Meer, and E. Enclona. 2004. "Biomass Estimations and Carbon Stock Calculations in the Oil Palm Plantations of African Derived Savannas Using IKONOS Data." *International Journal of Remote Sensing* 25 (23): 5447–5472.

Timothy, D., M. Onisimo, S. Cletah, S. Adelabu, and B. Tsitsi. 2016. "Remote Sensing of Aboveground Forest Biomass: A Review." *Tropical Ecology* 57 (2): 125–132.

Vaglio Laurin, G., Q. Chen, J. A. Lindsell, D. A. Coomes, F. Del Frate, L. Guerriero, F. Pirotti, and R. Valentini. 2014. "Aboveground Biomass Estimation in an African Tropical Forest with Lidar and Hyperspectral Data." *ISPRS Journal of Photogrammetry and Remote Sensing* 89: 49–58.

Vauhkonen, J., I. Korpela, M. Maltamo, and T. Tokola. 2010. "Imputation of Single-tree Attributes Using Airborne Laser Scanning-based Height, Intensity, and Alpha Shape Metrics." *Remote Sensing of Environment* 114: 1263–1276.

Vincenzi, S., M. Zucchetta, P. Franzoi, M. Pellizzato, F. PranovI, G. A. De Leo, and P. Torricelli. 2011. "Application of a Random Forest Algorithm to Predict Spatial Distribution of the Potential Yield of Ruditapes Philippinarum in the Venice lagoon, Italy." *Ecological Modelling* 222 (8): 1471–1478.

Walker, W., A. Baccini, D. Nepstad, N. Horning, D. Knight, E. Braun, and A. Bausch. 2011. *Field Guide for Forest Biomass and Carbon Estimation; Version 1.0*. Falmouth, MA: Woods Hole Research Center.

Xu, T., L. Cao, X. Shen, and G. H. She. 2015. "Subtropical Forest Biomass Estimation Based on the Airborne Laser Radar and Landsat 8 OLI Data." *Chinese Journal of Plant Ecology* 4: 309–321.

Yamada, M., and H. L. Gholz. 2002. "An Evaluation of Agroforestry Systems as a Rural Development Option for the Brazilian Amazon." *Agroforestry Systems* 55 (2): 81–87.

Yu, X., J. Hyyppä, M. Vastaranta, M. Holopainen, and R. Viitala. 2011. "Predicting Individual Tree Attributes From Airborne Laser Point Clouds Based on the Random Forests Technique." *ISPRS Journal of Photogrammetry and Remote Sensing* 66: 28–37.

Zanne, A. E., G. Lopez-Gonzalez, D. A. Coomes, J. Ilic, S. Jansen, S. L. Lewis, R. B. Miller, N. G. Swenson, M. C. Wiemann, and J. Chave. 2009. "Global Wood Density Database." Accessed 21 September 2015. http://hdl.handle.net/10255/dryad, 235.

Zhang, J., Lin, X., 2016. "Advances in Fusion of Optical Imagery and LiDAR Point Cloud Applied to Photogrammetry and Remote Sensing." *International Journal of Image and Data Fusion*. doi:10.1080/19479832.2016.1160960.

Zhao, P., D. Lu, G. Wang, L. Liu, D. Li, J. Zhu, and S. Yu. 2016a. "Forest Aboveground Biomass Estimation in Zhejiang Province Using the Integration of Landsat TM and ALOS PALSAR Data." *International Journal of Applied Earth Observation and Geoinformation* 53: 1–15. doi:10.1016/j.jag.2016.08.007.

Zhao, P., D. Lu, G. Wang, C. Wu, Y. Huang, and S. Yu. 2016b. "Examining Spectral Reflectance Saturation in Landsat Imagery and Corresponding Solutions to Improve Forest Aboveground Biomass Estimation." *Remote Sensing* 8: 469. doi:10.3390/rs8060469.

Zhou, J. J., Z. Zhao, J. L. Liu, J. Zhao, and Q. X. Zhao. 2015. "Different Methods of QuickBird Image Information to Estimate the Accuracy of Locust Forest Effective Leaf Area Index Comparison." *Forest Science* 51 (9): 24–34.

Zhou, J. J., Z. Zhao, Q. Zhao, J. Zhao, and H. Wang. 2013. "Quantification of Aboveground Forest Biomass Using QuickBird Imagery, Topographic Variables, and Field Data." *Journal of Applied Remote Sensing* 7 (1): 073484–073484.

Zolkos, S. G., S. J. Goetz, and R. Dubayah. 2013. "A Meta-analysis of Terrestrial Aboveground Biomass Estimation Using Lidar Remote Sensing." *Remote Sensing of Environment* 128: 289–298.