Forestry *An International Journal of Forest Research*

Institute of Chartered Foresters

Forestry 2016; **89**, 422–433, doi:10.1093/forestry/cpw016
Advance Access publication 10 March 2016

A principal component approach for predicting the stem volume in Eucalyptus plantations in Brazil using airborne LiDAR data

Carlos Alberto Silva[1,2]*, Carine Klauberg[1], Andrew T. Hudak[1], Lee A. Vierling[2], Veraldo Liesenberg[3], Samuel P. C. e Carvalho[4] and Luiz C. E. Rodriguez[4]

[1]US Forest Service (USDA), Rocky Mountain Research Station, RMRS, 1221 South Main Street, Moscow, ID 83843, USA
[2]Department of Forest, Rangeland, and Fire Sciences, College of Natural Resources, University of Idaho, (UI), 875 Perimeter Drive, Moscow, ID 83843, USA
[3]Santa Catarina State University (UDESC), Av. Luiz de Camões, 2090 – Conta Dinheiro, Lages, SC 88.520-000, Brazil
[4]Department of Forest Sciences, College of Agriculture Luiz de Queiroz (ESALQ), University of Sao Paulo (USP), Av. Pádua Dias, 11, PO Box 09, Piracicaba, SP 13418-900, Brazil

*Corresponding author. Tel: +1 208 5964510; E-mail: csilva@uidaho.edu

Received 11 August 2015

Improving management practices in industrial forest plantations may increase production efficiencies, thereby reducing pressures on native tropical forests for meeting global pulp needs. This study aims to predict stem volume (*V*) in plantations of fast-growing Eucalyptus hybrid clones located in southeast Brazil using field plot and airborne Light Detection and Ranging (LiDAR) data. Forest inventory attributes and LiDAR-derived metrics were calculated at 108 sample plots. The best LiDAR-based predictors of *V* were identified based on loadings calculated from a principal component analysis (PCA). After selecting these best predictors using PCA, we developed multiple regression models predicting *V* from selected LiDAR metrics. Metrics related to tree height and canopy depth were most effective for *V* prediction, with an overall model coefficient of determination (adj. R^2) of 0.87, and a root mean squared error (RMSE) of 27.60 $\mathrm{m^3\ ha^{-1}}$ (i.e. relative RMSE = 9.99 per cent). We used this model to map stem *V* of Eucalyptus hybrid clones across the full LiDAR data extent. The accuracy and precision of our results show that LiDAR-derived *V* is appropriate for updating Eucalyptus forest base maps and registries in the paper and pulp supply chain. However, further studies are necessary to evaluate and compare the cost of acquisition and processing of LiDAR data against conventional *V* inventory in this system.

Keywords: supply chain, LiDAR metrics, remote sensing, *Eucalyptus* spp., forest management, multivariate statistics

Introduction

Eucalyptus spp. are the most important short fibre source for pulp and paper production in southeast Brazil. Extensive *Eucalyptus* spp. plantations have been established in this region since the early 1970s due to their rapid growth rate (over 40 $\mathrm{m^3\ ha^{-1}\ year^{-1}}$). The share of the Brazilian GDP represented by the planted tree industry has grown each year, closing 2014 with 1.1 per cent of all the wealth generated in the country and 5.5 per cent of industrial GDP (Ibá, 2015). Plantations now cover an area of 3.18 million hectares and account for 57 per cent of the total reforested area in Brazil (Ibá, 2015). The stem volume (*V*) production of *Eucalyptus* spp. is extremely high compared with natural forests and contributes strongly to meet current wood fibre production demands, thereby reducing pressure on native forest exploitation (Vital, 2007). As the extent of these plantations has expanded, so too has the need for accurate monitoring of forest *V* in the pulp and paper supply chain.

Forest inventory in plantation of Eucalyptus hybrid clones is usually conducted annually to monitor *V* growth, identify problematic conditions (e.g. pathogens) during initial growth stages, and determine optimal harvest time later in the growth cycle. However, this procedure is expensive and time consuming. Therefore, approaches for deriving forest inventory information based on remotely sensed data are of great utility and interest (Ponzoni and Gonçalves, 1999; Gama *et al.*, 2010). Airborne laser scanning (ALS), also referred as airborne Light Detection and Ranging (LiDAR), is a powerful tool for forestry applications (Lefsky *et al.*, 2002; Næsset, 2002, 2004, Næsset and Gobakken, 2008; Hudak *et al.*, 2009). Key LiDAR applications include high accuracy retrieval of tree density, stem volume, above ground carbon, leaf area index and basal area (Naesset, 1997; 2002; Andersen *et al*. 2005; Roberts *et al.*, 2005; Hudak *et al.*, 2006; Coops *et al.*, 2007; Silva *et al.*, 2014).

LiDAR is very useful for providing high resolution, three-dimensional information of vertical and horizontal forest structures and the underlying topography. As a result, LiDAR data provide precise information of vegetation height, density and ground elevation. From these measurements, three-dimensional digital surface models (DSMs) and bare earth digital terrain models (DTMs; also more generally referred to as a digital elevation models or DEMs) can be generated. The difference between DSM and DTM elevations results in a topographically normalized digital surface model, also known as the canopy height model (CHM).

In the case of discrete return airborne LiDAR systems, the CHM is interpolated from points representing the three-dimensional (*x*,*y*,*z*) locations of top-of-canopy elements. Because the density of LiDAR returns can range from hundreds to many thousands of points per traditional forest plot area, these points can be analyzed to provide numerous canopy structure metrics useful in modeling forest stand variables. For example, in a given area, it is possible to calculate several metrics such as maximum height, mean height, height percentiles and canopy densities. Examples of the use of LiDAR metrics in forestry, including equations for metrics calculation, can be found in Næsset (2002, 2004), Hudak *et al.* (2006), García *et al.* (2010) and McGaughey (2014).

In order to derive forest parameters using LiDAR data, it is necessary to model field-collected *in situ* variables using the most important LiDAR metrics (i.e. predictor variables) within a statistical model framework. Because the number of candidate LiDAR metrics can be very large (e.g. >30 metrics), principal component analysis (PCA) may be one useful option to reduce the number of variables used in regression-based models (Li *et al*., 2008; Mutlu *et al*., 2008; Pascual *et al*., 2010). PCA is one of the most common multivariate statistical methods, and can be used to select the subset of variables (from a large number of predictor variables) that best explain the majority of the variation in a given forest biometric (Manly, 2004). In addition to indicating which metrics can be used in regression models, the PCs themselves can be included in regression models for predicting forest biometrics.

Stem volume is a forest inventory attribute directly related to the supply of fibre to pulp and paper companies. The development of better methods for regular inventory and monitoring of industrial forests will help pulp and paper production efficiency. However, to our knowledge, the use of PCA in LiDAR-based forest biometric prediction has been little studied. Therefore, our objectives were to: (i) select the best LiDAR-derived metrics for *V* modeling according to the PCA analysis; (ii) select the best model to predict and map *V*; and (iii) generate maps representing the spatial distribution of *V* in different plantations of Eucalyptus hybrid clones of different ages. This investigation was based on the hypothesis that LiDAR data and PCA analysis can facilitate precise and accurate inferences of *V* in Eucalyptus hybrid clones plantations in southeast Brazil.

Methods

Study area description

The study area consisted of six farms located within the Paraíba Valley in the state of São Paulo, Brazil (Figure 1). The climate of the region is characterized

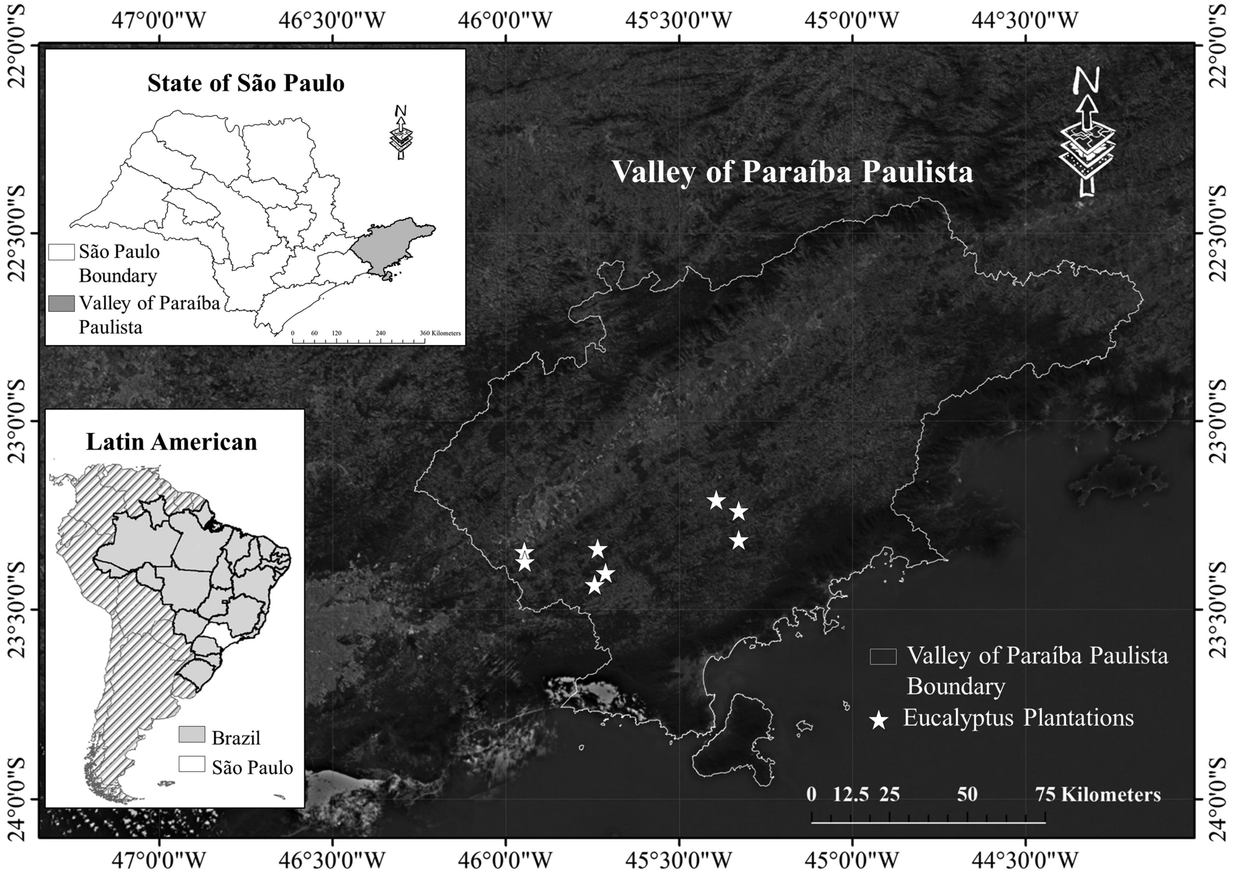

Figure 1 Location of the study area in the State of São Paulo, Brazil. The stars indicate the location of the Eucalyptus plantations.

as humid subtropical, with dry winters and hot summers (Cwa) (Köppen and Geiger, 1928). Annual average precipitation is ~1200 mm; average temperature ranges from a minimum of 17.1° C in the coolest month (July) to a maximum of 23.9° C in the hottest month (February). The topography in the selected plantations is complex with high relief, ranging from 578 to 1310 m in elevation.

The plantations contained seven hybrid clones of two species of eucalyptus, *Eucalyptus grandis* W. Hill ex Maid and *Eucalyptus urophylla* S.T. Blake. The plantations are managed by Fibria Celulose S/A, a pulp company. Stand age across the farms was variable and ranged from three to 8 years. All the trees were planted in a 3 × 2 m grid configuration, resulting in an average density of 1667 trees per ha.

Field data collection

A total of 108 circular plots of 400 m^2 each were established across the six farms for stand measurement. All plots were georeferenced with a geodetic GPS with differential correction capability (Trimble Pro-XR). For each GPS location, we recorded data for a time period ranging from 20–40 min, which allowed us to reduce the horizontal error to the level of 10 cm. In each sample plot, individual trees were measured for DBH and a subsample (15 per cent) of trees for maximum height (Ht). For trees in the plot that were not directly measured for Ht, the inventory team of the Fibria Celulose S/A company predicted heights from hypsometric models, which are models that use DBH as a predictor of Ht. The *V* (m^3 $tree^{-1}$) was obtained through the Spurr linear model (Spurr, 1952) adjusted for each farm, employing as independent variables the square of DBH and the Ht, and *V* as the dependent variable, following the model below.

$$V = \beta_0 + \beta_1 \times \mathrm{DBH}^2 \times \mathrm{Ht} + \varepsilon, \qquad (1)$$

where β_0 is the linear coefficient; β_1 is the slope coefficient; DBH is the diameter at breast height (1.30 m); Ht is the height and ε is the error.

The *V* models had adjusted coefficients of determination (adj. R^2) ranging from 0.96 to 0.99 and relative RMSE per cent ranging from 3.18 to 6.09 per cent (Table 1).

Differences in the average values of the DBH, Ht and *V* are mainly related to the age of the *Eucalyptus* spp. stands. However, differences may also depend on other factors such as the type and intensity of land use before the establishment of the stands; amount of soil compaction or reduction of edaphic fertility; and site index. Normally, at early stand ages there are small diameter DBH values and consequently small BA, Ht and *V* values. With increased stand age, specific site index values tend to increase, and well-defined vertical strata develop. The summed *V* content of all trees in the sample plot was multiplied by the plot area (0.04 ha) to calculate the *V* stored in the sample plot in m^3 ha^{-1}. Summary statistics of DBH, Ht and V measured in the stands are presented in Table 2.

Table 1 Summary of the *V* models

Eucalyptus spp. plantations	Model coefficients			RMSE	
	β_0	β_1 ($\times 10^{-5}$)	Adj. R^2	m^3 $tree^{-1}$	%
F986	0.006	3.280	0.968	0.006	4.813
F849	0.002	3.430	0.978	0.009	4.789
F950	0.014	3.090	0.976	0.010	6.006
F184	0.005	3.360	0.989	0.006	3.188
F166	0.003	3.330	0.984	0.012	5.116
F634	0.005	3.300	0.982	0.015	6.092

β_0= linear coefficient; β_1= slope coefficient; adj. R^2= adjusted coefficient of determination; RMSE= root mean square error.

LiDAR data acquisition and data processing

LiDAR data were obtained by a Riegl LMS-Q680I sensor mounted on a Piper Seneca II aircraft. The characteristics and precision of the LiDAR data are listed in Table 3. LiDAR data processing consisted of several steps that ingested the lidar point cloud data and provided four major outputs: the DTM, the digital surface model (DSM), the CHM, and the LiDAR-derived canopy structure metrics. All of the LiDAR processing phases were performed using US Forest Service FUSION/LDV 3.42 software (McGaughey, 2014).

Initially the *catalog* function in FUSION/LDV was used to evaluate the quality of the LiDAR data set. A classification algorithm based on Kraus and Pfeifer (1998) and available in the *groundfilter* function was applied to differentiate between ground and vegetation points. DTMs were generated using the classified ground points with a spatial resolution of one metre using *gridsurfacecreate*. The *canopymodel* tool was then used to interpolate the vegetation points and to generate DSMs with a spatial resolution of 1 m.

After generating the DSMs, the *clipdata* function was applied to normalize heights and to assure that the *z* coordinate for each point corresponded to the height above ground and not the orthometric elevation of the single point. The *canopymodel* function was applied again, but at this time it was used to create the CHM, which provided an estimate of vegetation height. The *polyclipdata* function was then used to subset of the LiDAR points within each of the 108 *in situ*-measured sample plots, and the *cloudmetrics* tool was applied to compute the LiDAR metrics as derived from the point cloud (McGaughey, 2014). Finally, the *gridmetrics* functions were used to generate the same LiDAR metrics as computed with *cloudmetrics*, but within grid cells of 3 m spatial resolution across the landscape.

In this study, we considered only the first returns to compute the LiDAR metrics. The first returns of the LiDAR pulses most likely reflect canopy tops and, according to Bater *et al.* (2011), the first returns are more stable than other returns in characterizing forest structure. Moreover, Silva *et al.* (2014) showed that LiDAR metrics derived from first returns strongly correlate with forest attributes, such as aboveground carbon in a plantation of Eucalyptus hybrid clones in southeast Brazil. Therefore, 31 LiDAR metrics calculated from first returns were considered for *V* modeling (Table 4). We decided to not use the entire suite of *cloudmetrics* in FUSION, because some of the metrics derived from this process did not have a reasonably intuitive physical meaning or relationship with forest attributes. Moreover, we considered only those metrics that have been frequently used as candidate predictors for forest attribute prediction in other studies (Næsset, 2002;2004; García *et al*., 2010; Hudak *et al*., 2012; Silva *et al*., 2014).

LiDAR metrics selection and regression model development

We used three data analysis approaches to model *V* of *Eucalyptus* spp. First, Pearson's correlation (*r*) was used to identify highly correlated predictor variables ($r > 0.9$) as presented in Hudak *et al.* (2012) and Silva *et al.* (2014). If a given group (two or more) of LiDAR metrics were highly correlated, we retained only one metric by excluding the others that were most highly correlated with the remaining metrics. Second, PCA was applied to the selected best predictor LiDAR metrics, and the metrics that were most likely to contribute to model development were identified by inspecting the eigenvectors in each PC. We then used those metrics with highest loading on the PCs as input variables in multivariate linear regression models predicting stand structure attributes.

An example of the use of PCA, including the equations used to obtain the eigenvalues, eigenvectors and the principal component (PC) scores, may be found in Jensen (2005). In the present study, PCA was applied over the selected LiDAR metrics for each of the 108 sample plots using the *prcomp* function from the stats package in *R* (R Core Team, 2014). A correlation matrix derived from the LiDAR metrics provided the basis for the eigenvalue and eigenvector calculations and for the subsequent determination of the

Table 2 Characteristics of the six plantations of Eucalyptus hybrid clones

Eucalyptus spp. plantation	Area (ha)	DBH	*H*	*V*	*N*	Age
F986	94.16	12.70 (1.75)	18.52 (1.43)	173.73 (73.23)	20	3.3
F849	138.96	14.13 (2.40)	22.16 (1.54)	272.73 (35.59)	26	4.7
F950	86.72	13.72 (2.55)	21.44 (3.44)	266.86 (44.14)	17	5.5
F184	58.34	14.55 (2.26)	23.67 (1.18)	291.16 (30.76)	20	5.9
F166	84.35	14.57 (3.25)	24.17 (1.42)	324.83 (44.35)	16	6.1
F634	84.80	14.35 (3.78)	25.89 (2.15)	374.79 (45.65)	14	8.0
Total	586.86	–	–	1755.58	108	–
Average	83.84	13.54 (2.98)	21.19 (0.80)	250.80 (17.57)	–	5.11

The values are based on *in situ* measured sample plots ($n = 108$). Standard deviation values are given in brackets.

Table 3 LiDAR flight characteristics

Parameter	Value
Average flight height	422.94 m
Average density	10 pulses m^{-2}
Pulse frequency	400 kHz
Scan angle	±45®
Laser wavelength	1055 nm
Average aircraft speed	57 m s^{-1} (205.20 km h^{-1})
Horizontal precision	0.1–0.15 m (1.0 sigma)

Table 4 LiDAR-derived canopy height metrics considered as candidate variables for predictive *V* models (McGaughey, 2014)

Variable	Description
HMIN	Height minimum
HMAX	Height maximum
HMEAN	Height mean
HMAD	Height median absolute deviation
HSD	Height standard deviation
HSKEW	Height skewness
HKURT	Height kurtosis
HCV	Height coefficient of variation
HMODE	Height mode
H01TH	Height 1st percentile
H05TH	Height 5th percentile
H10TH	Height 10th percentile
H15TH	Height 15th percentile
H20TH	Height 20th percentile
H25TH	Height 25th percentile
H30TH	Height 30th percentile
H35TH	Height 35th percentile
H40TH	Height 40th percentile
H45TH	Height 45th percentile
H50TH	Height 50th percentile
H55TH	Height 55th percentile
H60TH	Height 60th percentile
H65TH	Height 65th percentile
H70TH	Height 70th percentile
H75TH	Height 75th percentile
H80TH	Height 80th percentile
H90TH	Height 90th percentile
H95TH	Height 95th percentile
H99TH	Height 99th percentile
CR	Canopy relief ratio ((HMEAN − HMIN)/(HMAX − HMIN))
COV	Canopy cover (percentage of first return above 1.30 m)

PC scores. Each score represented a transformed metric from the linear combination of the LiDAR metrics of the sample plots. By analyzing the eigenvectors and the PC score, we could establish differences in the contribution of each LiDAR metric to the variability in the dataset, as well as the similarity in metrics calculated across the different aged stands (Manly, 2004; Li *et al*., 2008).

We used the *lm* linear model function in R statistical software (R Development Core Team, 2015) to develop the multiple linear regression models, and the Shapiro–Wilk (Shapiro and Wilk, 1965) and Breusch–Pagan (Breusch and Pagan, 1979) tests to evaluate the normality and heteroscedasticity of each model. In addition, the corrected Akaike information criterion (AICc) (Akaike, 1973, 1974) was calculated in order to measure the relative quality of each proposed model and to rank them accordingly (Hurvich and Tsai, 1989).

The precision and accuracy of estimates for each model were evaluated in terms of adjusted coefficient of determination (adj. R^2), absolute and relative root mean square error (RMSE), and absolute and relative bias:

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

$$\mathrm{BIAS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i), \qquad (3)$$

where n is the number of plots, y_i is the observed value for plot i, and $\hat{y}_i$ is the predicted value for plot i. Moreover, relative RMSE and biases were calculated by dividing the absolute values of RMSE and BIAS (equations 2 and 3) by the mean of the observations. We defined acceptable model precision and accuracy as a relative RMSE and bias of <15 per cent.

The best model was selected based on the AICc values, and its performance was evaluated by means of leave-one-out cross-validation (LOOCV). We also used the equivalence test (Robinson, 2015) to verify if the observed and predicted *V* values were statistically equivalent. Finally, we used the *AsciiGridPredict* function from the yaImpute package in *R* (Crookston and Finley, 2008) to apply the selected best model across the landscape to map the spatial distribution of *V* of *Eucalyptus* spp at the stand level, with spatial resolution of 3 m, for the benefit of forest managers. An overview of the methodology is outlined in Figure 2.

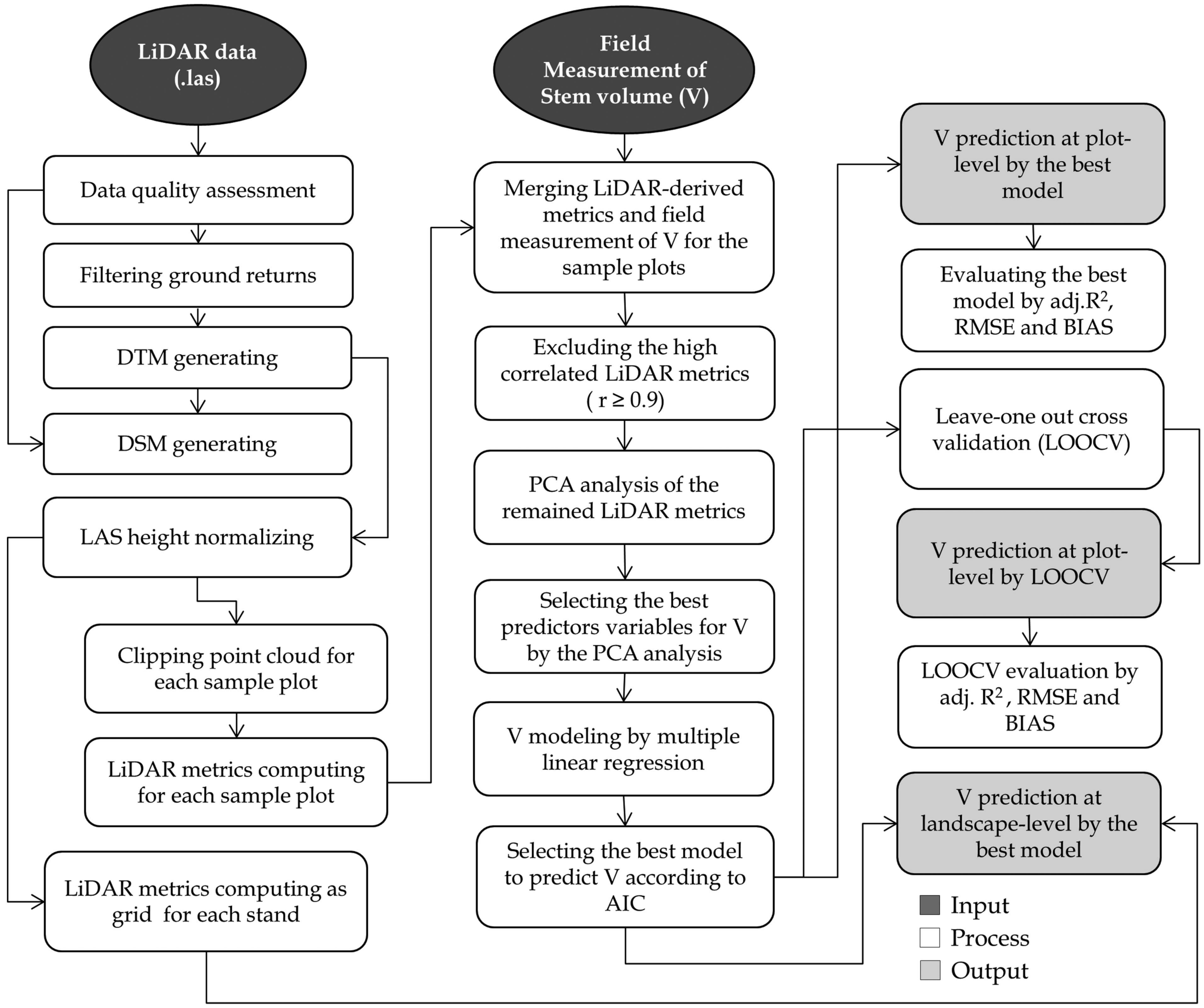
LiDAR data (.las)
Field Measurement of Stem volume (V)
Data quality assessment
Filtering ground returns
DTM generating
DSM generating
LAS height normalizing
Clipping point cloud for each sample plot
LiDAR metrics computing for each sample plot
LiDAR metrics computing as grid for each stand
Merging LiDAR-derived metrics and field measurement of V for the sample plots
Excluding the high correlated LiDAR metrics (r ≥ 0.9)
PCA analysis of the remained LiDAR metrics
Selecting the best predictors variables for V by the PCA analysis
V modeling by multiple linear regression
Selecting the best model to predict V according to AIC
V prediction at plot-level by the best model
Evaluating the best model by adj.R², RMSE and BIAS
Leave-one out cross validation (LOOCV)
V prediction at plot-level by LOOCV
LOOCV evaluation by adj. R², RMSE and BIAS
V prediction at landscape-level by the best model
■ Input
□ Process
▨ Output

Figure 2 Procedure for predicting stem volume (*V*) from LiDAR and inventory plot data in Eucalyptus plantations. AIC, Akaike Information Criterion; LOOCV, leave-one-out cross-validation; PCA, principal components analysis; RMSE, root mean squared error.

Results

Height variation of Eucalyptus spp. plantations

Variation in height for selected sample plots from Eucalyptus hybrid clones between early (i.e. 3.3 years), intermediate (i.e. 5.5 years), and advanced (i.e. 7.9 years) stages of development are shown in Figure 3. Although located at different plantations and therefore under distinct site indices, the LiDAR-derived height increased with age across all sites (Figure 3). On the other hand, the same was not observed for the number of LiDAR returns in the strata between 2 and 15 m in height, where the number of returns decreased with age.

In a plantation environment, young trees of Eucalyptus normally have a well-defined canopy with numerous branches, while mature trees, due to competition, higher canopy closure and light limitation, retain a decreased number of lateral branches. The number of branches was strongly reduced during advanced canopy growth ages (i.e. Figure 3c), resulting in fewer LiDAR returns at intermediate heights. In addition, as the stands approached harvest age, the ground floor had more small trees, bushes and grasses established due to the greater time since silvicultural treatment.

Highly correlated LiDAR metrics

Pearson's correlation test (*r*) showed that among the 31 candidate LiDAR metrics, 23 were highly correlated ($r > 0.9$). We kept one of the highly correlated metrics (H99TH), which along with seven other remaining metrics not highly correlated ($r < 0.9$) were included in PCA analysis. LiDAR metrics that were retained after correlation analysis included HSD, HCV, HSKEW, H01TH, H05TH, H99TH, CR and COV. The correlation structure of these eight metrics is shown in Table 5.

PCA of LiDAR metrics

The first five of eight PCs accounted for 97.7 per cent of the total variance contained in the selected set of eight LiDAR metrics.

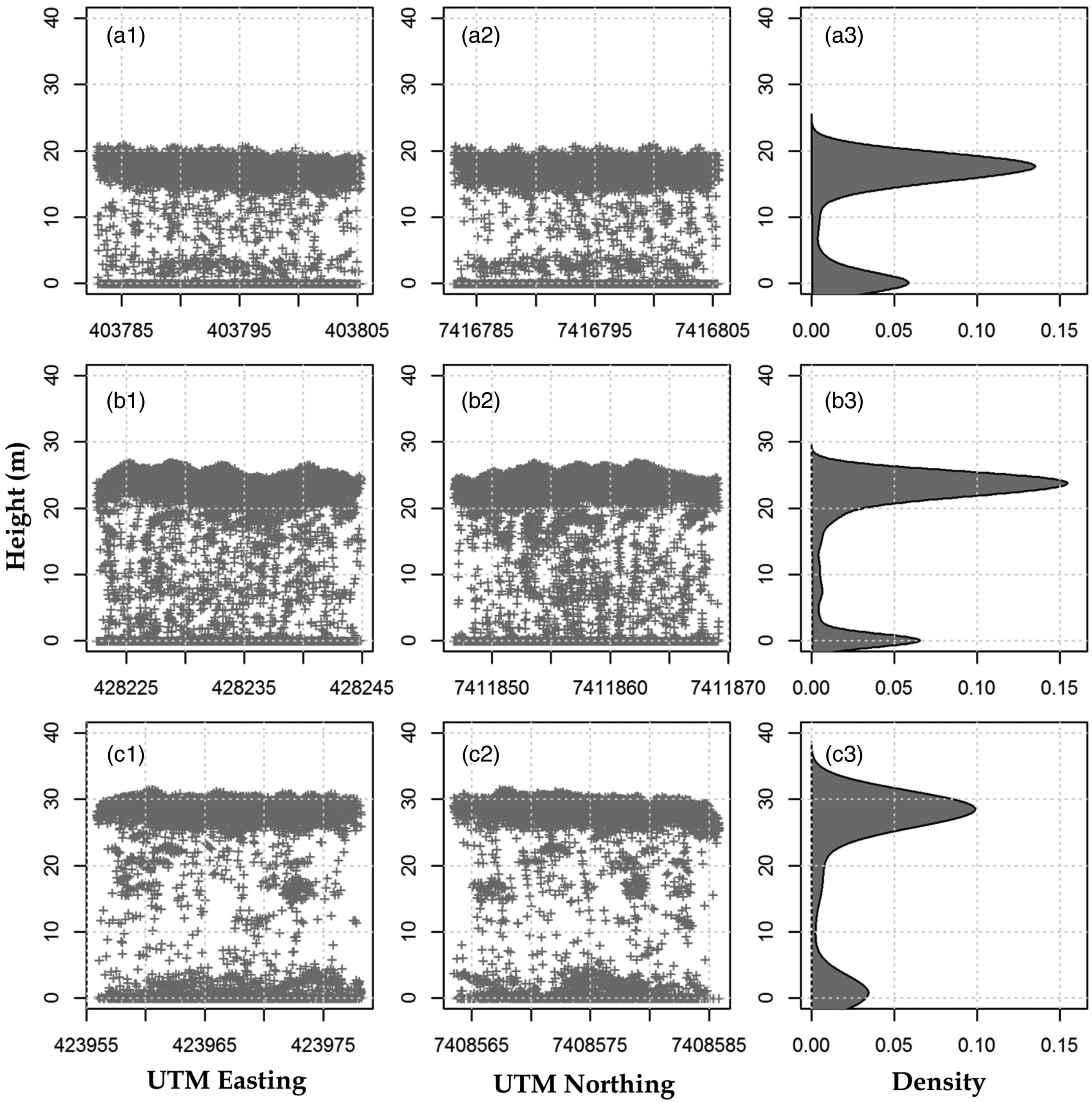

Figure 3 LiDAR profiles of selected sample plots of Eucalyptus representative of early (i.e. 3.3 years) (a), intermediate (i.e. 5.5 years) (b) and advanced (i.e. 7.9 years), (c) stages of development. (1) UTM Easting profile, (2) UTM Northing profile and (3) Density plot (canopy height profile).

Specifically, PC1, PC2, PC3, PC4 and PC5 accounted for 60.8, 16.9, 10.7, 6.9 and 2.3 per cent of the total variance, respectively (Figure 4). We opted to use the first five PCs to select the best LiDAR metrics for V modeling because PCs 6–8 explained a less than significant percentage (<2.5 per cent) of the remaining variance.

The PC eigenvector loadings (Table 6), which represented the contribution of each LiDAR metric toward the component, showed both negative and positive values. PC1 was expressed as positive loadings of HCV, followed by HSKEW (both with absolute $r > 0.90$ with PC1). On the other hand, PC2 showed positive loadings of H99TH ($r = 0.95$), whereas PC3, PC4 and PC5 showed positive loadings of COV ($r = 0.72$) and negative loadings of H01TH ($r = -0.51$) and H05TH ($r = -0.28$), respectively.

PC1 was highly correlated with HCV indicative of increasing canopy height variance associated with the ages of the Eucalyptus hybrid clones plantations and silvicultural treatments before harvesting, corroborating Figure 3. Three major groups are highlighted for the first two PCs (Figure 5). The first group representing PC1 highlights canopy height variation, while PC2 highlights canopy height measures and PC3 highlights canopy cover. LiDAR metrics selected from the first three PCs were not highly correlated with the selected metrics highlighted from the remaining PCs (Table 5). Therefore, the five main LiDAR metrics selected from the five first PCs were HCV, H99TH, COV, H01TH and H05TH (Table 6).

Stem volume modeling

Table 7 shows the performance of four multivariate linear regression models created to predict *V* of Eucalyptus hybrid clones based on the LiDAR metrics highlighted previously by PCA. Results from the Shapiro–Wilk test and Breusch–Pagan test reveal that the data were normally distributed, and heteroscedasticity had a 0.05 per cent level of significance. The addition of more terms into the models did not significantly improve model fit while increasing the AICc statistic (from 1031.4 to 1036.3). The model including just the HCV and H99TH metrics produced the lowest AICc statistic and was therefore selected as the best model.

Table 5 Pearson correlations among LiDAR metrics selected

R	HSD	HCV	HSKEW	H01TH	H05TH	H99TH	CR	COV
HSD	1.00							
HCV	0.89	1.00						
HSKEW	0.79	0.87	1.00					
H01TH	−0.54	−0.60	−0.56	1.00				
H05TH	−0.64	−0.81	−0.82	0.66	1.00			
H99TH	0.63	0.24	0.21	−0.01	0.12	1.00		
CR	−0.68	−0.88	−0.85	0.42	0.72	−0.06	1.00	
COV	−0.31	−0.48	−0.37	0.21	0.33	0.07	0.46	1.00

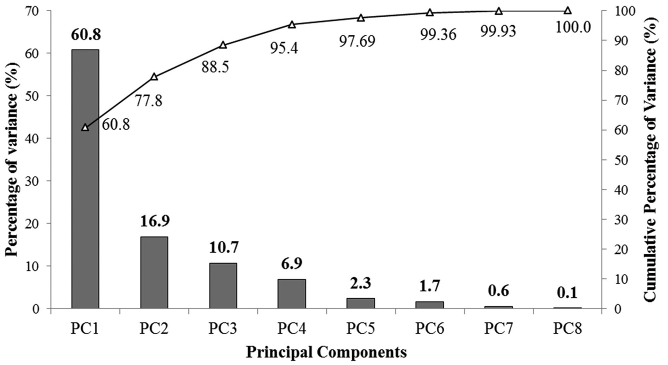

Figure 4 The percentage of variance and cumulative percentage of variance in *V* explained by the eight PCs.

The best model resulted in an adj. $R^2 = 0.84$, $r = 0.92$, RMSE = 27.6 m^3 ha^{-1} (9.99 per cent), and Bias = 0 (0 per cent) (Table 7). The LOOCV analysis revealed a highly stable model (Figure 6b), and the normal *Q*–*Q* and residual plots (Figure 6c and D) confirmed from a graphic perspective that this model met the parametric assumptions of normality and homoscedasticity.

Results from the statistical equivalence test between the best model and the LOOCV are presented in Figure 6a,b. The equivalence plot design presented here is an adaptation of equivalence plots presented by Robinson (2015). The grey polygon represents the ±25 per cent region of equivalence for the intercept, and the black vertical bar represents a 95 per cent confidence interval for the intercept. The predicted *V* from the model and the LOOCV are equivalent to the reference for the intercept because the black bar was completely within the grey polygon. If the grey polygon is lower than the black vertical bar, the predicted *V* would be biased low; if it is higher than the black vertical bar, the predicted *V* would be biased high. Moreover, the grey dashed line represents the ±25 per cent region of equivalence for the slope, and if the black vertical bar is contained completely within the grey dashed line, the pairwise measurements are considered to be equivalent. A bar that is wider than the region outlined by the grey dashed lines indicates highly variable predictions. The white dots are the pairwise measurements, and the solid line is a best-fit linear model for the pairwise measurements.

Predicted *V* of Eucalyptus clones for the 108 sample plots ranged from 117 to 427 m^3 ha^{-1}. The predicted *V* means for the six farms were 187 (F986), 254 (F950), 266 (F849), 288 (F184), 332 (F166) and 372 (F634) m^3 ha^{-1} (Figure 7). Predicted *V*s were well balanced overall, being slightly overpredicted during early and advanced stand ages and slighty underpredicted at intermediate ages. These differences may reflect varying site indices and management practices across the plantations. The Eucalyptus plantations containing younger stands showed the lowest *V* values (i.e. IDs F986 and F950). Advanced age stands contained the highest stem volumes (i.e. IDs F166 and F634). Figure 8

Table 6 Loadings and eigenvectors for the first five PCs

PCs	Ev	Eigenvectors (Eg)							
		HSD	HCV	HSKEW	H01TH	H05TH	H99TH	CR	COV
PC1	4.87	0.40	**0.44**	0.42	−0.31	−0.39	0.11	−0.40	−0.22
PC2	1.35	0.38	0.03	0.01	0.14	0.27	**0.82**	0.12	0.27
PC3	0.86	0.00	−0.06	0.05	−0.51	−0.27	−0.11	0.17	**0.79**
PC4	0.55	0.07	−0.07	−0.26	**−0.69**	0.11	0.15	0.44	−0.47
PC5	0.19	0.17	−0.01	0.03	0.38	**−0.66**	0.02	0.62	−0.11

PC is the given PC; Ev is the eigenvalues for each PC. Check Table 3 for the description of the LiDAR-derived metrics. Bold characters indicate the LiDAR metric with highest loading on the PC.
Bold values indicate the largest contributing LiDAR metric for a given PC.

shows the *V* maps with spatial resolution of 3 m over the six plantations predicted from the best model (Table 7; Eq. 1).

Discussion

Accurate estimates of *V* are critical for forest plantation inventory and planning. The development of methods that provide better estimates of *V* for regular monitoring of industrial forests is important for increasing forest management efficiency. This study presents a novel framework for predicting and mapping *V* in six fast-growing plantations of Eucalyptus hybrid clones using airborne LiDAR data and PCA.

LiDAR has been shown to be a powerful technology for inventory of Eucalyptus plantations (Silva *et al*., 2014; Carvalho *et al*., 2015); as expected, there is a significant relationship between V and LiDAR-derived metrics selected from the PCA analysis. In this study, the HCV and H99TH metrics were indicated by the PCA as the best predictor variables for determining *V* in plantation of Eucalyptus hybrid clones. This finding is consistent with previous studies that have shown LiDAR-derived metrics such as HCV and H99TH to be effective predictors of forest attributes, such as stem volume, height, basal area, and aboveground carbon in *Eucalyptus* spp. plantations (Packalen *et al*., 2011; Tesfamichael and Jan Van Aardt, 2010; Silva *et al*., 2014). As described in Li *et al.* (2008), these selected LiDAR metrics can succinctly describe 3D forest canopy structure because they capture most of the information contained in the canopy point cloud.

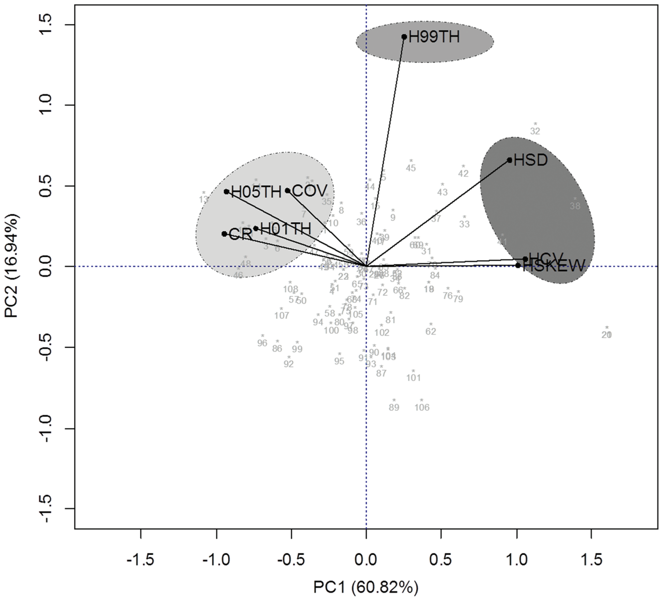

Figure 5 Projection of the first two PC scores from the selected LiDAR metrics. The grey ellipsoids represent the visual LiDAR metrics clusters. The grey points represent the values of PC1 *vs* PC2, and grey numbers represent the ID number of the sample plot ranging from 1 to 108 ($n = 108$). See Table 3 for the description of the LiDAR-derived metrics.

Using PCA analysis to guide LiDAR metrics selection can be advantageous relative to other methods such as stepwise variable selection. This is based on the fact that PCA can capture the covariance structure among candidate metrics in multiple dimensions. For example, each PC biplot (e.g. PC1 and PC2 in Figure 5) represents clusters of metrics, and the metrics in each cluster have magnitude and direction. The HCV and H99TH are orthogonal metrics that capture the majority of variation in PC1 and PC2 and were the most two important metrics used in the *V* modeling. In this case, HCV represents canopy depth variation and H99TH represents canopy height. Large HCV indicates that some trees are smaller than others (more variability), which would result in less volume for a plot with a similar H99TH value. The HCV term, therefore, adjusts the estimate of *V* downward to account for variation due to these smaller trees.

The most accurate method of predicting *V* in plantation of Eucalyptus hybrid clones is to physically sample it in the field using forest mensuration techniques. In a conventional inventory, one sample plot (300–500 m^2) is normally established and measured every 10 or 15 ha, with a goal to achieve a maximum acceptable relative RMSE of 10–15 per cent (Batista *et al*., 2014). However, this type of measurement over large areas is limited by budgets and time, making it impractical. Approaches for deriving forest inventory information based on LiDAR data are of great utility and interest owing to their promise for improving spatial sampling capabilities within plantations. In this study, we demonstrated that LiDAR can be used to predict *V* over large areas with RMSE of <15 per cent, which is equal to or less than the level of error that is traditionally accepted in a conventional field inventory.

LiDAR data have been used for forest inventory in countries such as Norway, USA and Canada (Næsset 1997, 2002, 2004; Hudak *et al*., 2006; Coops *et al*., 2007); however, the application of airborne LiDAR technology for Brazilian industrial forest management is relatively new. Zandoná (2006) and Macedo (2009) applied LiDAR data to detect individual trees and model stand volume in both *Pinus* spp. and *Eucalyptus* spp. plantations. Rodriguez *et al.* (2010) and Zonete *et al.* (2010) predicted diameter at breast height (DBH), height (Ht) and basal area (BA) in *Eucalyptus* spp.

Table 7 Adjusted coefficients of determination (adj. R^2), root mean square error (RMSE) and the corrected Akaike information criterion (AICc) of the regression models to predict stem volume (*V*) (m^3 ha^{-1})

Equation	Models	Adj. R^2	RMSE	RMSE%	AICc
1	**=−268.40−0.24HCV + 20.33H99TH**	**0.84**	**27.60**	**9.99**	**1031.41**
2	=−268.42−0.15HCV + 20.26H99TH + 0.25COV	0.84	27.62	9.99	1033.43
3	=−268.20−0.57HCV + 20.45H99TH + 0.12COV −2.85H01TH	0.84	27.45	9.94	1034.40
4	=−250.00−1.02HCV + 20.08H99TH −0.023COV −2.35H01TH −0.68H05TH	0.84	27.40	9.93	1036.31

BIAS = 0(0%) for all the models. Bold values represent the best model.

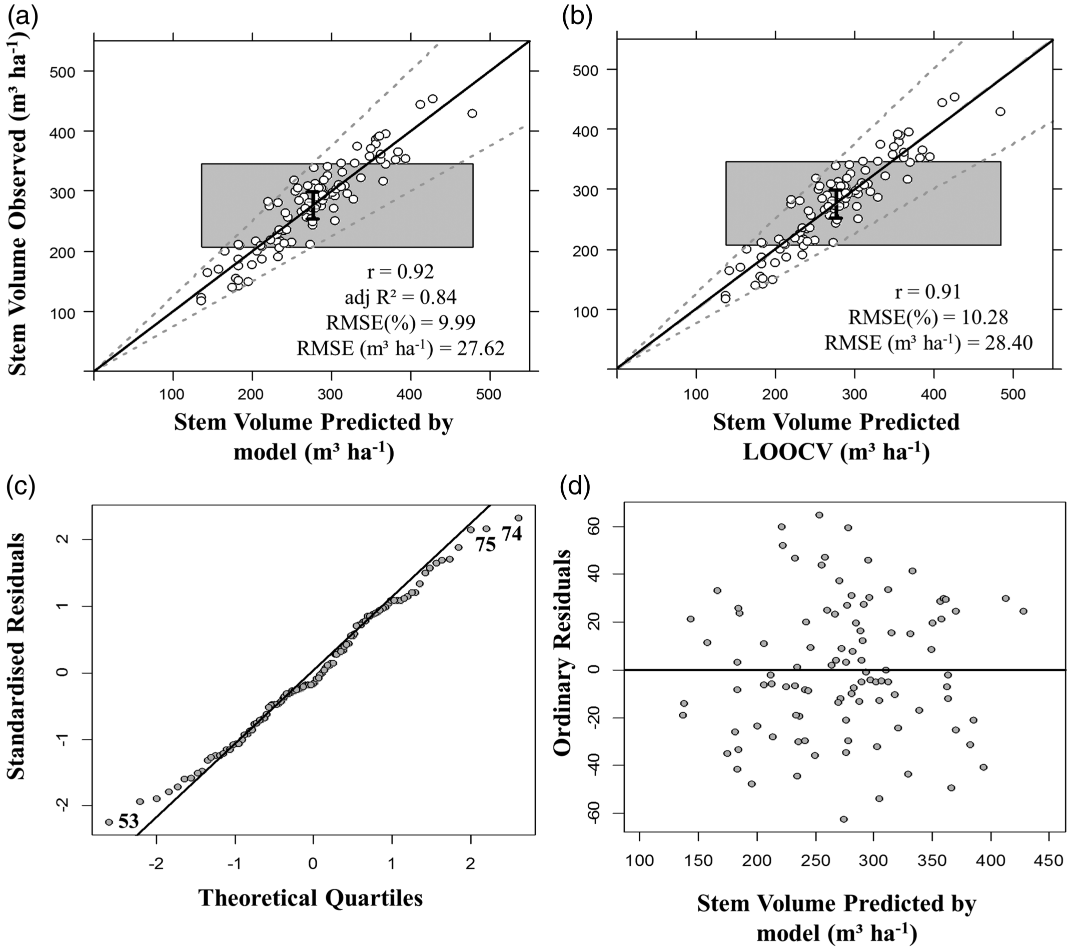

Figure 6 Equivalence plot of the predicted *vs* observed *V* (a); equivalence plot of the observed *vs* the LOOCV predicted *V* (b); normal *Q*–*Q* plot – standardized residuals *vs* theoretical quartiles; the numbers 53, 74 and 75 are the IDs of three sample plots that are outliers according to the Normal *Q*–*Q* analysis (c); and the ordinary residuals *vs* predicted *V* by the model (d); ($n = 108$).

plantations. More recently, Silva *et al.* (2014) showed that LiDAR measurements can also be used to model different components of aboveground carbon stocks (i.e. total, commercial, residuals) in *Eucalyptus* spp. plantations. Our findings are comparable to those of Zonete *et al.* (2010), who used multilinear regression models employing LiDAR-derived metrics as independent variables to predict DBH, height and BA in *Eucalyptus* spp. plantations in Bahia State (NE-Brazil).

Although the cost of LiDAR data acquisition was not a central objective to evaluate in this study, it is nonetheless an important factor to consider. Many factors influence the cost of LiDAR data. These factors include normal cost variables, such as project area size and location, the level of detail needed (pulse density – number of pulses sent by the sensor per m^2), as well as market variables, such as competition between LiDAR vendors. For *Eucalyptus spp.* inventory in southeast Brazil, the cost of LiDAR data acquisition is affordable because the large extent of plantation areas decreases the acquisition cost per unit area. Also, due to the low amount of canopy variability in this type of plantation, it is not necessary to acquire a point cloud with a high level of detail. Therefore, the cost of acquiring LiDAR data in a *Eucalyptus spp.* plantation can be further decreased by acquiring LiDAR data at low pulse

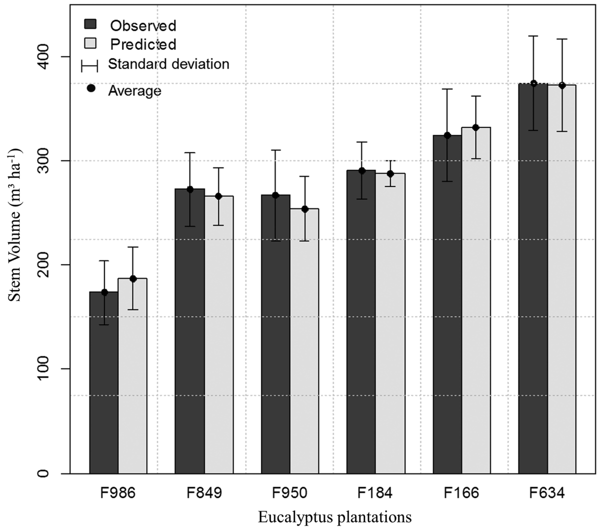

Figure 7 Comparison of observed and predicted *V* values of Eucalyptus plantations in the sample plots. See Table 1 for descriptions of the plantation codes.

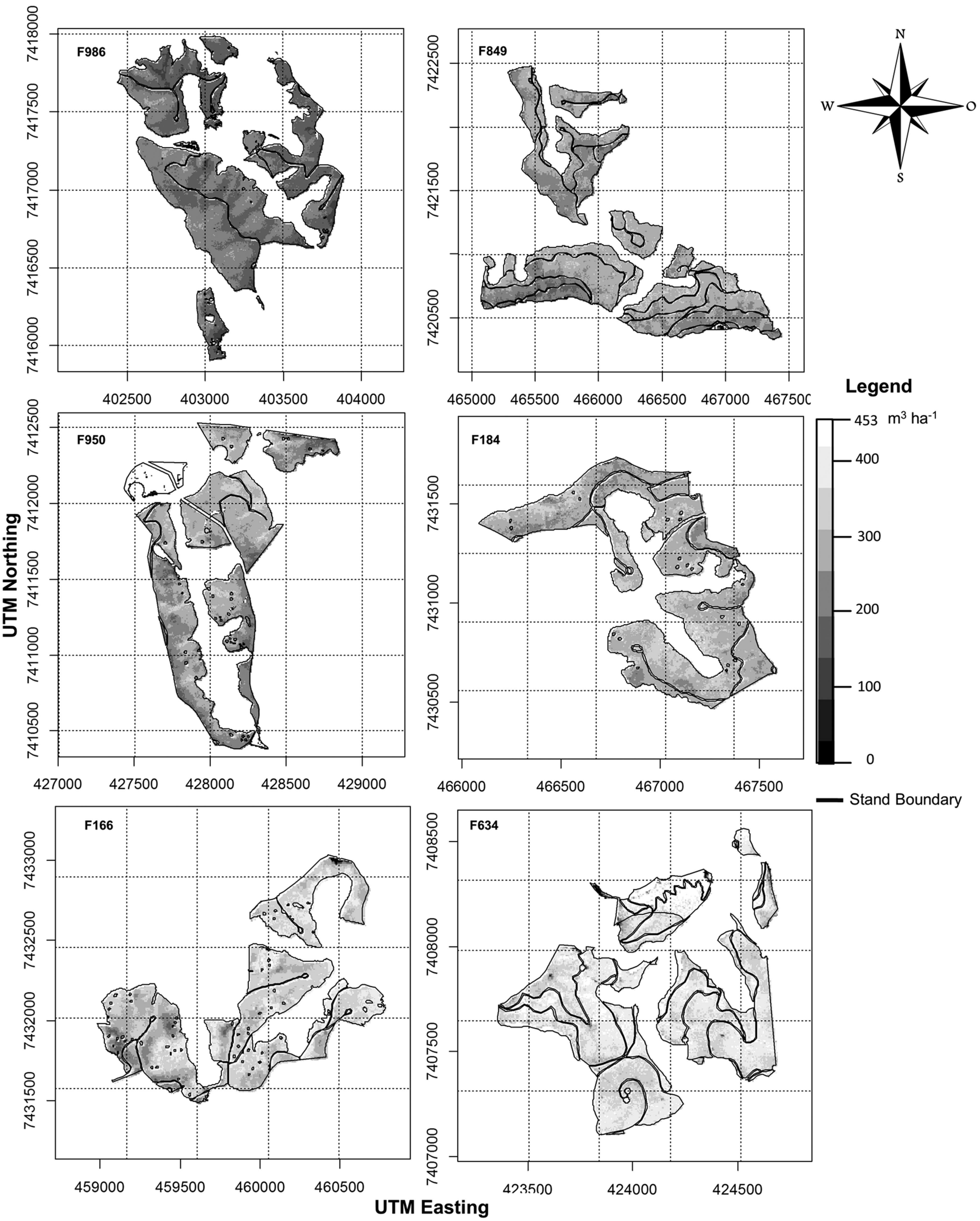

Figure 8 Predicted *V* for the six Eucalyptus plantations. See Table 1 for descriptions of the plantation codes.

density. An earlier study indicate that a low LiDAR pulse density provides similar results as a high LiDAR pulse density for predicting forest attributes in a *Eucalyptus* spp. plantation (Gonzalez-Ferreiro *et al*., 2013).

In addition to data acquisition cost, it is also important to take into account the cost of LiDAR data processing and modeling for deriving forest biophysical attributes. Several software packages currently exist for visualizing and processing LiDAR data, with both proprietary and open source options. Here, we presented a framework for processing LiDAR data and modeling *V* in plantation of Eucalyptus hybrid clones using the free and open source software packages FUSION and *R*. Specifically, we spent about three person-weeks processing the LiDAR data and modeling the *V* across the entire landscape. The time required to process field and LiDAR data and develop predictive models is directly influenced by the data volume, the presence of data outliers, the number of forest attributes modelled and the experience level of the technician.

The cost of using LiDAR technology for forest inventory could still be highly expensive in many situations; however, LiDAR has the advantage of predicting and mapping forest attributes at the landscape level with high accuracy, along with other natural resource management applications (Hudak *et al*., 2009). In a conventional inventory, spatial variability of forest attributes within stands is not normally considered. Hummel *et al*. (2011) found that the accuracy and cost of a LiDAR-based inventory summarized at the stand level was comparable to traditional stand-level, ground-based assessments for structural attributes. However, the LiDAR data were able to provide information across a much larger area and at a higher spatial resolution than stand-level, ground-based assessments alone. In this study, we mapped *V* across the landscape at a spatial resolution of 3 m. Therefore, the method employed to produce spatially explicit *V* predictions (Figure 8) would be applicable to support the supply chain of pulp and paper companies in Brazil or elsewhere. Although the range of structural variability represented in this study was large due to the many age classes and hybrid clone varieties sampled, it is possible that the framework developed here may need to be validated and further refined in other Eucalyptus plantation types.

Conclusion

In this study, we demonstrated the use of LiDAR and PCA analysis for *V* modeling in six plantations of Eucalyptus hybrid clones in southeast Brazil. We found that LiDAR measurements can be used to predict *V* across variable-age Eucalyptus plantations with adequate levels of precision and accuracy. Secondly, we found that PCA can be used to identify the best predictors to be included in multiple regression models. Thirdly, we found that the HCV and H99TH metrics were the most important LiDAR metrics for modeling *V* in this study. Finally, the spatial distribution of *V* stocks can be precisely mapped providing key information for the supply chain of a pulp and paper company. Even though we did not evaluate the cost of LiDAR data acquisition and processing, the framework presented herein can serve as a useful methodology, and we hope that the promising results for *V* modeling in this study will stimulate further research and applications not just in plantations of Eucalyptus hybrid clones in southeast Brazil, but in other plantation types elsewhere.

Funding

This research was possible due to the support of International Programs of the US Forest Service (USFS) during an internship by the first author at the USFS Rocky Mountain Research Station Forestry Sciences Laboratory in Moscow, Idaho. This research was supported by The State of São Paulo Research Foundation-FAPESP (Process 12/03176-0 and 13/05081-9). LiDAR and field inventory data collections were funded by FIBRIA S/A, a Brazilian pulp and paper company.

Conflict of interest statement

None declared.

References

Akaike, H. 1973 Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*. Petrov, B.N. and Csake, F. (eds). Akademiai Kiado, pp. 267–281.

Akaike, H. 1974 A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723.

Andersen, H., McGaughey, R.J. and Reutebuch, S.E. 2005 Estimating forest canopy fuel parameters using LIDAR data. *Remote Sensing Environ.* **94**, 441–449.

Bater, W.C., Wulder, M.A., Coops, N.C., Nelson, R.F., Hilker, T. and Nasset, E. 2011 Stability of sample-Based scanning-LiDAR-derived vegetation metrics for forest monitoring. *IEEE Trans. Geosci. Remote Sensing* **49**, 2385–2392.

Batista, J.L.F., Couto, H.T.Z. and Silva Filho, D.F. 2014 Quantificação de Recursos Florestais: árvores, Arvoredos E Florestas. 1st edn. Oficina de Textos, 384 pp.

Breusch, T.S. and Pagan, A.R. 1979 A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287–1294.

Carvalho, S.P.C., Rodriguez, L.C.E., Silva, L.D., Carvalho, L.M.T., Calegario, N., Lima, M.P. *et al*. 2015 Predição do volume de árvores integrando LiDAR e geoestatística. *Sci. Forestalis* **43**, 627–637.

Coops, N.C., Hilker, T., Wulder, M.A., Newnham, G. and Trofymow, J.A. 2007 Estimating canopy structure of Douglas-fir forest stands from discrete-return LiDAR. *Trees* **21**, 295–310.

Crookston, N.L. and Finley, A. 2008 Yaimpute: an R package for k-NN imputation. *J. Stat. Software* **23**, 1–16. http://www.jstatsoft.org/v23/i10/paper. Package URL: http://cran.r-project.org/web/packages/yaImpute/index.html (accessed on 20 March, 2015).

Dubayah, R.O. and Drake, J.B. 2000 Lidar remote sensing for forestry. *J. For.* **98**, 44–46.

ECOAR. 2003 *Greenhouse Effect*. 1st edn. São Paulo, 5 pp.

Gonzalez-Ferreiro, E.G., Miranda, D., Barreiro-Fernández, B.L., Bujan, S., Garcia-Gutierrez, J.G. and Dieguez-Aranda, U.D. 2013 Modelling stand biomass fractions in Galician *Eucalyptus globulus* plantations by use of different LiDAR pulse densities. *For. Syst.* **22**, 510–525.

Gama, F.F., dos Santos, J.R. and Mura, J.C. 2010 *Eucalyptus* biomass and volume estimation using interferometric and polarimetric SAR data. *J. Appl. Remote Sensing* **2**, 939–956.

García, M., Riaño, D., Chuvieco, E. and Danson, F.M. 2010 Estimating biomass carbon stocks for a Mediterranean forest in central Spain using LiDAR height and intensity data. *Remote Sensing Environ.* **114**, 816–830.

Hirata, T., Furuya, N., Suzuki, M. and Yamamoto, H. 2009 Airborne laser scanning in forest management: individual tree identification and laser pulse penetration in a stand with different levels of thinning. *For. Ecol. Manage.* **258**, 752–760.

Hudak, A.T., Crookston, N.L., Evans, J.S., Falkowski, M.J., Smith, A.M.S. and Gessler, P.E. 2006 Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return LiDAR and multispectral satellite data. *Canadian J. Remote Sensing* **32**, 126–138.

Hudak, A.T., Evans, J.S. and Smith, A.M.S. 2009 Review: liDAR utility for natural resource managers. *Remote Sensing* **1**, 934–951.

Hudak, A.T., Strand, E.K., Vierling, L. a., Byrne, J.C., Eitel, J.U.H., Martinuzzi, S. and Falkowski, M.J. 2012 Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. *Remote Sensing Environ.* **123**, 25–40.

Hummel, S., Hudak, A.T., Uebler, E.H., Falkowski, M.J. and Megown, K.A. 2011 A comparison of accuracy and cost of LiDAR versus stand exam data for landscape management on the malheur national forest. *J. For.* **109**, 267–273.

Hurvich, C.M. and Tsai, C.L. 1989 Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Ibá. 2015 *Brazilian Tree Industry*. http://www.iba.org/images/shared/iba_2015.pdf (accessed on 10 November, 2015).

Jensen, J.R. 2005 *Introductory Digital Image Processing*. 3rd edn. Prentice Hall, 544 pp.

Köppen, W. and Geiger, R. 1928 Klimate der Erde. Gotha: Verlag Justus Perthes. Wall-map 150cm × 200cm.

Korpela, I., Ørka, H.O., Maltamo, M., Tokola, T. and Hyyppa, J. 2010 Tree species classification using airborne LiDAR- effects of stand and tree parameters, downsizing of training set, intensity normalization, and sensor type. *Silva Fenn.* **44**, 319–339.

Kraus, K. and Pfeifer, N. 1998 Determination of terrain models in wooded areas with airborne laser scanner data. *ISPRS J. Photogrammetry Remote Sensing* **53**, 3193–3203.

Lefsky, M.A., Cohen, W.B., Harding, D.J., Parker, G.G., Acker, S.A. and Gower, S.T. 2002 LiDAR remote sensing of above-ground biomass in three biomes. *Global Ecol. Biogeogr.* **11**, 393–399.

Li, Y., Andersen, H.-E. and McGaughey, R.J. 2008 A comparison of statistical methods for estimating forest biomass from light detection and ranging (LiDAR). *West. J. Appl. For.* **23**, 223–231.

Loetsch, F., Zohrer, F. and Haller, K.E. 1973 *Forest Inventory*. 2nd edn. Verlagsgesellschaft, 905 pp.

Macedo, R.D.C. 2009 Estimativa volumétrica de povoamento clonal de Eucalyptus spp. através de laser scanner aerotransportado. Master's thesis. National Institute for Space Research, 143 pp.

Manly, B.F.J. 2004 *Multivariate Statistical Methods: A Primer*. 3rd edn. Chapman and Hall, 208 pp.

McGaughey, R.J. 2014 FUSION/LDV: *Software for LiDAR Data Analysis and Visualization*. 3rd edn. USDA, Forest Service Pacific Northwest Research Station, 15 pp.

Mutlu, M., Popescu, S.C., Stripling, C. and Spencer, T. 2008 Mapping surface fuel models using LiDAR and multispectral data fusion for fire behavior. *Remote Sensing Environ.* **112**, 274–285.

Næsset, E. 1997 Determination of mean tree height of forest stands using airborne laser scanner data. *ISPRS J. Photogrammetry Remote Sensing* **52**, 49–56.

Næsset, E. 2002 Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing Environ.* **80**, 88–99.

Næsset, E. 2004 Estimation of above- and below-ground biomass in boreal forest ecosystems. *Int Arch Photogrammetry, Remote Sensing Spatial Inf Sci* **36**, 145–148.

Næsset, E. and Gobakken, T. 2008 Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. *Remote Sensing Environ.* **112**, 3079–3090.

Ørka, H.O., Næsset, E. and Bollandsa, O.M. 2009 Classifying species of individual trees by intensity and structure features derived from airborne laser scanner data. *Remote Sensing Environ.* **113**, 1163–1174.

Packalen, P., Maltamo, M. and Mehtatalo, L. 2011 ALS-based estimation of plot volume and site index in a *Eucalyptus* plantation with a nonlinear mixed-effect model that accounts for the clone effect. *Ann. For. Sci.* **68**, 1085–1092.

Pascual, C., Garcia, A., Cohen, W.B. and Martin-Fernandez, S. 2010 Relationship between LiDAR-derived forest canopy height and Landsat images. *Int. J. Remote Sensing* **31**, 1261–1280.

Patenaude, G., Hill, R.A., Milne, R., Gaveau, D.L.A., Briggs, B.B.J. and Dawson, T.P. 2004 Quantifying forest above ground carbon content using LiDAR remote sensing. *Remote Sensing Environ.* **93**, 368–380.

Ponzoni, F.J. and Gonçalves, J.L.M. 1999 Spectral features associated with nitrogen, phosphorous and potassium deficiencies in *Eucalyptus saligna* seedling leaves. *Int. J. Remote Sensing* **20**, 2249–2264.

R Development Core Team. 2015 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.Rproject.org. (accessed on 20 June, 2015).

Roberts, S.D., Dean, T.J., Evans, D.L., McCombs, J.W., Harrington, R.L. and Glass, P.A. 2005 Estimation individual tree leaf area in loblolly pine plantations using LiDAR derived measurements of height and crown dimensions. *For. Ecol. Manage.* **2013**, 54–70.

Robinson, A. 2015 *Equivalence: Provides Tests and Graphics for Assessing Tests of Equivalence, Version 0.7.0.* https://cran.r-project.org/web/packages/ equivalence/ (accessed on 20 June, 2006).

Rodriguez, L.C.E., Polizel, J.L., Ferraz, S.F.B., Zonete, M.F. and Ferreira, M.Z. 2010 Inventário florestal com tecnologia laser aerotransportada de plantios de Eucalyptus spp. no Brasil. *Ambiência* **6**, 67–80.

SFB. 2011 *Brazils* Forests at A Glance - 2010: *Data From 2005 to 2010*. Serviço Florestal Brasileiro, p. 124.

Shapiro, S.S. and Wilk, M.B. 1965 An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611.

Silva, C.A., Klauberg, C., Carvalho, S.P.C., Hudak, A. and Rodriguez, L.C.E. 2014 Mapping aboveground carbon stocks using LiDAR data in *eucalyptus* spp. plantations in the state of São Paulo, Brazil. *Sci. Forestalis* **42**, 591–604.

Spurr, S.H. 1952 *Forest Inventory*. Ronald Press, 476 pp.

Stephens, P., Watt, P., Loubser, D., Haywood, A. and Kimberley, M. 2007 Estimation of carbon stocks in New Zealand planted forests using airborne scanning LiDAR. *ISPRS J. Photogrammetry Remote Sensing* **36**, 389–394.

Stephens, P.R., Kimberley, M.O., Beets, P.N., Paul, T.S.H., Searles, N., Bell, A. *et al*. 2012 Airborne scanning LiDAR in a double sampling forest carbon inventory. *Remote Sensing Environ.* **117**, 348–357.

Tesfamichael, S.G. and Jan Van Aardt, F.A. 2010 Estimating plot-level tree height and volume of *Eucalyptus grandis* plantations using small-footprint, discrete return LiDAR data. *Prog. Phys. Geography* **34**, 515–540.

Vital, M.H.F. 2007 Impacto ambiental de florestas de eucalipto. *Revista do BNDES* **14**, 235–276.

Yu, C.M. 2004 Sequestro Florestal do Carbono no Brasil*: Dimensões Políticas Socioeconômicas E Ecológicas*. IEB, 283 pp.

Zandoná, D.F. 2006 Potencial do uso de dados laser scanner aerotransportado para estimativa de variáveis dendrométricas. Master's thesis, Department of Agricultural Sciences, Federal University of Paraná, 92 pp.

Zonete, M.F., Rodriguez, L.C.E. and Packalén, P. 2010 Estimação de parâmetros biométricos de plantios clonais de eucalipto no sul da bahia: uma aplicação da tecnologia laser aerotransportada. *Sci. Florestalis* **38**, 225–235.