

# A machine learning approach to model leaf area index in *Eucalyptus* plantations using high-resolution satellite imagery and airborne laser scanner data

Andrés Hirigoyen<sup>1</sup>✉, Cristina Acosta-Muñoz<sup>2</sup>, Antonio Jesús Ariza Salamanca<sup>2</sup>, María Ángeles Varo-Martínez<sup>2</sup>, Cecilia Rachid-Casnati<sup>1</sup>, Jorge Franco<sup>3</sup>, Rafael Navarro-Cerrillo<sup>2</sup>

Hirigoyen A., Acosta-Muñoz C., Ariza Salamanca A.J., Varo-Martínez M.Á., Rachid-Casnati C., Franco J., Navarro-Cerrillo R., 2021. A machine learning approach to model leaf area index in *Eucalyptus* plantations using high-resolution satellite imagery and airborne laser scanner data. Ann. For. Res. 64(2): 165-183.

**Abstract** As a forest structural parameter, leaf area index (LAI) is crucial for efficient intensive plantation management. Leaf area is responsible for the energy absorption needed for photosynthetic production and transpiration, both affecting growth. Currently, LAI can be estimated either by remote-sensing methods or ground-based methods. However, unlike ground-based methods, remote estimation provides a cost-effective and ecologically significant advance. The aim of our study was to evaluate whether machine learning algorithms can be used to quantify LAI, using either optical remote sensing or LiDAR metrics in *Eucalyptus dunnii* and *Eucalyptus grandis* stands. First, empirical relationships between LAI and remote-sensing data using LiDAR metrics and multispectral high-resolution satellite metrics, were assessed. Selected variables for LAI estimation were: forest canopy cover, laser penetration index, canopy relief ratio (from among the LiDAR data), the green normalized difference vegetation index, and normalized difference vegetation index (from among spectral vegetation indices). We compared the accuracy of three machine learning algorithms: artificial neural networks (ANN), random forest (RF) and support vector regression (SVR). The coefficient of determination ranged from 0.60, for ANN, to 0.84, for SVR. The SVR regression methods showed the best performance in terms of overall model accuracy and RMSE (0.60). The results show that the remote sensing data applied through machine learning algorithms provide an effective method to estimate LAI in eucalypt plantations. The methodology proposed is directly applicable for operational forest planning at the landscape level.

**Keywords:** LAI canopy, machine learning algorithms, intensive silviculture.

**Addresses:** <sup>1</sup>National Institute of Agricultural Research (INIA) Tacuarembó, Uruguay| <sup>2</sup>Department of Forestry Engineering, Laboratory of Silviculture, Dendrochronology and Climate Change. DendrodatLab-ERSAF, University of Córdoba, Córdoba, Spain| <sup>3</sup>Faculty of Agronomy, University of the Republic, Paysandú, Uruguay.

✉ **Corresponding Author:** Andrés Hirigoyen (ahirigoyen@inia.org.uy).

**Manuscript** received October 27, 2020; revised December 14, 2021; accepted December 21, 2021.

## Introduction

The structure of the forest canopy is generally described through variables, such as the leaf area index (LAI), crown diameter, and crown length (Weiss et al. 2004). LAI is defined as the one-sided leaf area per unit of horizontal ground surface area (Watson 1947). LAI plays a key role at describing the canopy, which is responsible for the energy absorption needed for photosynthetic production and transpiration affecting growth. Therefore, LAI is one of the main variables used to evaluate many processes related to the physiological status of trees. However, LAI estimation is complex, since it involves factors related to plant size and architecture, species composition, growth, and canopy phenology. Methods of LAI estimation can be categorized as direct or indirect (Jonckheere et al. 2004). Direct methods use destructive sampling to determine the total number of leaves per tree and their area, taking into account their angles and distribution. This approach is time-consuming, labor-intensive, and costly (Jonckheere et al. 2004). In the indirect methods, LAI is inferred from observations of another variable. Contact and non-contact methods are used, mostly to measure some radiative feature related to the distribution of light inside the canopy (Morsdorf et al. 2006). Indirect approaches are faster and amenable to automation, and thereby allow a larger spatial sample to be obtained (Jonckheere et al. 2004). Methods based on differences in spectral reflection between vegetation and other coverage, and using remote sensing data, have been widely applied for LAI determination in forests during recent years (Yan et al. 2019).

Spectral vegetation indices (VIs) obtained from remote sensing data are valuable for the estimation of LAI, due to their enhanced resistance to atmospheric effects. Several authors have found positive correlations between LAI and VIs derived from satellite images (Pavithra et al. 1998, Megown et al. 1999, le Maire et al. 2012). The VIs most

frequently used to estimate LAI are the normalized difference vegetation index (*NDVI*) and the simple ratio (*SR*), applied in coniferous forests, grasslands, and deciduous forests (Staben et al. 2018), however, these indices can saturate in dense vegetation such as forest stands. Many studies report that VIs saturate at high LAI (Zhao & Popescu 2009). VIs are easily affected by atmospheric conditions and soil, VIs tend to saturate, decreasing the accuracy of prediction of forest parameters (Luo et al. 2019) and leading to underestimation of LAI. The spectral reflectance of vegetation can be similar to the background reflectance, even with the use of hyperspectral remote sensing systems that boast high spectral (Tesfamichael et al. 2018).

Moreover, multispectral sensors do not have robust means of characterization of the vertically distributed forest attributes, since they provide information about plants using two dimensions (Tesfamichael et al. 2018). There are many satellite-derived LAI products, but their coarse-moderate spatial resolution (Modis 500 m, Landsat 30 m) has made it difficult to apply them at a regional scale (Zhou et al. 2020). This research is part of the new trends in precision forestry that is being used in intensive *Eucalyptus* plantations. This type of forestry requires continuous field data, high precision, at a reasonable cost, which can only be achieved by integrating high-resolution data processing from remote sensors. Planet Labs launched a constellation of CubeSats that cover the entire globe daily with a spatial resolution of 3 to 5 m (Planet Labs 2018). The use of CubeSat constellations that provide frequent and high-resolution images is one way to achieve both high review frequency and high spatial resolution (Zhou et al. 2020).

Active sensors employing light detection and range (LiDAR) technology can fill the gaps of multispectral sensors in LAI estimation by providing data on forest structural attributes, specifically canopy vertical profiles. Airborne Laser Scanner (ALS) is a LiDAR that operates

from aerial platforms, while the Terrestrial Laser Scanning (TLS) uses ground platforms. It can capture more detailed information on 3-D structure, and produces data points in a three-dimensional cloud, which permits the variation of the forest canopy features distributed vertically (Peduzzi et al. 2012). Unlike traditional optical data, ALS provide a direct measurement of the vertical structure of the forest canopy, ALS metrics has been used to predict stand variables such as volume or biomass (Jensen et al. 2008). Due to the ability to penetrate canopies and its rapid coverage of large areas, ALS has been used for LAI mapping at landscape or regional scales (Yan et al. 2019). ALS metrics (ALm) contain information on canopy structure and LAI; thus, the selection of ALm for LAI estimation modeling is a key procedure. ALS metrics provide information on the canopy structure, the correct selection of ALm that will integrate the LAI estimation model (Zhao & Popescu 2009). Different ALm such as ALS height percentiles, laser interception index (LII) and laser penetration index (LPI) have been used to predict LAI in previous research (Morsdorf et al. 2006, Peduzzi et al. 2012, Tseng et al. 2016).

The LPI is an approximation to the canopy gap fraction, it is expected to be useful for estimating the LAI. Combination ALm with hyperspectral and multispectral imagery can improve prediction accuracy of forest canopy parameters (Luo et al. 2019). Non-parametric machine learning (ML) approaches represent an alternative to discover unusual non-linear relationships, as well as operate high-dimensional and collinearity problems, handle non-linearity and non-normality of data (Lary et al. 2016). Since non-parametric statistical methods do not make a priori assumptions about the data, they are more suitable to establish and describing relationships between remote sensing images and the forest variables of interest (Ingram et al. 2005). A wide range of ML algorithms have been used effectively in forest applications for both regression and

classification of non-linear systems (Lary et al. 2016). ML including K-nearest neighbors, random forest (RF), tree-based models such as decision trees (DT), artificial neural networks (ANN), genetic algorithm, support vector machines (SVM), case-based reasoning (Lary et al. 2016, Jain et al. 2020).

LAI is a meaningful index for intensive plantation management, it provides information on current conditions and the possible evolution of stand growth and yield. Hence, it has been used to identify stand fertilization needs, thinning, or other management practices seeking to improve tree growth and yield (Peduzzi et al. 2012). Thus, several studies have applied spectral vegetation indices for the easy retrieval of LAI in *Eucalyptus* plantations (le Maire et al. 2012). However, few studies have integrated vegetation indices and ALS data for the estimation of LAI in *Eucalyptus* stands (but see Tesfamichael et al. 2018).

In this study, we compared several commonly used ML algorithms for integrating ALS metrics and vegetation indices derived from Planet imagery for predicting LAI across stands of *Eucalyptus grandis* Hill ex Maiden and *Eucalyptus dunnii* Maiden in Uruguay. Accordingly, the specific objectives were: i) to evaluate the efficiency of ALS and Planet spectral data as covariates to estimate LAI; ii) to test the performance of machine learning algorithms such as SVM, ANN, and RF in LAI prediction; iii) to analyze the spatial consistency among products and between products and field measurements; and iv) to assess whether Planet multispectral data can improve LAI estimates in conjunction with ALS.

## Materials and Methods

### Site description

The present research was established in five forest areas in Uruguay owned by Forestal Oriental SA, representing a total area of 2448 ha. We selected five zones (Figure 1): zones b1 and b2 (midpoint coordinates: 43°13'97"S

– 64°24'78"W) with 861 ha; zones b3 and b4 (midpoint coordinates: 63°07'44"S – 52°22'64"W) with 870.1 ha; and zone b5 (midpoint coordinates: 63°44'33.05"S – 63°95'58"W) with 716.9 ha. The study zones have mean annual rainfall between 1300 and 1400 mm (Castaño et al. 2011). Climate in the region is temperate subtropical, with a mean annual temperature of 18 °C (12 °C in the coldest month, 24 °C in the warmest month). Soils were classified as siliceous, fine-loamy, active, thermic Mollic Hapludalfs (Bentancor 2017). All zones are planted with *E. dunnii* and *E. grandis* (Table 1).

### Field measurement of LAI

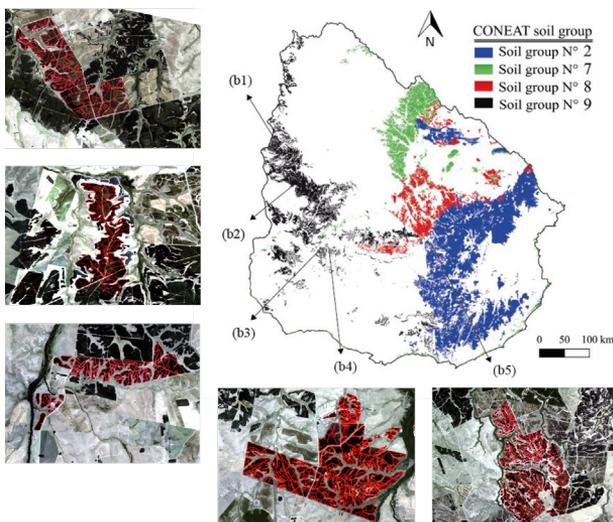
The LAI was measured in May-June 2018 in 39 plots of *E. dunnii* and 45 plots of *E. grandis*, each with a fixed radius of 10 m (314.16 m<sup>2</sup>), using a ceptometer (AccuPAR Model 80; Decagon Devices, Germany) (Table 1). The ceptometer was calibrated according to the instruction manual (Decagon Devices 2016). The measurements inside the plots were made by holding the instrument at a height of 1.3 m. For all plots, 6 sampling points were systematically located within the measurement area, 3 in the

tree rows and 3 between rows. LAI measurement points were systematically assigned along a transect perpendicular to the two central rows of trees in each plot. In each sampling point, 4 measurements were taken (north, south, east, west), 24 values of radiation in total. Between plots, radiation was measured on top of the canopy, in an area outside the plantation where the radiation reached the Ceptometer.

**Table 1** Silvicultural characteristics of *Eucalyptus grandis* and *Eucalyptus dunnii* in Uruguay.

	Stand attributes	min	mean	max	stdev
<i>E. grandis</i>	Dbh (cm)	10.65	15.49	19.33	1.81
	TH (m)	12.57	21.69	33.3	4.46
	LAI (m <sup>2</sup> m <sup>-2</sup> )	1.2	2.95	4.78	1.05
	Age (year)	4	6	9	1.4
	N (tree ha <sup>-1</sup> )	634	1041	1336	144
<i>E. dunnii</i> AB	Dbh (cm)	10.83	15.27	19.38	1.53
	TH (m)	9.52	19.55	29.22	4.23
	LAI (m <sup>2</sup> m <sup>-2</sup> )	1.17	3.26	5.07	1.07
	Age (year)	4	6	8	1.1
	N (tree ha <sup>-1</sup> )	669	1070	1432	159

Note: Standard deviation (stdev); minimum (min) and maximum (max) values for: stems density (N, trees ha<sup>-1</sup>); diameter at breast height (Dbh, cm); total height (TH, m); leaf area index (LAI, m<sup>2</sup> m<sup>-2</sup>), and Age (years).



**Figure 1** Location of the study sites (b1-b5) and Soils prioritized for forestry in Uruguay.

### Planet data and spectral vegetation indices

Planet is a company that operates a constellation of more than 175 small standardized CubeSat 3U nanosatellites. Planet provides daily nadir-pointing high-resolution land surface imaging of the entire Globe (Planet Labs 2018). Planet's satellites provide multispectral high-resolution satellite images (MS HR) at a spatial resolution of 3–4 m (nadir ground sampling distance) with red, green, and blue (RGB) and near-infrared (NIR) data. We used a total of 11 Planetscope Ortho Tile products (Planet Labs 2018). An orthomosaic of multiple scenes

orthorectified in a single strip was merged and divided into a defined grid (25 x 25 km). The images were acquired from February to April 2018 and underwent radiometric, geometric, and sensor correction (Planet Labs 2018). The Ortho Scene–Analytic (level 3B) have a rigorous geometric correction (Frazier & Hemingway 2021). The analytical format involved a 4-band,

16-bit multispectral image (RGB-NIR), having an orthorectified spatial resolution of 3.125, with daily review. Five spectral indices were calculated (Table 2). Image pre-processing and processing were performed with ENVI (Exelis 2015), and post-processing with QGIS (QGIS Development Team 2009) and R software (R Core Development Team 2013).

**Table 2** Vegetation indices used to model leaf area index (LAI) for *Eucalyptus* plantations in Uruguay.

Index	Formulation	Reference
Enhanced vegetation index	$EVI = 2.5 \frac{\rho_{NIR} - \rho_{Red}}{1 + \rho_{NIR} + 6\rho_{Red} - 7.5\rho_{Blue}}$	(Huete & Van Leeuwen 1999)
Green normalized difference vegetation index	$GNDI = \frac{\rho_{NIR} - \rho_{Green}}{\rho_{NIR} + \rho_{Green}}$	(Gitelson & Merzlyak 1998)
Normalized Difference Vegetation Index	$NDVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}}$	(Rouse Jr et al. 1973)
Simple Ratio	$SRI = \frac{\rho_{NIR}}{\rho_{Red}}$	(Jordan 1969)
Greenness Index	$GI = \frac{\rho_{Green}}{\rho_{Red}}$	(Xue & Su 2017)

### Acquisition and processing of ALS data

Airborne LiDAR acquisition was carried out during March 2018, using a Riegl VUX-1 laser scanner (Riegl, Germany) installed on a helicopter, with a pulse repetition rate of 550 kHz, an angular step width of 0.0687°, and a FOV of 55°, at a flight altitude of 110 m.a.s.l. The resulting point density was 20 pulses m<sup>-2</sup>, geo-referenced in the WGS84 UTM S21 coordinate system. For all echoes *x* and *y* (planimetric coordinates), plus ellipsoidal height values are calculated. The FUSION software (McGaughey 2013) was used to check and process ALS point clouds.

After the ALS data was verified and validated, a digital terrain model (DTM), a canopy height model, a digital surface model, and ALS metrics were generated. We used the *GroundFilter* function to identify and filter the returns that reached the ground. To create a DTM with a pixel

size of 1 m<sup>2</sup>, the *GridSurfaceCreate* function was used (González-Jaramillo et al. 2018). To normalize the heights, the *Clipdata* algorithm was applied to ensure that the *z* coordinate of each point corresponds to the height above the ground. *Cover* was used to calculate the canopy cover fraction (CCF), which represents the proportion of soil that is covered by the crown vertical projection and allows the density and horizontal distribution of the vegetation to be determined. Point data stored in *las* format were converted to ASCII text format by *LDA2ASCII*. The folders obtained from these procedures included the variables *x*, *y*, and *z*, and intensity values for each laser pulse return (Mesas-Carrascosa et al. 2012).

The ALS normalized point-cloud data pre-plots were filtered according to (Peduzzi et al. 2012) and the points were categorized into three groups, depending on the height above the ground (hag): vegetation returns (Nv),

ground returns (Ng) and all returns (Na). The ceptometer was held at approximately 1.3 m above the ground, so the vegetation

returns were classified using a 1.3-m threshold (Sumnall et al. 2016b). Table 3 shows the LiDAR metrics (ALm).

**Table 3** Height and intensity distribution metrics of ALS point cloud returns for each plot level area.

ALS metrics	Symbols
All returns (return hag > 0.2 m)	Elev.total, Elev.max, Elev.mean, Elev.median, Elev.sd, Elev.cv, Elev.sk, Elev.kt, Elev.P05 ... Elev.P99
Total number and first of ground returns	Gr.total, Gr.firs
Vegetation returns (return hag > 1.3 m)	Veg.total, Veg.max, Veg.mean, Veg.median, Veg.sd, Veg.cv, Veg.sk, Veg.kt., Veg.P5, ..., Veg.P95
Intensity value summary from vegetation returns (returns hag > 1.3 m)	Sd, Cv, Sk, Kt

Note: Hag= height above the ground. Subscripts were as follows: frequency (total), mean, median, minimum (min), maximum (max), standard deviation (sd), coefficient of variation (cv), skewness (sk) kurtosis (kt), and height percentiles (Elev.P10, Elev.P20,... Elev.P99).

To record the energy of the returns reflected back to the sensor, the intensity values are stored in ALS data. This recording is a function of several variables (such as angle of incidence, reflectivity and target area, laser power, atmospheric absorption, and distance between the sensor and the target). The intensity values need to be calibrated before their comparison; this process allows to compare different flights and scans (Mesas-Carrascosa et al. 2012). A simple method to correct the intensity proposed by Mesas-Carrascosa et al. (2012), consists of normalizing the range to a standard range established by the user, and the effect of the angle of incidence is corrected by applying the cosine.

$$I_{normalized} = I \frac{R^2}{(R_s^2 \cos\alpha)} \quad (1)$$

where  $I_{normalized}$  is the normalized intensity,  $I$  is the raw intensity value,  $R$  is the range (sensor-target distance),  $R_s$  is the standard range, and  $\alpha$  is the angle of incidence. As information about the flight lines was not available, we used an approximation to determine the  $R$  values, using the difference between the height of each point and the mean flight altitude (Mesas-Carrascosa et al. 2012). Due to the smooth relief of the study area, we assumed that the scan angle was equivalent to the angle of incidence (Mesas-Carrascosa et al. 2012).

### ALS pseudo-waveforms construction

The waveform vertical distribution curve contains more information than the waveform-derived metrics and the returns (Zhou & Qiu 2015). We used discrete-return data to create pseudo-waveforms (ALS-PWF) as the vertical frequency distribution of the sum of the intensity, as a function of height bins (Zhou & Qiu 2015). Waveforms can be uniform Gaussian-shaped functions, asymmetric, peaked, or flattened depending on the backscattering surface material (trees, grass, ground, roofs) (Zaletnyik et al. 2010). ALS-PWF data were constructed by modifying the method proposed by Luo et al. (2019). In each plot, ALS point clouds were separated by elevation values and assigned to height bins (from hag = 0.2 m at canopy height maximum) with intervals of 0.5 m. The relative returns intensity in each height bin were calculated by dividing the sum of the intensity in each height bin by the total intensity (Peduzzi et al. 2012). ALS-PWF data were obtained by connecting the intensity corresponding to the elevation of each height bin, through a smooth curve. Each ALS-PWF series was analyzed to identify the number of peaks and its shape as a density function. To distinguish between different ALS-PWF from backscattering surface material as trees, shapes statistical measures such as

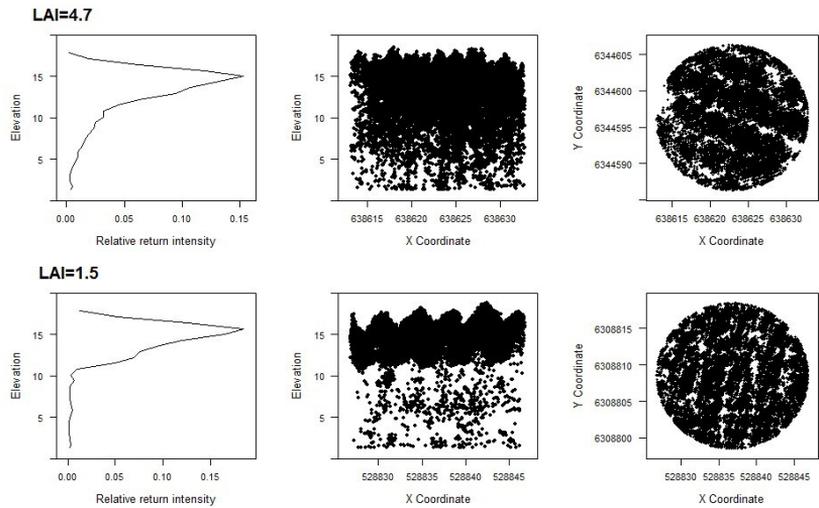
skewness, kurtosis, and standard deviations can be calculated and compared (Zhou & Qiu 2015). If the shape of a ALS-PWF is different from a normal distribution, the waveforms have to be separated according to the number of peaks, because skewness and kurtosis have meaning only for a normal (unimodal) distribution (Zaletnyik et al.

2010). The number of peaks in the ALS-PWF was determined with the local maximum point in a simple peak detection method by *findpeaks* function of the *pracma* package of R software. A process of Gaussian decomposition was not applied because we expected that after ALS-PWF creation, as described above, only a major peak corresponding to the tree crowns zone would be identified. The plots area did not have undergrowth vegetation, so peaks under the crown were not expected.

Figure 2 presents a visual analysis of the relationship between the shapes of the ALS-PWF and the plots from which they were obtained.

#### Laser penetration index

In LPI calculation, DTM and non-ground points are taken into account. The canopy penetration and corresponding LAI can be derived from the fraction of echoes located on and below the canopy (Solberg et al. 2006). LPI is based on the same principles as the instruments that measure LAI indirectly, measuring the solar light transmission or reflectance through vegetation (Peduzzi et al. 2012). The LPI can be related to the principles used by the IAF indirect measurement instruments; these apply a vertical stratification to



**Figure 2** *Eucalyptus grandis* plots with maximum LAI (top) and minimum LAI (bottom); ALS pseudo-waveform (left), sample cross section of a point cloud of a plot (middle) and sample of top view of canopy return of a plot (right).

determine the amount of returns above or below a height threshold (Sumnall et al. 2016b).

We calculated five LPI (Table 4) based on points data and three based on intensity information, as previously proposed (Peduzzi et al. 2012, Sumnall et al. 2016a, Tseng et al. 2016). In these works, LPI showed good correlation with the measured LAI.

#### Variable selection

To reduce data redundancy and improve overall interpretability, a multicollinearity analysis was performed. Spearman's correlation was used to identify highly correlated ALm variables. A collinearity analysis was then performed using the Variance Inflation Factor (VIF). Independent variables were selected and collinear variables were deleted, using a VIF >10 as the critical threshold (Ariza Salamanca et al. 2019). To select the best explanatory variables from among the ALm, LPI, and VI, several authors have employed RF techniques (Luo et al. 2019). The RF selection procedure uses a percentage increase in model mean square error (%MSE for regression) to select variables that explain the highest amount of variation.

**Table 4** Computation of ALS penetration indices (LPI).

LPI calculation	Proposed by
$LPI_1 = \frac{N_g}{N_g + N_v}$	(Solberg et al. 2006)
$LPI_2 = \frac{N_g}{N_g + N_a}$	(Peduzzi et al. 2012)
$LPI_3 = \frac{N_g + N_{sg}}{N_g + N_a}$	(Tseng et al. 2016)
$LPI_4 = \frac{N_v}{N_g + N_a}$	(Luo et al. 2019)
$LPI_5 = \frac{I_v}{I_{all}}$	(Hopkinson & Chasmer 2009)
$LPI_6 = \frac{I_g}{I_{all}}$	(Sumnall et al. 2016a)
$LPI_7 = \frac{I_v}{I_g}$	(Sumnall et al. 2016a)

Note: ( $N_g$ ) total numbers of ground points, ( $N_{sg}$ ) single-return ground points, ( $N_v$ ) numbers of vegetation points ( $N_g$ ), numbers of ground points, the sum of intensity of all returns ( $I_{all}$ ), sum of intensity for returns below and above the ground threshold ( $I_g, I_v$ )

This variable importance score is computed as the difference in predictions on the out-of-bag observations before and after permutation in the data withheld from each tree (Pearse et al. 2017). However, importance scores computed in this way have been shown to overestimate the importance of highly correlated predictors (Pearse et al. 2017).

In our study, we used permutation importance (PIMP) proposed by Altmann et al. (2010) for reducing bias, based on a permutation test, returning significant *p-values* for each feature.

PIMP is available in the *vita* package of R software (R Core Development Team 2013).

### Machine learning algorithms

Once the variables were selected, ML was used

as an empirical approach for determination of the regression between the LAI and predictors. ML methods require a training dataset and a random subset of the data for a completely independent validation. ML methods require for independent validation a random subset of the data, and a random training data set for fit. (Lary et al. 2016). We randomly divided the dataset into two sets for training and validation, with 70% and 30% of the data, respectively. We carried out a comparison of the most used regressions in ML approaches (SVM, ANN, and RF) in a common framework to estimate LAI.

### Support vector regression (SVR)

SVM can be used for classification tasks and to solve regression problems, where a continuous prediction output is expected (Mountrakis et al. 2011, Jain et al. 2020). SVR is based on the principle of structural risk minimization, and looks for a hyperplane with the greatest margin to divide samples into two classes with the largest interval (Jain et al. 2020). SVM operate by assuming that each set of input parameters will have a unique relationship to its response variable (Gleason & Im 2012). The grouping of these predictors and their relationships are sufficient to define rules that can be applied for predicting variables (Gleason & Im 2012). This method has proven its robustness to determine dimensionality and outliers in the training data, its generalization ability, and its ability to reduce overfitting (Mountrakis et al. 2011, Gleason & Im 2012). For SVR, a linear function was used as the kernel function because it showed better behavior than polynomial and sigmoidal kernel functions. To select cost and gamma parameters, we used the *svm* and *tune.svm* functions in R to find the optimal values (implemented with package *e1071*).

### Artificial neural networks

ANN are an alternative approach for modeling non-linear and complex phenomena in forest science. ANN are typically composed of an input layer, an output layer, and one or more hidden layers (Ingram et al. 2005). Our ANN structure

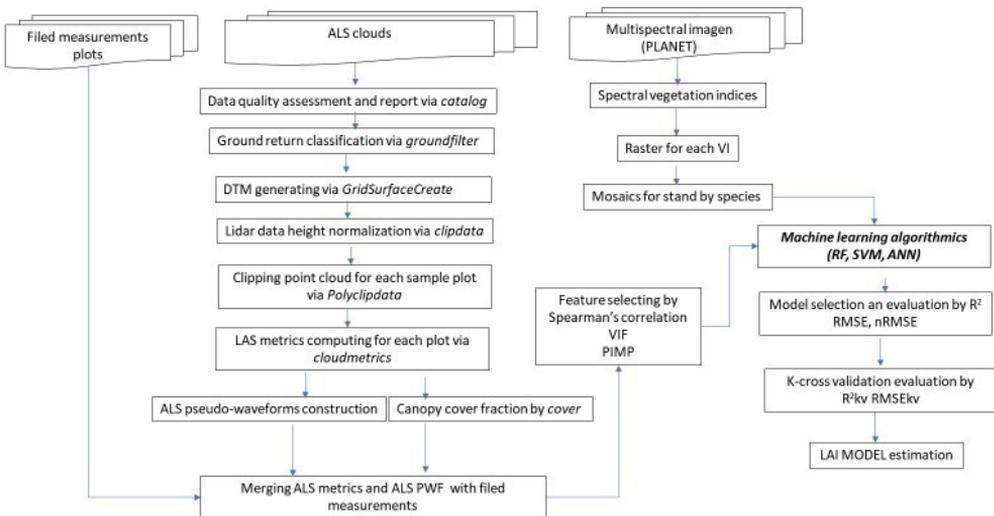
included an input layer, an output layer, and five hidden layers (a ranging from 1 to 15 neurons in the hidden layers were evaluated).

Random Forest

RF has been widely used to estimate dendrometric variables. This bagging-based method, creates and calculate the average of a large group of correlated trees. To determine how an input is related to a predicting variable, RF operated with a series of binary rule-based decisions based on classification and regression trees (Pearse et al. 2017). The samples for training each component tree were selected using bagging. RF regression output is determined by averaging the individual tree output (Jain et al. 2020). The RF algorithm has some parameters that must be configured before training the model; the number of predictors taken into account at each bifurcation of the tree (*mtry*) and the number of random trees assembled during the construction of the model (*n tree*).

Model assessment and validation

The set of observations for each group was randomly separated into training (70%) and testing (30%) sets. In order to determine the goodness of fit, the determination coefficient ( $R^2$ ) of the linear regression between the observed and predicted values was used during the model development stages (Fassnacht et al. 2014). Model performance was assessed by the root-mean-square error (RMSE) and the RMSE normalized by the mean observed values of the dependent variables (nRMSE, %). RMSE indicates the absolute value of the error, whereas nRMSE represents the relative value of the error with respect to the average plot-level. We compare the models through the  $R^2$ , RMSE, and nRMSE. According to Fassnacht et al. (2014), when describing the model performance it is necessary to report at least the RMSE and the correlation between the predictions and observations.



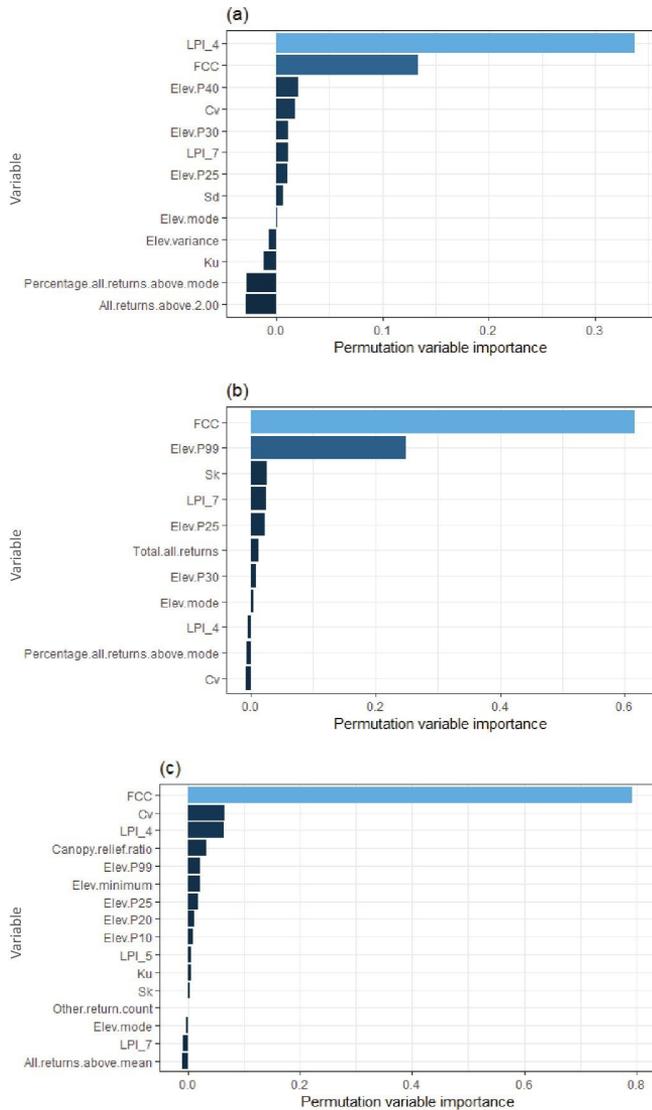
**Figure 3** Flowchart describing the methodological steps used to prepare and process data and model variables and to evaluate the results. Where: root Mean Squared Error (RMSE), normalized root mean squared error (nRMSE), determination coefficient ( $R^2$ ), root-mean-squared error of the k-cross-validation (RMSEkv), determination coefficient of the k-cross-validation ( $R^2kv$ ), Artificial neuronal networks (ANN), Support vector machine (SVM), Random forest (RF) approaches, Variance Inflation Factor (VIF) and permutation importance (PIMP).

To achieve accurate and reasonable comparisons of different LAI estimates, all LAI values were modeled using the same variables. Statistical analyses were carried out separately for ALS metrics with ALS-PWF, and for ALS metrics with ALS-PWF and vegetation indices together; first for each species individually (Group *E. dunnii* and Group *E. grandis*), then for both species together (Group *Eucalyptus*).

Due to the limited field data available, we used k-fold cross-validation (k-cv) with a 20-fold approach to validate the models. A model was fitted using 19 of the 20 subsets (training dataset) and then it was used to predict the remaining subset (validation dataset). This process was repeated until all subsets had been used once. For k-fold cross-validation root-mean-squared (RMSE<sub>kv</sub>), determination coefficient (R<sup>2</sup><sub>kv</sub>) was calculated. The comparison of the results was expressed in terms of RMSE (m<sup>2</sup> m<sup>-2</sup>), nRMSE (%), and R<sup>2</sup>, plus the coefficients obtained in a process of k-cv RMSE<sub>kv</sub> (m<sup>2</sup> m<sup>-2</sup>), R<sup>2</sup><sub>kv</sub> for each method. Higher R<sup>2</sup> values and lower RMSE and nRMSE values indicate higher model precision and accuracy.

The present study employed several data sets and required the development of remote sensing metrics and indices, as well as various data analysis procedures. A flowchart describing these steps and the relationships between each stage is provided in Figure 3.

We performed all analyses using R software (R Core Development Team 2013). The R packages used for SVM were: *e1071* and *kernelab*; for ANN: *nnet*; and for RF: *randomForest* and *caret*.



**Figure 4** Conditional inference trees-derived importance ranking of the ALSm of: (a) *Eucalyptus grandis* (b) *Eucalyptus dunnii* and (c) both species. Variables are arranged in descending order of mean importance.

## Results

### Prediction of LAI using ALS and ALS-PWF data

After performing multicollinearity analysis and Spearman’s correlation test, the remaining ALS metrics were used as input in the PIMP function. Variable importance was calculated using the corrected RF model, based on the PIMP scores of the features. In most cases, it was superior in accuracy to the conditional inference trees implemented in previous approaches. To select suitable tuning parameters for RF before it was employed in the PIMP function, the number of predictors trialed in each split (parameter *mtry*) varied from 3 up to the maximum number of predictor variables. With *mtry* = 5, minimum MSE and stability of importance scores were achieved. Then, we varied the number of trees (*ntree*) to assess the stability of the results. A *ntree* value of 120 produced the lowest out-of-bag-error for the models and RF results. *ALS<sub>m</sub>* with a higher increase in the permutation variance import was the best variable for LAI estimation.

The four best ranked variables for estimation of the LAI of *E. grandis* were Laser penetration 4 (LPI 4), Canopy relief ratio (CCF), Percentile 40 (Elev.P40) , and coefficient of variation of ALS-PWF (Cv) (Figure 4 a); for *E. dunnii*, best ranked variables were CCF, Percentile 99 (Elev.P99), skewness of ALS-PWF (Sk), and laser penetration 7 (LPI 7) (Figure 4 b). When considering both species together, the four best ranked variables were CCF, Cv, laser penetration 4 (LPI 4), and the CCF (Figure 4 c).

SVR was the model with the highest R<sup>2</sup> (78%) and lowest nRMSE values for estimation of LAI in *E. dunnii*, followed by RF and ANN (Table 5). For *E. grandis*, SVM achieved the highest median overall accuracy (79%) and smaller RMSE and nRMSE values, followed by RF and ANN. Due to the

low precision values of the *Eucalyptus* group (R<sup>2</sup> < 40%), each predictor variable was evaluated individually. The canopy relief ratio was eliminated because it did not improve the results. So, SVR achieved the highest overall accuracy (84%) and smaller RMSE and nRMSE values for LAI estimation in the *Eucalyptus* group, followed by RF and ANN (Table 5). The SVM and RF models explained more than 72% of the variability in LAI and ANN more than 60%.

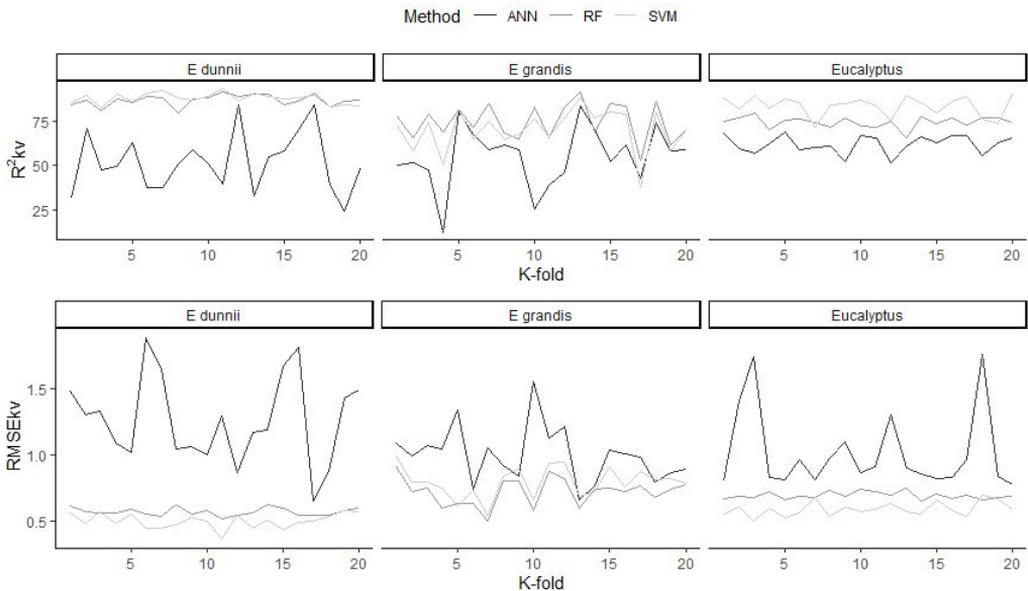
By changing the training dataset, we were able to evaluate the accuracy for each ML algorithm for a mean of 20 k-fold and for each k-fold. The k-cv values in the same k-executions were recorded for the three methods, to obtain a comparable trend (Figure 5).

ANN had the highest variability by species. ANN gave acceptable results (R<sup>2</sup> > 60%) in a few estimations and exhibited higher variation in R<sup>2</sup>kv and RMSEkv terms between k-fold. These fluctuations of prediction accuracy and absolute value of error in ANN were more pronounced than for SVR and RF (these models were more stable). Also, within the stability of the SVM and RF methods, the k-fold variations were more stable in *E. dunnii* than in the other groups.

**Table 5** Statistical criteria for the estimation of LAI in *Eucalyptus dunnii*, *Eucalyptus grandis* and *Eucalyptus* (sp.) in Uruguay.

	Model	RMSE	nRMSE	R <sup>2</sup>	RMSEkv	R <sup>2</sup> kv
<i>E. dunnii</i>	SVR	<b>0.60</b>	<b>0.19</b>	<b>0.78</b>	<b>0.49</b>	<b>0.88</b>
	ANN	1.07	0.33	0.69	1.26	0.51
	RF	0.65	0.20	0.76	0.56	0.86
<i>E. grandis</i>	SVR	<b>0.79</b>	<b>0.27</b>	<b>0.79</b>	<b>0.71</b>	<b>0.75</b>
	ANN	1.83	0.62	0.60	0.99	0.55
	RF	0.96	0.33	0.72	0.79	0.70
<i>Eucalyptus</i>	SVR	<b>0.66</b>	<b>0.21</b>	<b>0.84</b>	<b>0.58</b>	<b>0.83</b>
	ANN	0.91	0.29	0.70	1.11	0.62
	RF	0.71	0.23	0.77	0.76	0.70

Note: Root Mean Squared Error (RMSE), normalized Root Mean Squared Error (nRMSE), determination coefficient (R<sup>2</sup>), root-mean-squared error of the k- cross-validation (RMSEkv), determination coefficient of the k-cross-validation (R<sup>2</sup>kv).



**Figure 5** Trend of root-mean-square error (RMSEkv - bottom) and determination coefficient ( $R^2_{kv}$  - top) of k-cross validation for *Eucalyptus dunnii*, *Eucalyptus grandis*, and both species (*Eucalyptus*). ANN Artificial neuronal networks (black line), SVM Support vector machine (soft grey line), RF Random forest (grey line) approaches.

For a single k-folder, variability of RMSEkv was observed (e.g. *E. dunnii* in  $k=6$  had ANN=1.89, RF=0.55, and SVR=0.44; *E. grandis* in  $k=10$  had ANN=1.55, SVR=0.65, and RF=0.58; *Eucalyptus* in  $k=3$  had ANN=1.7, SVR=0.77, and RF=0.67). Also, *E. dunnii* in  $k=19$  had ANN=0.24, SVR=0.84, and RF=0.86; *E. grandis* in  $k=4$  had ANN=0.11, SVR=0.50, and R=0.69; and *Eucalyptus* in  $k=12$  had ANN=0.51, SVR=0.65, and RF=0.74.

The standard deviation of the  $R^2$  and the RMSE of validation were calculated to represent the model stability averaged over k-iterations. Values close to 0 indicate higher model stability. The SVR method exhibited the lowest standard deviation of both the RMSE (0.056, 0.140, and 0.055 for *E. dunnii*, *E. grandis*, and *Eucalyptus*, respectively) and the validation coefficient of determination (0.031, 0.105, and 0.055 for *E. dunnii*, *E. grandis*, and *Eucalyptus*, respectively). In all

cases, the  $R^2_{kv}$  and RMSEkv indicators during validation confirmed the testing results. The SVR methods showed high precision in LAI estimation, in terms of model median overall accuracy and nRMSE, for the three groups, while ANN was the approach with the lowest precision, accuracy, and stability.

### Prediction of LAI using combined MS HR vegetation indices and ALS metrics

With the models selected in the previous section, the accuracy gain obtained by adding vegetation indices was evaluated. Due to multicollinearity between indices, each vegetation index was evaluated independently as a predictor variable. Table 6 shows the RMSE, nRMSE,  $R^2$ , RMSEkv, and  $R^2_{kv}$  for the models and VI. The best models for LAI were chosen based on their highest coefficient of determination ( $R^2$ ) and lowest RMSE, and were corroborated using k-cross validation.

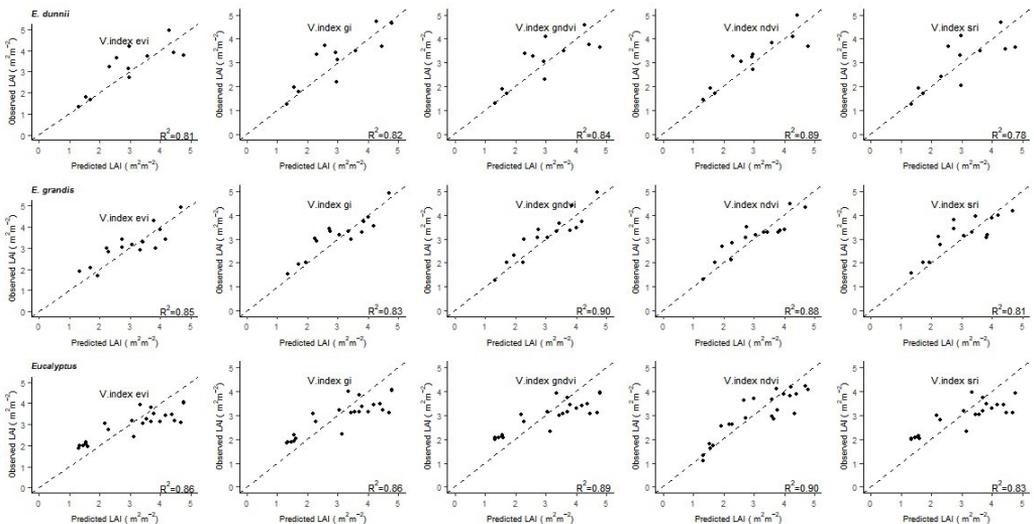
**Table 6** Statistical comparison between models using different vegetation indexes (VI) to estimate LAI in *Eucalyptus dunnii*, *Eucalyptus grandis* and *Eucalyptus* (sp.) in Uruguay by Support vector regression (SVR).

	Model	VI	RMSE	nRMSE	RMSEkv	R <sup>2</sup> kv
<i>E. dunnii</i>	SVR	EVI	0.70	0.23	0.61	0.80
		GI	0.66	0.22	0.64	0.79
		<b>GNDVI</b>	0.56	0.20	0.62	0.83
		<b>NDVI</b>	0.53	0.16	0.60	0.85
		SRI	0.69	0.21	0.68	0.73
<i>E. grandis</i>	SVR	EVI	0.50	0.17	0.68	0.79
		GI	0.60	0.20	0.68	0.79
		<b>GNDVI</b>	0.40	0.13	0.57	0.82
		<b>NDVI</b>	0.43	0.14	0.59	0.80
		SRI	0.69	0.23	0.64	0.76
<i>Eucalyptus</i>	SVR	EVI	0.75	0.20	0.64	0.83
		GI	0.74	0.20	0.63	0.82
		GNDVI	0.78	0.20	0.61	0.82
		<b>NDVI</b>	0.50	0.16	0.50	0.88
		SRI	0.78	0.19	0.61	0.83

Note: Root Mean Squared Error (RMSE, m<sup>2</sup> m<sup>-2</sup>), normalized root mean squared error (nRMSE, %), root-mean-squared error of the k-cross-validation (RMSEkv, m<sup>2</sup> m<sup>-2</sup>) and determination coefficient of the k-cross-validation (R<sup>2</sup>kv). Values in bold show the best performing model per vegetation indices.

The inclusion of *SRI* gave smaller R<sup>2</sup> and R<sup>2</sup>kv values in the estimation of LAI than the other VIs, whereas *NDVI* achieved the highest overall accuracy (89% in the fit and 85% in the validation) and smaller RMSE, nRMSE, and RMSEkv values (Figure 6). For *E. grandis*, *GNDVI* was incorporated as a predictor variable, exhibiting lower RMSE, nRMSE and RMSEkv values and higher R<sup>2</sup> and R<sup>2</sup>kv values than the other VI. For both species together, the accuracy was higher than 90% when the best VI was incorporated (*NDVI*).

The indices with the best performances were the normalized infrared differences, either with red or green. Again, spectrally, *E. grandis* behaved in a slightly different way from *E. dunnii*, although in the joint modeling (*Eucalyptus* group) it had no weight, as happened with LPI 4 and LPI 7.



**Figure 6** Observed versus predicted LAI values from models using different Vegetation indexes (Vindex) for *Eucalyptus dunnii*, *Eucalyptus grandis* and *Eucalyptus* (sp.) by Support vector regression (SVR).

Regarding the model performance assessed in terms of nRMSE, the model of LAI estimation for *E. grandis* that included *GNDVI* showed the lowest relative value error (0.13%) with respect to the average for all cases (Groups and indices). The relative error values were highest for *E. dunnii*. For *E. grandis*, graphical analysis revealed an underestimation behavior for all models (Figure 6).

Overall, the LAI prediction was best for all Groups using *ALSm+ALS-PWF* and *VI*, and the LAI predictions were acceptable. Nevertheless, *SRI* performed relatively poorly in LAI estimation and yielded the lowest RMSE, nRMSE, and  $R^2$  values for the *E. dunnii* and *E. grandis* groups. According to these results, CCF, Elev.P99, Sk, LPI 7, and *NDVI* (for *E. dunnii*), LPI 4, CCF, Elev.P40, Cv, and *GNDVI* (for *E. grandis*), and CCF, Cv, LPI 4, and *NDVI* (for *Eucalyptus*) were the most suitable indices combination for LAI estimation because they showed higher accuracy and precision than the other indices.

## Discussion

Leaf area index is a key physiological variable for understanding forest dynamics and growth, and as such it has been widely used in forest models. To improve LAI estimation in intensive *Eucalyptus* plantations, we have considered an integrated two-step method, including ALS and high spatial resolution vegetation indices, based on ML algorithms. We have demonstrated the ability of this approach to estimate LAI with high accuracy and stability, which is consistent with previous research (Yuan et al. 2017). The temporal gap between the acquisition of ALS, MS HR, and field data increases the uncertainty in LAI estimates (Zhao & Popescu 2009). In our study, the temporal difference was less than 2 months, avoiding uncertainty.

### LAI values and ALS variable selection

The *Eucalyptus* LAI observed values varied between 1.17 and 5.07  $m^2 m^{-2}$ , with mean 178

values of 2.95 and 3.26  $m^2 m^{-2}$  for *E. grandis* and *E. dunnii*, respectively. Those values are within the range reported by Scurlock et al. (2001), who reported LAI value for *Eucalyptus* species of 3.51  $m^2 m^{-2}$ . Observed values also agree with Whitehead & Beadle (2004), who reported a value of 3.19  $m^2 m^{-2}$ . However, in Uruguay Alonso (2011) determined a mean LAI value of 1.82  $m^2 m^{-2}$  in *Eucalyptus globulus* stands. This difference could be explained by differences in the species (*E. grandis* vs *E. globulus*), age (6-year-old plantations in our study vs over 10 years old for *E. globulus*), and estimation method (ceptometer vs fish-eye photography), among other factors.

When ranking the importance of *ALSm* and *ALS-PWF* variables in LAI estimation, the LPI and a shape parameter were always ranked at or near the top, showing the importance of the height distribution metric. The CCF metric was the most important *ALSm* in LAI estimation, and it was included in all the LAI models selected. The CCF corresponded to the upper canopy cover, where we expected the greatest density of foliage. The number of *ALSm* selected for each model was a balance between parsimony and relevance, determined by the permutation variable importance. The difference in the branching between the two species meant that LAI models for *E. dunnii* selected middle height percentiles (Elev.P40) while in *E. grandis* they included high height percentiles (Elev.P99).

The LPI is related to the canopy development; the closer and denser the vegetation, the fewer the laser pulses that penetrate to reach the ground (Peduzzi et al. 2012). Thus, LPI has been used to predict LAI, and models including this variable were able to explain 80% or more of LAI variation (Peduzzi et al. 2012). The penetration metric selected for *E. grandis* and *Eucalyptus* was a metric of return heights (LPI 4), while for *E. dunnii* the most important penetration metric focused on intensity (LPI 7). This may be related to the characteristics of the two species; the spectral response of

the leaves in *E. dunnii* was more sensitive to infrared radiation, related to a healthier status of the plantations. *Canopy relief ratio* was deleted for the *Eucalyptus* group, likely because the biological information provided by this variable was already included through the CCF.

### LAI models

In our study, an SVM approach was chosen to predict the LAI in *E. dunnii*, *E. grandis*, and *Eucalyptus*. Ability to handle non-linearity and to quantify the importance of independent variables makes SVR an effective algorithm. SVR has been implemented in a wide range of analyses with ALS and MS HR in agricultural and forest investigations (Mountrakis et al. 2011). However, in our study ANN showed poor LAI prediction, which contrasts with other studies where ANN proved more efficient than RF and SVR in predicting forest structural parameters (increment in diameter, height, dbh) (Tavares Júnior et al. 2020).

The best models, with the highest  $R^2$  ( $>0.78$ ) and the lowest nRMSE ( $0.137 \text{ m}^2 \text{ m}^{-2}$ ), included four (*E. grandis* and *E. dunnii*) or three (*Eucalyptus*) ALSm. These results are consistent with previous studies. Assessing *E. grandis*, (Tsfamichael et al. 2018) achieved  $R^2=0.67$  with two LiDAR metrics (height and standard deviation of points below the 10th height) and obtained  $R^2=0.83$  when employing 14 metrics (with collinearity problems). Peduzzi et al. (2012) obtained models with  $R^2$  values between 0.61 (two model inputs) and 0.83 (six model inputs). Jensen et al. (2008) reported estimated LAI  $R^2$  values of 0.86 (RMSE = 0.76) and 0.69 (RMSE = 0.61). Morsdorf et al. (2006) estimated returned  $R^2$  values of 0.69 (RMSE = 0.01), but it should be noted that the LAI values in their study were within a narrower range (0–2) than ours.

We expected that ALS height metrics would be present in all the LAI models, because increases in canopy height should occur in tandem with increases in LAI, and this was observed for *E. dunnii* (Elev.99) and *E. grandis*

(Elev.P40). In this study, LAI was sensitive to the inclusion of LPI, and LPI 7 and LPI 4 related the vegetation returns with the total returns and the intensity of the returns with the intensity of the ground returns. Sumnall et al. (2016a) reported an  $R^2$  value of 0.89 when employing LPII in LAI estimation. We computed the vegetation return and intensity above a threshold of 1.3 m, at the same height as the LAI data (ceptometer measurements, dbh height). By implementing this minimum height threshold for the within-plot ALS, the returns and the corresponding metrics were limited to the vertical space in which the greatest amount of foliage was distributed. In addition, the link between field LAI measurement height and the threshold of what is defined as ‘ground’ and ‘canopy’ within LiDAR returns, as suggested in Sumnall et al. (2016b), implies that estimation of the vertical distribution of LAI through the canopy is possible.

### Prediction of LAI using combined ALS and MS HR variables

The integration of Planet-derived VIs markedly improved the LAI estimation with respect to the results obtained by ALSm+PWF. The integration of ALSm, ALS-PWF metrics, and VIs involved four different approaches. The use of high spatial resolution, multispectral VIs can provide continuous and detailed spectral signatures of forest biophysical attributes (Shen et al. 2018). The ANN, RF, and SVR models contained a stochastic element that resulted in a different LAI model each time they were applied to the same training data. With the addition of VIs, the improvement in the accuracy of the predictive models was significant. In general, the NIR-based VIs (i.e., *NDVI* and *GNDVI*) performed better in terms of LAI assessment than the other indices evaluated, and *NDVI* for *E. dunnii*, *GNDVI* for *E. grandis*, and *NDVI* for *Eucalyptus* were the best-fitted parameters, the improvement in nRMSE being 16%, 51%, and 39%, respectively. The gain in explained variance ( $R^2$ ) was 14 % (*E. dunnii* and *E.*

*grandis*) and 7 % (*Eucalyptus*) when using these VIs. The  $R^2_{kv}$  and  $RMSE_{kv}$  obtained from 20-fold cross-validation also agreed well. For a *Eucalyptus* clone (*grandis* x *urophylla*) in Brazil,  $R^2$  values between 0.68 and 0.77 were reported when LAI was estimated with *NDVI* models, using Landsat (De Almeida et al. 2015) and RapidEye sensors (Alvares et al. 2015).

The plots of the observed against the estimated values for the VIs (Figure 5) demonstrate that the *NDVI* showed less data dispersion with respect to the (1:1) trend line and gave lesser trends of overestimation or underestimation of LAI for the *Eucalyptus* group. *NDVI* is the most widely used vegetation index, and it is based on the red and near-infrared wavelength bands (Xue & Su 2017). On the other hand, *GNDVI* uses the green band instead of the red and can be more sensitive to chlorophyll; thus, it may be more accurate than *NDVI* over areas with dense vegetation cover (Gitelson & Merzlyak 1998). In Rio Grande do Sul, De Godoy Goergen et al. (2016) using Landsat 5, found that *GNDVI*, *NDVI*, and *SRI* were correlated with the age of *E. dunnii* trees and that *GNDVI* could be used to differentiate *E. dunnii* from *E. urograndis*.

Our results demonstrate that the combination of the ALS<sub>m</sub> and PWF datasets with VIs was suitable for the estimation of LAI. Together, these variables yielded the highest prediction accuracies in the k-cross-validation, with  $R^2_{kv} > 0.85$  and  $RMSE_{kv} < 0.6$  ( $m^2 m^{-2}$ ). The standard deviation of  $RMSE_{cv}$  was small for SVR, ranging from  $0.05 m^2 m^{-2}$  (*Eucalyptus*) to  $1.23 m^2 m^{-2}$  (*E. dunnii*). The ANN models produced substantially higher  $R^2_{kv}$  and  $RMSE_{kv}$  standard deviations than SVR and RF. When both species were treated as a single group (*Eucalyptus*), an increase in the accuracy ( $R^2_{kv}=0.88$ ), precision ( $RMSE_{kv}=0.5 m^2 m^{-2}$ ), and stability (standard deviations of  $R^2_{kv} < 0.05$ ) was observed.

Previous studies demonstrated that the integration of these two complementary

datasets can improve the prediction of forest structural parameters (Shen et al. 2018). Although multispectral data have limitations regarding the quantification of the vertical structure of forests, multispectral metrics such as VIs have the capability to improve the accuracy of the prediction of forest structural parameters. This indicates that multispectral metrics have great potential in the prediction of forest structural parameters.

There are some potential limitations in our study. First, the temporal gap in the acquisition of the ALS, MS HR, and field data may increase the uncertainty in the LAI estimates (Zhao & Popescu 2009). However, in our study, the temporal difference was less than 2 months, thus minimizing the uncertainty. Second, there was no spatial mismatch between the field measurements (plots of  $314 m^2$ ) and the CubeSat pixels (3 m resolution). This may have resulted in a homogeneous pixel covering one stand with homogeneous growth conditions. Differences among stands would have been included in the different pixels. The pixel effect can cause LAI underestimation at coarse spatial resolution because of area aggregation that produces averaged values.

The results allowed us to suggest a new approach to LAI estimation in intensive *Eucalyptus* plantations that can be applied over distinct forest test sites, selecting and applying the best features of each source of data to improve LAI estimation.

## Conclusion

In this study, we applied a modeling approach to estimate LAI in *Eucalyptus* plantations, integrating several remote sensing data sources. The SVM and RF models were more accurate than the ANN model. The SVM model, using ALS metrics, ALS-PWF shape parameters, and Vis from MS HR, performed well for *E. grandis*, *E. dunnii*, and both species together. The most informative ALS metrics for LAI estimation were related to the upper and middle tree height percentiles and the

canopy cover fraction. The most important pseudo-waveforms parameters shapes were the coefficient of variation (Cv, for *E. grandis* and *Eucalyptus*) and skewness (Sk, for *E. dunnii*). The *NDVI* and *GNDVI* from the Planet data were proven to be highly appropriate for the assessment and prediction of LAI in *Eucalyptus* plantations. The selected models, incorporating ALSm, ALS-PWF, and multispectral data, have great potential for LAI values. This study provides an improved foundation for LAI estimation in intensive *Eucalyptus* plantations, which is crucial for forest management and research.

## Acknowledgments

The authors thank the Instituto Nacional de Investigaciones Agropecuarias (INIA-Uruguay) for supporting our research work and for help during the fieldwork. We are particularly grateful to Roberto Scoz, Demian Gomez, Leonidas Carrasco and Alicia Peduzzi for their assistance during this research. RMNC acknowledge the institutional support of the Ministerio de Ciencia, Innovación y Universidades (Spain), through the ESPECTRAMED (CGL2017-86161-R), show significant changes.

## References

- Alonso J., 2011. Modelación de procesos hidrológicos asociados a la forestación con *Eucalyptus* en el Uruguay. Facultad de Ingeniería, Universidad de la República.
- Altmann A., Toloşi L., Sander O., Lengauer T., 2010. Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10): 1340–1347. <https://doi.org/10.1093/bioinformatics/btq134>
- Alvares C.A., Mattos E.M. de, Campoe O.C., Marrichi A.H.C., Stape J.L., 2015. Uso de sensoriamento remoto na estimativa do índice de área foliar em *Eucalyptus*. XVII Simpósio Brasileiro de Sensoriamento Remoto, pp. 6429–6436.
- Ariza Salamanca A.J., Navarro-Cerrillo R.M., Bonet-García F.J., Pérez-Palazón M.J., Polo M.J., 2019. Integration of a Landsat time-series of NBR and hydrological modeling to assess *Pinus pinaster* Aiton. Forest Defoliation in South-Eastern Spain. *Remote Sensing*, 11(19): 2291. <https://doi.org/10.3390/rs11192291>
- Castaño J.P., Giménez A., Ceroni M., Furest J., Aunchayna R., Bidegain M., 2011. Caracterización agroclimática del Uruguay 1980-2009. Serie Técnica INIA, 193: 33.
- De Almeida A.Q., Ribeiro A., Delgado R.C., Rody Y.P., De Oliveira A.S., Leite F.P., 2015. Índice de área foliar de eucalyptus estimado por índices de vegetação utilizando imagens TM - landsat 5. *Foresta e Ambiente*, 22: 368–376. <https://doi.org/10.1590/2179-8087.103414>
- De Godoy Goergen L.C., De Vargas Kilca R., Da Silva Narvaes I., Silva M.N., Silva E.A., Pereira R.S., Adami M., 2016. Distinção de espécies de eucalipto de diferentes idades por meio de imagens TM/Landsat 5. *Pesquisa Agropecuária Brasileira*, 51: 53–60. <https://doi.org/10.1590/S0100-204X2016000100007>
- Fassnacht F.E., Hartig F., Latifi H., Berger C., Hernández J., Corvalán P., Koch B., 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, 154: 102–114. <https://doi.org/10.1016/j.rse.2014.07.028>
- Frazier A.E., Hemingway B.L., 2021. A technical review of planet smallsat data: Practical considerations for processing and using planetscope imagery. *Remote Sensing*, 13(19): 3930. <https://doi.org/10.3390/rs13193930>
- Gitelson A.A., Merzlyak M.N., 1997. Remote estimation of chlorophyll content in higher plant leaves. *International Journal of Remote Sensing*, 18: 2691–2697.
- Gitelson A.A., Merzlyak M.N., 1998. Remote sensing of chlorophyll concentration in higher plant leaves. *Advances in Space Research*, 22(5): 689–692. [https://doi.org/10.1016/S0273-1177\(97\)01133-2](https://doi.org/10.1016/S0273-1177(97)01133-2)
- Gleason C.J., Im J., 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment*, 125: 80–91. <https://doi.org/10.1016/j.rse.2012.07.006>
- Hopkinson C., Chasmer L., 2009. Testing LiDAR models of fractional cover across multiple forest ecozones. *Remote Sensing of Environment*, 113(1): 275–288. <https://doi.org/10.1016/j.rse.2008.09.012>
- Huete A., Van Leeuwen W., 1999. MODIS vegetation index (MOD13). Algorithm Theoretical Basis Document, 3(213), 295-309.
- Ingram J.C., Dawson T.P., Whittaker R.J., 2005. Mapping tropical forest structure in southeastern Madagascar using remote sensing and artificial neural networks. *Remote Sensing of Environment*, 94(4): 491–507. <https://doi.org/10.1016/j.rse.2004.12.001>
- Jain P., Coogan S.C.P., Subramanian S.G., Crowley M., Taylor S., Flannigan M.D., 2020. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4): 478-505. <https://doi.org/10.1139/er-2020-0019>
- Jensen J.L.R., Humes K.S., Vierling L.A., Hudak A.T., 2008. Discrete return lidar-based prediction of leaf area index in two conifer forests. *Remote Sensing of Environment*, 112(10): 3947–3957. <https://doi.org/10.1016/j.rse.2008.07.001>
- Jonckheere I., Fleck S., Nackaerts K., Muys B., Coppin

- P., Weiss M., Baret F., 2004. Review of methods for in situ leaf area index determination Part I. Theories, sensors and hemispherical photography. *Agricultural and Forest Meteorology*, 121(1-2): 19–35. <https://doi.org/10.1016/j.agrformet.2003.08.027>
- Jordan C.F., 1969. Derivation of leaf-area index from quality of light on the forest floor. *Ecology*, 50(4): 663–666.
- Lary D.J., Alavi A.H., Gandomi A.H., Walker A.L., 2016. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1): 3–10. <https://doi.org/10.1016/j.gsf.2015.07.003>
- Le Maire G., Marsden C., Nouvellon Y., Stape J.L., Ponzoni F.J., 2012. Calibration of a species-specific spectral vegetation index for leaf area index (LAI) monitoring: Example with MODIS reflectance time-series on eucalyptus Plantations. *Remote Sensing*, 4(12): 3766–3780. <https://doi.org/10.3390/rs4123766>
- Luo S., Wang C., Xi X., Nie S., Fan X., Chen H., Yang X., Peng D., Lin Y., Zhou G., 2019. Combining hyperspectral imagery and LiDAR pseudo-waveform for predicting crop LAI, canopy height and above-ground biomass. *Ecological Indicators*, 102: 801–812. <https://doi.org/10.1016/j.ecolind.2019.03.011>
- Megown R.A., Webster M., Jacobs S., 1999. Using Landsat TM imagery to estimate LAI in a *Eucalyptus* plantation. 1–13.
- Mesas-Carrascosa F.J., Castillejo-González I.L., De la Orden M.S., Porras A.G.F., 2012. Combining LiDAR intensity with aerial camera data to discriminate agricultural land uses. *Computers and Electronics in Agriculture*, 84: 36–46. <https://doi.org/10.1016/j.compag.2012.02.020>
- Morsdorf F., Kötz B., Meier E., Itten K.I., Allgöwer B., 2006. Estimation of LAI and fractional cover from small footprint airborne laser scanning data based on gap fraction. *Remote Sensing of Environment*, 104(1): 50–61. <https://doi.org/10.1016/j.rse.2006.04.019>
- Mountrakis G., Im J., Ogole C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3): 247–259. <https://doi.org/10.1016/j.isprsjprs.2010.11.001>
- Pavithra B., Kalaivani K., Ulagapriya K., 1998. Remote sensing techniques for mangrove mapping. *International Journal of Engineering and Advanced Technology*, 8: 27–30.
- Pearse G.D., Morgenroth J., Watt M.S., Dash J.P., 2017. Optimising prediction of forest leaf area index from discrete airborne lidar. *Remote Sensing of Environment*, 200: 220–239. <https://doi.org/10.1016/j.rse.2017.08.002>
- Peduzzi A., Wynne R.H., Fox T.R., Nelson R.F., Thomas V.A., 2012. Estimating leaf area index in intensively managed pine plantations using airborne laser scanner data. *Forest Ecology and Management*, 270: 54–65. <https://doi.org/10.1016/j.foreco.2011.12.048>
- Planet Labs, 2018. Precision Ag insights from frequent imaging smarter farming throughout the season. 22.
- R Core Development Team, 2013. A language and environment for statistical computing. 1.
- Rouse jr. J.W., Haas R.H., Schell J.A., Deering D.W., 1973. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation.
- Scurlock J.M.O., Asner G.P., Gower S.T., 2001. Worldwide historical estimates of leaf area index, 1932–2000. *ORNL/TM-2001/268*, 34. <https://doi.org/ORNL/TM-2001/268>
- Shen X., Cao L., Chen D., Sun Y., Wang G., Ruan H., 2018. Prediction of forest structural parameters using airborne full-waveform LiDAR and hyperspectral data in subtropical forests. *Remote Sensing*, 10(11): 1729. <https://doi.org/10.3390/rs10111729>
- Solberg S., Næsset E., Hanssen K.H., Christiansen E., 2006. Mapping defoliation during a severe insect attack on Scots pine using airborne laser scanning. *Remote Sensing of Environment*, 102(3-4): 364–376. <https://doi.org/10.1016/j.rse.2006.03.001>
- Staben G., Lucieer A., Scarth P., 2018. Modelling LiDAR derived tree canopy height from Landsat TM, ETM+ and OLI satellite imagery — A machine learning approach. *International Journal of Applied Earth Observation and Geoinformation*, 73: 666–681. <https://doi.org/10.1016/j.jag.2018.08.013>
- Sumnall M., Peduzzi A., Fox T.R., Wynne R.H., Thomas V.A., Cook B., 2016a. Assessing the transferability of statistical predictive models for leaf area index between two airborne discrete return LiDAR sensor designs within multiple intensively managed Loblolly pine forest locations in the south-eastern USA. *Remote Sensing of Environment*, 176: 308–319. <https://doi.org/10.1016/j.rse.2016.02.012>
- Sumnall M.J., Fox T.R., Wynne R.H., Blinn C., Thomas V.A., 2016b. Estimating leaf area index at multiple heights within the understorey component of Loblolly pine forests from airborne discrete-return lidar. *International Journal of Remote Sensing*, 37: 78–99. <https://doi.org/10.1080/01431161.2015.1117683>
- Tavares Júnior I da S., Torres C.M.M.E., Leite H.G., Castro N.L.M. de, Soares C.P.B., Castro R.V.O., Farias A.A., 2020. Machine learning: Modeling increment in diameter of individual trees on Atlantic Forest fragments. *Ecological Indicators*, 117: 106685. <https://doi.org/10.1016/j.ecolind.2020.106685>
- Tesfamichael S.G., van Aardt J., Roberts W., Ahmed F., 2018. Retrieval of narrow-range LAI of at multiple lidar point densities: Application on *Eucalyptus grandis* plantation. *International Journal of Applied Earth Observation and Geoinformation*, 70: 93–104. <https://doi.org/10.1016/j.jag.2018.04.014>
- Tseng Y.H., Lin L.P., Wang C.K., 2016. Mapping CHM and LAI for heterogeneous forests using airborne full-waveform LiDAR data. *Terrestrial, Atmospheric and Oceanic Sciences*, 27: 537–548. [https://doi.org/10.3319/TAO.2016.01.29.04\(ISRS\)](https://doi.org/10.3319/TAO.2016.01.29.04(ISRS))
- Watson D.J., 1947. Comparative physiological studies on the growth of field crops: I. Variation in net assimilation rate and leaf area between species and varieties, and

- within and between years. *Annals of Botany*, 11: 41–76. <https://doi.org/10.1093/oxfordjournals.aob.a083148>
- Weiss M., Baret F., Smith G.J., Jonckheere I., Coppin P., 2004. Review of methods for in situ leaf area index (LAI) determination Part II. Estimation of LAI, errors and sampling. *Agricultural and Forest Meteorology*, 121: 37–53. <https://doi.org/10.1016/j.agrformet.2003.08.001>
- Whitehead D., Beadle C.L., 2004. Physiological regulation of productivity and water use in *Eucalyptus*: A review. *Forest Ecology and Management*, 193: 113–140. <https://doi.org/10.1016/j.foreco.2004.01.026>
- Xue J., Su B., 2017. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, 2017. <https://doi.org/10.1155/2017/1353691>
- Yan G., Hu R., Luo J., Weiss M., Jiang H., Mu X., Xie D., Zhang W., 2019. Review of indirect optical measurements of leaf area index: Recent advances, challenges, and perspectives. *Agricultural and Forest Meteorology*, 265: 390–411. <https://doi.org/10.1016/j.agrformet.2018.11.033>
- Yuan H., Yang G., Li C., Wang Y., Liu J., Yu H., Feng H., Xu B., Zhao X., Yang X., 2017. Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: Analysis of RF, ANN, and SVM regression models. *Remote Sensing*, 9. <https://doi.org/10.3390/rs9040309>
- Zaletnyik P., Laky S., Toth C., 2010. LIDAR waveform classification using self-organizing map. *American Society for Photogrammetry and Remote Sensing Annual Conference 2010: Opportunities for Emerging Geospatial Technologies*, 2: 1055–1066.
- Zhao K., Popescu S., 2009. Lidar-based mapping of leaf area index and its use for validating GLOBCARBON satellite LAI product in a temperate forest of the southern USA. *Remote Sensing of Environment*, 113: 1628–1645. <https://doi.org/10.1016/J.RSE.2009.03.006>
- Zhou H., Wang C., Zhang G., Xue H., Wang J., Wan H., 2020. Generating a spatio-temporal complete 30 m leaf area index from field and remote sensing data. *Remote Sensing*, 12: 2394. <https://doi.org/10.3390/rs12152394>
- Zhou Y., Qiu F., 2015. Fusion of high spatial resolution WorldView-2 imagery and LiDAR pseudo-waveform for object-based image analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 101: 221–232. <https://doi.org/10.1016/j.isprsjprs.2014.12.013>