

VANESSA SOUSA DA SILVA

**LiDAR-DERIVED METHODS FOR VOLUME ESTIMATION AND INDIVIDUAL
TREE DETECTION IN *Eucalyptus* spp. PLANTATIONS**

RECIFE
Pernambuco – Brasil
Fevereiro – 2019

VANESSA SOUSA DA SILVA

LiDAR-DERIVED METHODS FOR VOLUME ESTIMATION AND INDIVIDUAL TREE
DETECTION IN *Eucalyptus* spp. PLANTATIONS

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Florestais da Universidade Federal Rural de Pernambuco, como parte das exigências para obtenção do título de Mestre em Ciências Florestais, Área de Concentração: Biometria e Manejo Florestal.

Orientador: Prof. Dr. Emanuel Araújo Silva

Co-orientador: Dr. Carlos Alberto Silva

Co-orientadora: Dr^a. Gabrielle Hambrecht Loureiro

RECIFE
Pernambuco – Brasil
Fevereiro – 2019

VANESSA SOUSA DA SILVA

**LIDAR-DERIVED METHODS FOR VOLUME ESTIMATION AND INDIVIDUAL
TREE DETECTION IN *Eucalyptus* spp. PLANTATIONS**

APROVADO EM: 26/02/2019

Banca Examinadora:



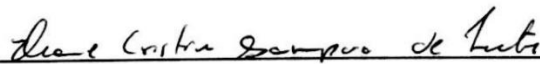
Prof. Dr. Emanuel Araújo Silva

(Orientador – Departamento de Ciências Florestais/UFRPE)



Prof. Dr. Eric Bastos Gorgens

(Membro Titular – Departamento de Engenharia Florestal/ UFVJM)



Prof.ª Dr.ª Eliane C. Sampaio de Freitas

(Membro Titular – Departamento de Ciências Florestais/UFRPE)

RECIFE – PE

FEVEREIRO – 2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

S586l Vanessa Sousa da Silva
LiDAR-Derived methods for volume estimation and individual
tree detection in *Eucalyptus* spp. plantations / Vanessa Sousa da Silva.
– 2019.
84 f. : il.

Orientador: Emanuel Araújo Silva.

Coorientadores: Carlos Alberto Silva e Gabrielle Hambrecht
Loureiro.

Dissertação (Mestrado) – Universidade Federal Rural de
Pernambuco, Programa de Pós-Graduação em Ciências Florestais,
Recife, BR-PE, 2019.

Inclui referências.

1. LiDAR 2. Aprendizado do computador
3. Levantamentos florestais 4. Sensoriamento Remoto I. Silva,
Emanuel Araújo, orient. II. Silva, Carlos Alberto, coorient.
III. Loureiro, Gabrielle Hambrecht, coorient. IV. Título

CDD 634.9

AGRADECIMENTOS

À minha família, em especial minha mãe Helenice, pelo amor e dedicação incondicional de sempre. Ao meu irmão Victor agradeço pelos ensinamentos e incentivo na busca de aperfeiçoamento pessoal e intelectual. À João Tanajura (minha família em Recife), por ser mais do que um amigo, um verdadeiro irmão.

Ao meu orientador, Prof. Dr. Emanuel Araújo Silva, a quem agradeço pela oportunidade, confiança, amizade, paciência e todo suporte. Ao meu co-orientador Dr. Carlos Alberto Silva agradeço pela inspiração profissional, por todo o auxílio fundamental na concepção e desenvolvimento da dissertação, elaboração de scripts e análise de dados. Aos membros de todas as bancas examinadoras pelas valiosas contribuições. Meu muito obrigada.

À Universidade Federal Rural de Pernambuco, ao Programa de Pos-Graduação em Ciências Florestais, pela estrutura física e de corpo docente, componentes essenciais para minha formação acadêmica. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pela bolsa de estudos concedida.

À empresa Suzano S.A. pelo apoio na realização desse trabalho e cedimento dos dados, especialmente nas pessoas da Dr. Gabrielle Hambrecht Loureiro como co-orientadora e Pedro Resende Silva.

À University of Idaho e ao College of Natural Resources na pessoa do Prof. Dr. Arjan Meddens pelo suporte no desenvolvimento desta pesquisa e auxílio no processamento dos dados.

Aos pesquisadores Franciel Rex, Mikey Mid Mohan e Elisiane Alba por todo auxílio mais que essenciais na elaboração de scripts, análise de dados e escrita. Aos meus amigos e colegas de Pós-Graduação, em especial à todos do LBMF e LASER, principalmente à Djailson Costa, Gabriela Salami, Anderson Silva e Mayara Pessoa pela troca de conhecimentos, companheirismo e momentos de descontração durante esse período.

À Isis pela luz de amor e alegria em todos os dias da minha vida. Obrigada pela sua amizade, incentivo e zelo, todo o meu amor e gratidão. À Paulo, Felipe, Tawany, Joe, Victor, Laura, Rafaela, Cláudia, Marcella, e Rosa por todo o companheirismo e carinho. À todas as pessoas que contribuíram de alguma forma para a conclusão de mais esta etapa.

Muito Obrigada!

SILVA, VANESSA SOUSA DA. LiDAR-derived methods for volume estimation and individual tree detection in *Eucalyptus* spp. plantations. 2019. Advisor: Emanuel Araújo Silva. Co-advisors: Carlos Alberto Silva and Gabrielle Hambrech Loureiro.

GENERAL ABSTRACT

Accurate and spatially explicit measurements of forest attributes are considered of great importance for sustainable forest management and environmental protection. Improvements in the management of eucalyptus plantations result in multiple industrial and environmental benefits. Remote sensing techniques can increase planting management efficiency by reducing or replacing field sampling that requires a longer time and therefore higher costs. Airborne Light Detection and Ranging (LiDAR) systems have become an important remote sensing technique for forest inventory, mainly because this technology can provide high accuracy and spatially detailed information on forest attributes across entire landscapes. Remote sensing data from LiDAR combined with machine learning techniques and automated individual tree detection algorithms present great potential for modeling forest attributes. This dissertation is focused on the comparison of predictive methods of total stem volume and number of individual trees in plantations of *Eucalyptus* spp. from LiDAR-derived data. More specifically evaluating: 1- the combined impact of sample size and parametric and non-parametric modeling techniques; 2- the accuracy of algorithms for automatic individual trees detection. The modeling technique that presented the best performance was verified for the OLS method, which was able to provide results comparable to the traditional approaches of forest inventory using only 40% of the total field plots, followed by the Random Forest (RF) algorithm for the same sample size. The Dalponte e Silva automatic detection algorithms presented more accurate results with the lowest commission and omission errors, and consequently better F-scores in most of the sampled plots, obtaining comparable results.

Key words: LiDAR, Machine Learning, Forest Inventory, Remote Sensing.

SILVA, VANESSA SOUSA DA. Métodos derivados de LiDAR para estimativa de volume e detecção individual de árvores em plantios de *Eucalyptus* spp. 2019. Orientador: Emanuel Araújo Silva. Co-orientadores: Carlos Alberto Silva e Gabrielle Hambrecht Loureiro.

RESUMO GERAL

Medições acuradas e espacialmente explícitas de atributos florestais são consideradas de suma importância para o manejo florestal sustentável e a proteção ambiental. Melhorias no manejo de plantios de eucalipto resultam em múltiplos benefícios industriais e ambientais. As técnicas de sensoriamento remoto podem aumentar a eficiência do gerenciamento de plantios, reduzindo ou substituindo a amostragem de campo que demanda um maior tempo e conseqüentemente maiores custos. Os sistemas LiDAR (Airborne Light Detection and Ranging) tornaram-se uma importante técnica de sensoriamento remoto para o inventário florestal, principalmente porque essa tecnologia pode fornecer informações de alta precisão e espacialmente detalhadas sobre os atributos da floresta em paisagens inteiras. Dados de sensores remotos LiDAR combinados com técnicas de aprendizado de máquina e algoritmos automatizados de detecção de árvores individuais apresentam grande potencial para modelagem e delineamento de atributos florestais em larga escala. Esta dissertação está focada na comparação de métodos preditivos de volume total e de número de árvores individuais em plantios de *Eucalyptus* spp. à partir de dados derivado de sensor LiDAR. Mais especificamente avaliando: 1- o impacto combinado do tamanho da amostra e técnicas de modelagem paramétricas e não-paramétricas; 2- a acurácia de algoritmos de detecção automática de árvores individuais. A técnica de modelagem que apresentou o melhor desempenho foi verificado para o método OLS, que foi capaz de fornecer resultados comparáveis às abordagens tradicionais de inventário florestal usando apenas 40% do total de parcelas de campo, seguido pelo algoritmo Random Forest (RF) para o mesmo tamanho de amostra. Os algoritmos de detecção automática Dalponte e Silva apresentaram resultados mais precisos com os menores erros de comissão e omissão, e conseqüentemente melhores F-scores na maioria das parcelas amostradas, obtendo resultados comparáveis.

Palavras-chave: LiDAR, Aprendizagem de máquinas, Inventário Florestal, Sensoriamento Remoto.

TABLE OF CONTENTS

1. GENERAL INTRODUCTION	1
2. LITERATURE REVIEW	4
2.1 The <i>Eucalyptus</i> Genus.....	4
2.2 Forest growth and production modeling.....	5
2.3 Remote Sensing applied to the study of vegetation.....	6
2.4 LiDAR.....	7
2.5 Individual Trees Detection.....	9
2.6 Machine Learning.....	10
3. REFERENCES	13

CHAPTER I

COMBINED IMPACT OF SAMPLE SIZE AND MODELING TECHNIQUES FOR PREDICTING VOLUME IN *Eucalyptus* spp. PLANTATIONS FROM LiDAR DATA

ABSTRACT	20
RESUMO	21
1. INTRODUCTION	22
2. MATERIALS AND METHODS	25
3. RESULTS	33
4. DISCUSSION	40
5. CONCLUSIONS	45
6. REFERENCES	46

CHAPTER II

COMPARISON OF ALGORITHMS FOR INDIVIDUAL TREE DETECTION IN *Eucalyptus* spp. PLANTATIONS FROM LiDAR DATA

ABSTRACT	51
RESUMO	52
1. INTRODUCTION	53
2. MATERIALS AND METHODS	54
3. RESULTS	58
4. DISCUSSION	64
5. CONCLUSIONS	66
6. REFERENCES	68
GENERAL CONCLUSION AND RECOMMENDATIONS	70

LIST OF FIGURES

CHAPTER I

Figure 1. Location map of study area and plots.....	25
Figure 2. Principal Component Analysis.....	34
Figure 3. Boxplots of modeling methods performance measures.....	36
Figure 4. Predictive Maps.....	39

CHAPTER II

Figure 1. Location map of study area and plots.....	55
Figure 2. Detection of trees on plots.....	61

LIST OF TABLES

CHAPTER I

Table 1. Summary statistics of forest attributes.....	27
Table 2. Airborne LiDAR survey specifications	28
Table 3. LiDAR-derived structure metrics	28
Table 4. Pearson correlations among selected LiDAR metrics	33
Table 5. Loadings and eigenvectors for the first five PCs.....	33
Table 6. Modeling methods performance measures (PM) averages and standard deviations.....	38

CHAPTER II

Table 1. Airborne LiDAR survey specifications	56
Table 2. Results of tree detection by the different methods.....	59
Table 3. Analysis of the tree detection quality by the different methods.....	60
Table 4. Summary results of the tree detection with all methods tested.....	64

LIST OF ABBREVIATIONS AND ACRONYMS

PCA Principal component analysis

ALS Airbourne Laser Scanner

ANN Artificial Neural Network

BIAS Bias

CFI Continuous Forest Inventory

CHM Canopy Height Model

DBH Diameter at breast height

DSM Digital Surface Model

DTM Digital Terrain Model

DGPS Differential Global Positioning System

GPS Global Position System

IMU Inertial Measurement Unit

k-NN *k*-nearest neighbors

LASER Light Amplification by Stimulated Emission of Radiation

LiDAR Light Detection and Ranging

ML Machine Learning

OLS Ordinary least-squares

R² Coefficient of determination

RF Random Forest

RMSE Root Mean Square Error

SVM Support Vector Machine

1. GENERAL INTRODUCTION

Approximately 54% of the Brazilian territory corresponds to an area with forest cover, of which about 60% is composed of natural forests and 40% from planted forests. Currently most of the timber production comes from forests planted for industrial purposes, whose area in 2017 represented 7.84 million hectares (IBÁ, 2018). The cultivation of the genus *Eucalyptus* spp., for the production mainly of cellulose, wood panels, charcoal and firewood, occupied 5.7 million hectares of the area of planted trees in the country; they are mostly located in the states of Minas Gerais (24%), São Paulo (17%), and Mato Grosso do Sul (15%) (IBÁ, 2018).

The growth of the forest sector in Brazil makes the short, medium- and long-term planning of great importance, requiring its optimization to ensure the flow of wood over time. In this scenario, it is evident that the forest manager needs tools to generate accurate information about the current forest inventory and, consequently, provide the correct modeling of future productivity (AVARENGA, 2012). Methodologies for the evaluation of forest potential have been improved by researchers and companies, with the aim of minimizing errors, reducing costs and processing time in obtaining data.

The determination of the total or commercial volume of wood remains one of the most important variables for the diagnosis of the potential of a forest. The volumetric estimation is a requirement for the adequate management of a forest population, being the most used variable in the management, industry and commercialization (SANQUETTA; BALNINOT, 2004). The evaluation of the growth and production of a given species in a given site depends directly on variables such as the productive capacity of the site, the age of the standstill, as well as the dendrometric variables (CAMPOS; LEITE, 2013).

Dendrometric variables are measurements obtained in trees, generally seeking to improve knowledge about the structure of the stands in which they are inserted. The most broadly used variables are: basal area and height. The basal area is calculated from the diameter at breast height (DBH) (SCOLFORO; MELLO, 2006). Knowing the basal area and the height of the trees, it is possible to estimate the volume of the individuals. Volumetric equations can be constructed and adjusted from other dendrometric variables, but an analysis is necessary to evaluate the importance of these variables in the proposed model (KÖHL et al., 2006).

One of the main collection method of dendrometric parameters in Brazil is through a Continuous Forest Inventory (CFI), which is based on the installation of permanent plots, where

data such as the DBH and total height are periodically measured (SCOLFORO; MELO, 2006). However, it is a method considered laborious, demanding a lot of time and financial resources. In conventional forest inventories, errors are still included, such as bias in the measurement of diameters and heights, error in the manipulation of data, error in the measurement of plot area, which reflect in the imprecision of growth models and production, directly impacting the accuracy of volumetric estimates (OLIVEIRA et al., 2014).

In order to facilitate the acquisition of more accurate dendrometric data, reducing the participation of the inventory in the costs of forest production, several remote sensing techniques have been studied to estimate the characteristics traditionally obtained through the CFI (ANDRADE, 2013). In the past decade, advances in remote sensing have provided new tools, techniques, and technologies to support forest management. The use of remote sensing for forest inventory today can be considered of extreme importance, because from the use of sensor images and their processing, it is possible to obtain more detailed information that allows to adjust mathematical models that express the relation of the variables of remote sensing with the variables of conventional inventories (OLIVEIRA, 2011).

It is in this context that the concept of precision forestry, defined as the use of geospatial information tools in order to enable repeatable measurements, actions and processes to manage and harvest forest stands, supporting economic, environmental and sustainable decisions, is inserted. Through the precision forestry management, it is possible to combine geotechnologies with the conventional forest inventory, making possible the integration of previous forest planning and logging operations through a remote system (RIBEIRO, 2002).

Remote sensing technologies have been widely utilized to characterize forest structure at both local and global scales. For instance, in the past two decades, LiDAR (light detection and ranging) remote sensing has emerged as a technology well-suited to providing accurate estimates of forest attributes including height, volume, basal area and biomass both in natural and industrial plantation forest ecosystems (HUDAK et al., 2006; LEFSKY et al., 2002; SILVA et al., 2014). However, even though LiDAR can quickly provide forest attributes across extensive landscapes, it is still mostly used for research purposes, mainly due to the high cost of data acquisition and lack of optimized and accessible tools and methods for processing and modeling LiDAR data for forestry applications. Moreover, accurate prediction of forest attributes from LiDAR is highly dependent on methods.

A relevant aspect in the development of the estimates refers to the method used to construct the predictive models. Regression is the standard model adjustment tool for the various forest measurement tasks. However, in the last few years an important statistical development has been taking advantage of the current computational power, as well as making use of the various information available in large databases, which is not possible with most models based on traditional regression, due to their low flexibility and rigidity (MONTAÑO, 2016). The use of machine learning techniques, a subdivision of artificial intelligence, that uses sophisticated architecture algorithms such as Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Networks Neural Network - ANN), have demonstrated great capacity for constructing more complex models, such as multivariate nonlinear regression and nonparametric regression (LARY et al., 2016).

In view of the above, the search of methodologies that provide an improvement in the data acquisition for vegetation studies is a necessity, in order to establish alternatives that guarantee results satisfactorily similar to those obtained by traditional methods, reducing the time and quantity of resources. This is particularly true in developing nations such as Brazil, where applications of lidar are in the early stages. The airborne LiDAR system proves to be a promising tool for surveying qualitative-quantitative metrics of forest stands, thus justifying this type of study as a viability strategy for the remote forest inventory, in order to allow improvements in the production planning processes in Eucalyptus stands. The research presented in this dissertation is therefore focused upon further development of strategies to promote sustainable management of industrial eucalyptus plantation forests. Specifically, the main goal of this dissertation is to assess novel methods for forest inventory attributes prediction and mapping at individual tree and plot levels from lidar remote sensing data.

2. LITERATURE REVIEW

2.1. The *Eucalyptus* Genus

About 5.7 million hectares in Brazil are covered by eucalyptus. The planted forests in the country are basically destined to the production of pulp and paper, firewood and charcoal, reconstituted wood panels, laminate floors and solid wood products (IBÁ, 2018). The cultivation of *Eucalyptus* genus on an economic scale in Brazil occurred in 1904, with the introduction of 144 species to meet the demands of wood for railway roads construction (VALVERDE, 2007). As of 1965, with the law of fiscal incentives for reforestation, the area of Eucalyptus planted in Brazil increased from 500 thousand to three million hectares (TRUGILHO, et al., 2001).

The *Eucalyptus* genus belongs to the Myrtaceae family and is originated from the islands of Oceania. The species belonging to this genus are considered as fast growing and are suitable for management in coppice, allowing the regrowth to be driven for more than two rotations. The genus has more than 500 identified species that adapts and grows satisfactorily in the most diverse regions of the world in different environmental conditions (HASELEIN et al., 2005; IBÁ, 2018; VALVERDE, 2007). In tropical regions such as Brazil, the cutting cycles of Eucalyptus plantations range from 5 to 7 years (GUEDES et al., 2015; SCOLFORO et al., 2016). As a result of its characteristics, eucalyptus is one of the most planted trees in the world, as it is a species of easy adaptation to the most diverse climatic and soil conditions, besides the many forms of wood utilization by the industry (MARTINS et al., 2006; SILVA et al., 2009).

As a consequence of the economic importance of these species, it was necessary its genetic improvement, in order to increase its productivity through more productive genotypes, to adapt the raw material to its final destination, to increase disease resistance and the tolerance to abiotic and climatic stresses (GOLLE et al., 2009; GONÇALEZ et al., 2014; PEREIRA et al., 2000). Hybridization and cloning have been the main drivers of Brazilian forest development, since this is considered a viable solution for the country's many demands for wood. The most widespread hybrid in Brazil comes from the crossing of *Eucalyptus urophylla* x *Eucalyptus grandis*. A breeding adapted to the most different conditions due to its tolerance to water deficit (characteristic of *E. urophylla*) and potential of rooting and field growth (characteristics of *E. grandis*) (NEVES et al., 2011; BRISOLA; DEMARCO, 2011).

2.2. Forest growth and production modeling

A mathematical model is a mathematical formulation based on hypotheses, which attempts to represent physical or biological phenomena, in order to generate an equation that can quantitatively estimate such phenomenon at a given level of probability (SILVA, 2015). The origins of forest growth and production models date back to the late eighteenth century in Germany with the development of volume tables. At present, prognoses are performed with equations or systems of interrelated mathematical equations rather than as volume tables, a change driven by the popularization of computational systems (ALDANA, 2010).

The application of mathematical modeling to the forests growth and production, whether planted or native, give results well known in the literature. These models help researchers and managers especially with the prediction or prognosis of forest yields, with the purpose of selecting better management options, more adequate silvicultural alternatives or forest harvesting planning (BURKHART; TOMÉ, 2012). Different methods can be employed to the estimation of structural parameters through the cloud of LiDAR points, all of them are based on allometric relationships between statistical metrics derived from LiDAR points related to forest canopy height and structural measurements obtained in the field (ANDERSON et al., 2006).

The main method used when modeling the relationship between LiDAR and field data is the parametric regression (multiple linear regression). The main advantage of using this type of methodology is the simplicity and clarity of the resulting model. In contrast, this method also has some drawbacks: this process provides a set of highly correlated predictors with little physical justification and, as a parametric technique, it is only recommended when assumptions such as normality, homoscedasticity, independence and linearity are met (OSBORN; WATERS, 2002).

Recent studies have shown that non-parametric techniques such as machine learning tools can be successfully used for the estimation of forest attributes (HUDAK et al., 2008; LATIFI et al., 2010; SIMARD et al., 2011). The speed and ease of implementation of these approaches, absence of restrictive assumptions, and the ability of some to include categorical dependent and (or) independent variables are contributing to their increased popularity (WITTEN; FRANK, 2000; GARCÍA-GUTIÉRREZ et al., 2015; GÖRGENS et al., 2015; SHIN et al., 2016; LEE et al., 2018; AYREY; HAYES, 2018).

2.3. Remote Sensing applied to the study of vegetation

In terms of definition, Meneses et al. (2012) present remote sensing as "a science that aims to develop the imaging of the Earth's surface by quantitatively detecting and measuring the responses of electromagnetic radiation interactions with the terrestrial materials ". However, the use of remote sensing techniques does not always allow obtaining images, but other types of data. In general, remote sensing can be understood as the set of techniques for obtaining data about a given target, so that the sensor is at a remote distance from it and therefore, there is no physical contact between the two (FIGUEIREDO, 2005).

Novo (2008) comments that currently available remote sensing systems provide consistent data of the Earth's surface, which are of great utility for various applications. Concerning ecological and vegetation applications, Blaschke and Kux (2007) point out that reliable spatial data and landscape ecology parameters are of great importance for the tasks of protecting and developing the environment and nature. Research related to vegetation enables the understanding of the structure and dynamics of plant formations, besides supporting the planning of actions that subsidize the management and preservation of natural environments (FLORENZANO, 2011).

Formigoni et al. (2011) state that the monitoring of vegetation cover using remote sensing products and techniques is based on the need to analyze plant resources, contributing to the temporal monitoring and obtaining information such as the distribution of vegetation types, phenology, canopy structure, stress conditions and changes in soil use. The development of remote sensing techniques has allowed the acquisition of diverse information about the terrestrial surface, contributing mainly to the investigations of the biophysical parameters of the vegetation, such as Foliar Area Index (LAI), percentage of green cover, chlorophyll content and even its detailed three-dimensional configurations, supporting the temporal, edaphic and phenological analyzes of the vegetation (VIGANÓ et al., 2011).

To Cintra (2007), forest management based on accurate spatial information favors the decision-making process in the following aspects: increasing the availability and quality of information, facilitating the developed activities; greater agility in the understanding of phenomena and processes; reduces the risk of errors, increasing the reliability of decisions; generate faster and more precise decisions, in addition to localized interventions, becoming possible elements of cost reduction.

2.4. LiDAR

The function of a remote sensor is to capture and measure the amount of energy reflected and/or emitted by targets, and thus obtain information about the nature and conditions of these targets, associating them with the real world (CENTENO, 2004). The spectral, spatial and texture information of the acquired images of passive sensors (i.e. aerial photographs and satellite imagery) are the main components used in the characterization of the forest. However, these attributes are related only to the horizontal structures of the forest, being insensitive to the measurements of the vertical structures (GOETZ et al., 2009). Another use limitation of passive sensors is its dependence on the sun as a source of illumination, thus data analysis can be impaired by the presence of clouds, which may be constant in some regions (D'OLIVEIRA et al., 2014).

One solution to these limitations is the use of active sensors, which use their own energy source, such as Light Detection and Ranging (LiDAR). In case of LiDAR, the light source is the laser, which emits short-wave electromagnetic radiation (1-10 μm), considered a direct method in data capture. The differential of this sensor is in its ability to obtain three-dimensional characteristics of the analyzed targets, which allows to calculate both horizontal and vertical structures of the forest, such as height, topography below the canopy of trees and distribution (HOLMGREN et al., 2003). Another important feature of this sensor is that laser pulses can penetrate through small openings in the forest canopy and provide accurate topography maps with high resolution, precise estimates of height and vegetation overlap, among other aspects of treetops (COOPS et al., 2007).

The technique of airborne LiDAR imaging consists of the emission of laser pulses directed to the ground by a mirror in a direction transverse to the direction of the flight line and the simultaneous measurement of the round-trip time of the energy of a pulse between the sensor and the target. This incident energy pulse interacts with the tree canopy (leaves, branches and trunk) and the ground surface, returning to the instrument. The time interval from the movement of the pulse from the beginning to its return to the sensor is measured, providing the distance between the instrument and the object (JENSEN, 2011).

When you reach objects without a well-defined surface, a laser signal can produce multiple reflection registers (returns or pulses). For forest studies the most important are the first and the last. The first return refers to the canopy top surface of the forest, it provides information from the higher portion of the objects and is used to model the canopy surface. The last return is reflected

at the lowest level reached by the laser and is used for the terrain modeling. This allows the separation of the vegetation from the soil surface to generate the digital height models (COELHO; VARGAS, 2007; JENSEN, 2011).

Each laser point projected on the ground has its planimetric coordinates and surface elevation measurements recorded. The planimetric coordinates of latitude and longitude of the laser points are obtained by means of the exact synchronization of an integrated orientation and position system, composed by a Differential Global Positioning System (DGPS) an Inertial Measurement Unit (IMU), and the data of the laser (CASTRO, CENTENO, 2005). In addition to the planimetric coordinates, the system stores, for each mapped point, the reflectance value of the target materials. It is possible, through this information, to generate orthoimages of intensity and hypsometry (intensity combined with altimetric information) (ZANDONÁ et al., 2008).

As a result of the survey, a cloud of laser scan points, or data set with XYZ coordinates, can be used to generate a computational model of the reflected surface below. After clipping the LiDAR points cloud (based on georeferenced field plots), statistics metrics are then extracted from this cloud and linked to the forest attributes measured in field plots. These extracted statistics are referred to specific metrics grouped in terms of height and density. The height metrics consist primarily of measures of location (i.e. mean, median, percentiles) and measures of dispersion (e.g. standard deviation, interquartile range, variance). The density metrics compute ratios of returns above a height break (WHITE et al., 2013).

From the cloud of points, it is also possible to obtain the Digital Surface Model (DSM) and the Digital Elevation Model (DEM), which, by their difference, provide accurate height data, generating the Digital Heights Model (DHM) or Digital Vegetation Model (DVM), considered the most difficult and time-consuming information to obtain in the field (CROW et al., 2007). Through the Digital Heights Model (DHM) it is possible to obtain direct measurements and estimates of important dendrometric variables of both native and planted forests. The direct measurements obtained from the LiDAR data are the counting of individual trees; individual tree height; crown diameter (characteristic with high correlation with volume and biomass) (NUTTO; SPATHELF, 2008). In addition, it is possible to estimate other important structural features for forest management, such as biomass, basal area, diameter and volume through modeling techniques combined with direct measurements (DUBAYAH et al., 2000; NUTTO; SPATHELF, 2008). It is also possible to carry out a forest stratification by means of height, which can be used as an

auxiliary tool for the definition and allocation of field plots, reducing sample error of estimates (SCOLFORO; MELO, 2006).

2.5. Individual Tree Detection

LiDAR derived data have demonstrated great potential for estimating forest parameters at both plot level and individual tree level (VAUHKONEN et al., 2014). Data processing is usually initiated by filtering the point cloud for the terrain classification and the generation of a digital terrain model (DTM). After the generation of the DTM, the model is used for the normalization of the point cloud or the generation of a canopy height model (CHM), frequently used in the individual tree detection. The CHM is generated by the difference between the digital surface model (DSM), given by the interpolation of the highest points (first return of the pulse), by the DTM (HYYPÄ et al., 2015).

From the CHM, individual trees can be detected, usually through some local maxima algorithm (higher points), watershed segmentation, region growth, among others (FAVORSKAYA; JAIN, 2017). An important point to note is that it is not usually possible to detect all trees, especially the ones with intertwined or dominated canopies (VAN LEEUWEN; NIEUWENHUIS, 2010; VAUHKONEN et al., 2014), which is the main error from the individual tree-based methods. The detection success depends on several factors, such as the forest conditions (density and spatial pattern), algorithms and parameters, among others (VAUHKONEN et al., 2014).

The crowns delimitation is usually based on some type of segmentation, edge detection, or local minima algorithms (HYYPÄ et al., 2015; VAN LEEUWEN; NIEUWENHUIS, 2010). In addition to the canopy delimitation, the height of the detected trees is usually extracted, based on the heights of the CHM, but it is common to observe an underestimation of the heights (HYYPÄ et al., 2008; VAN LEEUWEN; NIEUWENHUIS, 2010). This underestimation is explained by the fact that LiDAR pulses reach the sides of the crowns more often than the crown's maximum point, especially in conifers (BALDAUF; GARCIA, 2016).

Area-based methods (usually in plots) present a slightly different procedure, usually as a basic process: delimitation of the area, execution of field inventory, acquisition of LiDAR data and processing, adjustment of model between LiDAR data and field, and parameter estimation for the whole area (NÆSSET, 2014). The area-based methods are the most consolidated, and

considered standard in forest inventories, allowing the estimation of variables such as height, volume and basal area of plots (WHITE et al., 2016). Despite this, individual tree methods have been widely applied in the recent years and have the advantage of not requiring as much field data volume as it is necessary for area methods (due to the need for calibration for each case), besides having a good biological relation with parameters as volume (HYYPÄ et al., 2015). Area based methods, on the other hand, are easier to integrate with the current system of field-applied plot inventories and allow for more cost-effective ALS data collection (WULDER et al., 2012), since individual tree methods require high densities of points (WHITE et al., 2016).

2.6. Machine Learning

In parallel to the advances in remote sensing, computational techniques, such as machine learning algorithms (MLA), have been increasingly used to model spectral and biological data. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention (BREIMAN, 2001). Typically, machine learning algorithms try to describe the procedure of learning and show what is learned and express it as a set of rules. One of the most widely used learning procedure is the so-called supervised learning, where data are splitted in to input and target group in which it tries to map input data to target values in which input data are training data. A model which presented into learning process try to make a prediction and is corrected when those predictions are wrong. In this way training process continues until the model achieves a desired level of accuracy on the training data (BROWNLEE, 2013).

These techniques are able to overcome the difficulties of classical statistical methods such as spatial correlation, non-linearity of data, and overfitting (WERE et al., 2015). An additional advantage is that machine learning allows the user to implement a continuous learning process. Previous remote sensing studies have shown a superior or promising level of performance by machine learning techniques over more classical methods (FANG et al., 2003; ATZBERGER, 2004; DURBHA et al., 2007; ZHAO et al., 2008; ZHAO et al., 2011). These studies highlight the benefits of applying more robust techniques in solving problems previously resolved by traditional statistical modelling.

Machine Learning contains different algorithms in base of type of learning. In this study four different supervised machine learning algorithms were analyzed: *k*-nearest neighbors (k-NN), random forest (RF), support vector machine (SVM) and artificial neural networks (NNT):

- i. The k-NN algorithm uses a set of predictor feature variables (*X*) to match each target pixel to a number (*k*) of most similar (nearest neighbors or NN) reference pixels for which values of response variables (*Y*) are known (MCROBERTS, 2012). It allows estimating a variable of interest through a weighted average of the known variables of the *k*-nearest neighboring plots. The weighted average could be done using either an inverse distance weighting or the square of the inverse distance (MCROBERTS, 2013). Examples of forest applications using k-NN imputation, including mathematical formulation may be found in Hudak et al. (2008), Racine et al. (2014), and Fekety et al. (2014).
- ii. The RF algorithm, initially proposed by Breiman (2001), is an ensemble method that generates a set of individually trained decision trees and combines their results. The greatest advantage of these decision trees as regression methods is that they are able to accurately describe complex relationships among multiple variables, and by aggregating these decision trees, more accurate solutions are generated (GLEASON; IM, 2012). In addition to these characteristics, RF is an easy parameterization method (IMMITZER et al., 2012). This method has shown great potential in regression studies, in some cases generating better results than conventional techniques (GARCÍA-GUTIÉRREZ et al., 2015; GÖRGENS et al., 2015; WU et al., 2016).
- iii. SVMs operate by assuming that each set of inputs will have a unique relation to the response variable and that the grouping and the relation of these predictors to one another is sufficient to identify rules that can be used to predict the response variable from new input sets. For this, SVMs project the input space data into a feature space with a much larger dimension, enabling linearly non-separable data to become separable in the feature space. This method has been successfully used in forestry classification and regression problems (SHAO; LUNETTA 2012; GARCÍA-GUTIÉRREZ et al., 2015; WU et al., 2016).
- iv. NNTs are a parallel-distributed information processing system that simulates the working of neurons in the human brain, being able to learn from examples. Artificial neural networks are widely used to model complex and non-linear relations between inputs and

outputs or to determine patterns in data (DIAMANTOPOULOU, 2012). The use of this technique in conjunction with remote sensing data is consolidated in several studies (CLUTER et al., 2012; GARCÍA-GUTIÉRREZ et al., 2015; RODRIGUEZ-GALIANO et al., 2015; WERE et al. 2015).

3. REFERENCES

AYREY, E.; HAYES, D. J. The use of three-dimensional convolutional neural networks to interpret LiDAR for forest inventory. **Remote Sensing**, v. 10, n. 4, p. 1–16, 2018.

ALDANA, E.P. **Medición forestal**. La Habana: Editorial Felix Varela, 2010. 266 p.

ANDERSON, J., et al. The use of waveform lidar to measure northern temperate mixed conifer and deciduous forest structure in New Hampshire. **Remote Sensing of Environment**, v. 105, p. 248–261, 2006.

ANDRADE, M. C. R. **Proposta de redução de custos em inventários florestais por meio do uso de técnicas de geoprocessamento**. São José dos Campos: Inpe, 2013.

ATZBERGER, C. Object-based retrieval of biophysical canopy variables using artificial neural nets and radiative transfer models. **Remote Sensing of Environment**, Amsterdam, v. 93, n. 1/2, p. 53-67, 2004.

AVARENGA, L. H. V. **Imagens de alta resolução e geoestatística na estratificação da fisionomia cerrado para inventários florestais**. 92 f. Dissertação (Mestrado), Universidade Federal de Lavras, Lavras - MG, 2012.

BALDAUF, T.; GARCIA, M. Image processing of radar and Lidar in tropical forestry. In: PANCEL, L.; KÖHL, M. (Eds.). **Tropical Forestry Handbook**. 1. ed. Berlin: Springer-Verlag Berlin Heidelberg, p. 635–661, 2016.

BLASCHKE, T.; KUX, H. **Sensoriamento remoto e SIG avançados: novos sistemas e sensores inovadores**. 2ª. Edição, São Paulo: Oficina de Textos, 304 p., 2007.

BREIMAN, L., **Random forests**. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

BRISOLA, S. H.; DEMARCO, D. Análise anatômica do caule de *Eucalyptus grandis*, *E. urophylla* e *E. grandis* x *urophylla*: desenvolvimento da madeira e sua importância para a indústria. **Scientia Forestalis**, v. 39, n. 91, p. 317-330, 2011.

BROWNLEE, J. **A tour of machine learning algorithms**. *Machine Learning Mastery*, 2013.

BURKHART, H.E.; TOME, M. **Modeling forest trees and stands**. Springer, Dordrecht, The

Netherlands, 2012. 132 p.

CAMPOS, J. C. C.; LEITE, H.G. **Mensuração Florestal: Perguntas e Respostas**. 4 Ed. Viçosa: UFV, 2013. 605 p.

CASTRO, F. C.; CENTENO, T. M. Segmentação de imagens geradas por perfilamento a laser para delimitação de árvores individuais em uma área de reflorestamento de eucaliptos. In: SBSR, 12., 2005, Goiânia. **Anais...** Goiânia: SBSR, 2005.

CENTENO, J. A. S. **Sensoriamento Remoto e Processamento de Imagens Digitais**. Curitiba: Curso de Pós-Graduação em Ciências Geodésicas, Universidade Federal do Paraná, 2004. 219 p.

CINTRA, D. P. **Classificação de estágios sucessionais florestais por meio de imagens de alta resolução (IKONOS) no Parque estadual da Pedra Branca**, RJ. 2007. 64f. Dissertação – Universidade Federal Rural do Rio de Janeiro, Seropédica – RJ, 2007. Disponível em: <http://www.if.ufrj.br/pgcaf/pdfdt/Dissertacao%20Danielle%20Cintra.pdf> Acesso em: 12 de jul. de 2018.

CLUTER, M. et al. Estimating tropical forest biomass with a combination of SAR image texture and Landsat TM data: An assessment of predictions between regions. **ISPRS Journal of Photogrammetry and Remote Sensing**, 70,66–77, 2012.

COELHO, A. H.; VARGAS, R. M. A. Geração de modelos digitais de terreno a partir de dados de laser scanner aerotransportado em área de floresta usando o software livre GRASS. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 13., 2007, Florianópolis. **Anais...** Florianópolis: [s. n.], p. 3653-3660, 2007.

COOPS, N. C. et al. Estimating canopy structure of Douglas-Fir forest stands from discrete-return LiDAR, **Trees - Structure and Function**, New York, v. 21, p. 295-310, 2007.

CROW, P. et al. Woodland vegetation and its implications for archaeological survey using LiDAR. **Forestry**, v. 80, n. 3, 2007.

DIAMANTOPOULOU, M. Assessing a reliable modeling approach of features of trees through neural network models for sustainable forests. **Sustainable Computing: Informatics and Systems**, v. 2, p. 190–197, 2012.

D'OLIVEIRA, M. V. N. et al. Uso do Lidar como Ferramenta para o Manejo de Precisão em Florestas Tropicais. Rio Branco, AC.: **Embrapa**, 2014. 130 p.

DUBAYAH, R. et al. Land surface characterization using LiDAR remote sensing. In: HILL, M. J. & ASPINALL, R.J. (eds), **Spatial Information for Land Use Management**. Gordon & Breach Science Publishers, Amsterdam. p. 25-38, 2000.

DURBHA, S.S. et al. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. **Remote Sensing of Environment**, Amsterdam, v. 107, n.

1/2, p. 348-361, 2007.

FANG, H.L. et al. Retrieving leaf area index using a genetic algorithm with a canopy radiative transfer model. **Remote Sensing of Environment**, Amsterdam, v. 85, n. 3, p. 257-270, 2003.

FAVORSKAYA, M. N.; JAIN, L. C. Handbook on Advances in Remote Sensing and Geographic Information Systems Paradigms and Applications in Forest Landscape Modeling. **Boca Raton: CRC Press**, 2017.

FEKETY P.A. et al. Temporal transferability of LiDAR-based imputation of forest inventory attributes. **Canadian Journal Forest Resources**, v. 45, p. 422-435, 2014.

FIGUEIREDO, D. **Conceitos básicos de sensoriamento remoto**. Companhia Nacional de Abastecimento-CONAB. Brasília-DF, 2005. Disponível em: http://www.conab.gov.br/conabweb/download/SIGABRASIL/manuais/conceitos_sm.pdf. Acesso em: 21 de jul. de 2018.

FLORENZANO, T. G. Geotecnologias na geografia aplicada: difusão e acesso, **Revista do Departamento de Geografia**, v. 17, p. 24-29, 2011. Disponível em: <http://www.revistas.usp.br/rdg/article/view/47272/51008>. Acesso em: 27 de jul. de 2018.

FORMIGONI, M. H. et al. Análise temporal da vegetação na região do Nordeste através de dados EVI do MODIS. **Ciência Florestal**, v. 21, n. 1, p. 1-8, 2011. Disponível em: <https://periodicos.ufsm.br/cienciaflorestal/article/view/2740/1667> Acesso em: 13 de ago. de 2018.

GLEASON, C.J.; IM, J. Forest biomass estimation from airborne LiDAR data using machine learning approaches. **Remote Sensing of Environment**, Amsterdam, v. 125, p. 80-91, 2012

GOETZ, S. J. et al. Mapping and monitoring carbono stocks with satellite observations: a comparison of methods. **Carbon Balance and Management**, Heidelberg, v. 4, n. 1, p. 2-9, 2009.

GOLLE, D. P. et al. Melhoramento florestal: ênfase na aplicação da biotecnologia. **Ciência Rural**, Santa Maria, v. 39, n. 5, p. 1607-1614, 2009.

GONÇALEZ, J. C. et al. Relações entre dimensões de fibras e de densidade da madeira ao longo do tronco de *Eucalyptus urograndis*. **Scientia Forestalis**, Piracicaba, v. 42, n. 101, p. 81-89, 2014.

GÖRGENS, E. B. et al. A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. **Computers and Electronics in Agriculture**, v. 116, p. 221–227, 2015.

GUEDES, I. C. L. et al. Spatial continuity of dendrometric characteristics in clonal cultivated *Eucalyptus* sp. throughout the time. **Cerne**, v. 21, n. 4, p. 527–534, 2015.

HASELEIN, C. R. et al. Características tecnológicas da madeira de árvores matrizes de *Eucalyptus grandis*. **Ciência Florestal**, Santa Maria, v. 14, n. 2, p. 145-155, 2005.

HOLMGREN, J.; NILSSON, M.; OLSSON, H. Estimation of tree height and stem volume on plots using airborne laser scanning. **Forest Science**, v. 49, p. 419-428, 2003.

HYAMS, D. Curve Expert Version 1.37. **A comprehensive curve fitting package for Windows**, 2005.

HYYPPÄ, J. et al. Remote sensing of forests from LiDAR and Radar. In: THENKABAIL, P. S. (Ed.). *Land Resources Monitoring, Modeling, and Mapping with Remote Sensing*. **Boca Raton, USA: CRC Press**, p. 397–428, 2015.

HYYPPÄ, J. et al. Review of methods of small - footprint airborne laser scanning for extracting forest inventory data in boreal forests. **International Journal of Remote Sensing**, v. 29, n. 5, p. 37–41, 2008.

HUDAK, A.T. et al. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from lidar data. **Remote Sensing of Environment**, v. 112, p. 2232–2245, 2008.

HUDAK, A. T. et al. Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. **Canadian Journal of Remote Sensing**, v. 32, n. 2, p. 126–138, 2006.

IMMITZER, M. et al. “Tree species classification with random forest using very high spatial resolution 8- bandWorldView-2 satellite data.” **Remote Sensing**, v. 4, p. 2661– 2693, 2012.

INDÚSTRIA BRASILEIRA DE ÁRVORES. **Relatório Iba 2018**, Brasília, p. 80, 2017. Disponível em: https://iba.org/images/shared/Biblioteca/IBA_RelatorioAnual2017.pdf. Acesso em: 02 junho 2018.

JENSEN, J. R. **Sensoriamento Remoto do Ambiente: Uma perspectiva em recursos terrestres**. Tradução de José Carlos N. Epiphanyo [et al.]. 1 Ed. São José dos Campos: Parêntese, 2011. 672 p.

KÖHL, M. et al. **Sampling methods, remote sensing and GIS multiresource forest inventory**. Springer Science & Business Media, 2006.

LATIFI, H. et al. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. **Forestry**, London, v. 83, n. 4, p. 395-407, 2010.

LARY, D. J. et al. Geoscience Frontiers Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, v. 7, n. 1, p. 3–10, 2016.

LEE, J. et al. Machine Learning Approaches for Estimating Forest Stand Height Using Plot-Based Observations and Airborne LiDAR Data. **Forests**, v. 9, n. 5, p. 268, 2018.

- LEFSKY, M. A. et al. Lidar Remote Sensing for Ecosystem Studies. v. 52, n. 1, p. 19–30, 2002.
- MARTINS, I. S et al. Alternativas de índices de seleção em uma população de *Eucalyptus grandis* Hill ex Maiden. **Cerne**, Lavras, v. 12, n. 3, p. 287-291, 2006.
- MCROBERTS, R. E., et al. Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications. **Remote Sensing of Environment**, v. 125, p. 157–166, 2012.
- MCROBERTS, R. E., et al. Inference for lidar-assisted estimation of forest growing stock volume. **Remote Sensing of Environment**, v. 128, p. 268–275, 2013.
- MENESES, P. R. et al. **Introdução ao processamento de imagens de sensoriamento remoto**. Brasília, 2012.
- MONTAÑO, R. A. N. R. **Aplicação de Técnicas de Aprendizado de Máquina na Mensuração Florestal**. 102 p. Phd Thesis. Universidade Federal do Paraná. [S.l.]: [s.n.], 2016.
- NÆSSET, E. Area-Based Inventory in Norway – From innovation to an operational reality. In: MALTAMO, M.; NÆSSET, E.; VAUHKONEN, J. (Eds.). **Forestry Applications of Airborne Laser Scanning: Concepts and Case Studies**. Dordrecht: Springer, 2014, 240 p.
- NEVES, T. A. et al. Avaliação de clones de *Eucalyptus* em diferentes locais visando a produção de carvão vegetal. **Pesquisa Florestal Brasileira**, Colombo, v. 31, n. 68, p. 319-330, 2011.
- NOVO, E. M. L. M. **Sensoriamento Remoto: princípios e aplicações**. 3. ed. São Paulo: Edgard Blucher, v. 01. p. 363, 2008.
- NUTTO, L.; SPATHELF, P. **Modelling diameter growth of plantation grown eucalypts based on crown projection area**. Porto Seguro: IPEF, 2008, 360 p.
- OLIVEIRA, L. T. et al. Determinação do volume de madeira em povoamento de eucalipto por escâner a laser aerotransportado. **Pesquisa Agropecuária Brasileira (Impressa)**, v. 49, p. 692-700, 2014.
- OLIVEIRA, L. T. **Aplicação do lidar no inventário de florestas plantadas**. 109 p. Phd Thesis – Universidade Federal de Lavras, Lavras, 2011.
- OSBORNE, J.; WATERS E., **Four assumptions of multiple regression that researchers should always test**, Pract. Assess. Res. Eval. 8 (2), 2002.
- PEREIRA, J. C. D. et al. **Características da madeira de algumas espécies de eucalipto plantadas no Brasil**. Colombo: Embrapa Florestas, 2000, 114 p.
- R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org/>. 2014.

RACINE, E.B. et al. Estimating forest stand age from LiDAR-derived predictors and nearest neighbour imputation. **Forest Science**, v. 60, n. 1, p. 128–136, 2014.

RIBEIRO, C. A. A. S. **Floresta de precisão**. In: MACHADO, C. C. (Ed.) Colheita florestal. Viçosa, Ed. Universidade Federal de Viçosa, 2002. 468 p.

RODRIGUEZ-GALIANO, V. et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. **Ore Geology Reviews**, v. 71, p. 804–818, 2015.

SANQUETTA, C. R.; BALBINOT, R. Metodologias para Determinação de Biomassa Florestal (p. 77-94). In: SANQUETTA, C. R.; BALBINOT, R.; ZILLIOTTO, M. A. (Eds.) **Fixação de carbono: atualidade, projetos e pesquisas**. Curitiba/PR: [s.n], 2004. 205 p.

SCOLFORO, J. R. S.; MELLO, J. M. **Inventário florestal**. Lavras: FAEPE, 2006. 561 p.

SCOLFORO, H.F. et al. Modeling dominant height growth of Eucalyptus plantations with parameters conditioned to climatic variations. **Forest Ecology and Management**, v. 380, p.182–195, 2016.

SHAO, Y.; LUNETTA, R. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 70, p. 78–87, 2012.

SHIN, J. et al. Comparing Modeling Methods for Predicting Forest Attributes Using LiDAR Metrics and Ground Measurements. **Canadian Journal of Remote Sensing**, v. 42, n. 6, p. 739–765, 2016.

SILVA, A. L. L. et al. Tolerância ao resfriamento e congelamento de folhas de eucalipto. **Biociências**, Porto Alegre, v. 17, n. 1, p. 86-90, 2009.

SILVA, C. A. et al. Mapping aboveground carbon stocks using LiDAR data in *Eucalyptus* spp. plantations in the state of São Paulo. **Scientia Forestalis**, v. 42, p. 591–604, 2014

SILVA, J.A.A. da. Conceitos e princípios básicos de modelagem matemática em ciências florestais. **Anais da Academia Pernambucana de Ciência Agronômica**, Recife, v. 11/12, p.195-215, 2015.

SIMARD, M. et al. Mapping forest canopy height globally with spaceborne lidar. **Journal of Geophysical Research: Biogeosciences**, New York, v. 116, p. 1-12, 2011.

TRUGILHO, P. F. et al. Avaliação de clones de *Eucalyptus* para a produção de carvão vegetal. **Cerne**, Lavras, v. 7, n. 2, p. 104-114, 2001.

VALVERDE, S. R. Plantações de eucalipto no Brasil. **Revista da Madeira**, n.107, 2007. Disponível em: <http://www.remade.com.br/revistadamadeira>. Acesso em: jul. 2018.

VAN LEEUWEN, M.; NIEUWENHUIS, M. Retrieval of forest structural parameters using LiDAR remote sensing. **European Journal of Forest Research**, v. 129, n. 4, p. 749–770, 2010.

VAUHKONEN, J. et al. Introduction to forestry applications of airborne laser scanning. In: MALTAMO, M.; NÆSSET, E.; VAUHKONEN, J. (Eds.). **Forestry Applications of Airborne Laser Scanning**. Dordrecht: Springer, p. 1–16, 2014.

VIGANÓ, H. A. et al. **Análise do desempenho dos Índices de Vegetação NDVI e SAVI a partir de imagem Aster**. Curitiba - PR, p.1828, 2011. Disponível em: <http://www.dsr.inpe.br/sbsr2011/files/p1364.pdf>Acesso em: 24 de set. de 2018.

WERE, K. et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afrotropical landscape. **Ecological Indicators**, v. 52, p. 394–403, 2015.

WITTEN, I. H.; FRANK, E. Data mining : practical machine learning tools and techniques with Java implementations. Morgan Kaufmann series in data management systems. **ST-Data mining: practical machine**, p. 369, 2000.

WHITE, J.C. et al. The utility of image-based point clouds for forest inventory: a comparison with airborne laser scanning. **Forests**, Basel, v. 4, n. 3, p. 518-536, 2013.

WHITE, J. C. et al. Remote sensing technologies for enhancing forest inventories: a review. **Canadian Journal of Remote Sensing**, v. 42, n. 5, p. 619–641, 2016.

WU, C. et al. Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery. **Journal of Applied Remote Sensing**, v. 10, p. 3, 2016.

WULDER, M. A. et al. Lidar sampling for large-area forest characterization: a review. **Remote Sensing of Environment**, v. 121, p. 196–209, 2012.

ZANDONÁ, D. F. et al. Varredura a Laser aerotransportado para estimativa de variáveis dendrométricas. **ScientiaForestalis**, Piracicaba, v. 36, n. 80, p. 295-306, dez. 2008.

ZHAO, K.; et al. Bayesian learning with Gaussian processes for supervised classification of hyperspectral data. **Photogrammetric Engineering and Remote Sensing**, Bethesda, v. 74, n. 10, p. 1223-1234, 2008

ZHAO, K. et al. Characterizing forest canopy structure with lidar composite metrics and machine learning. **Remote Sensing of Environment**, Amsterdam, v. 115, n. 8, p. 1978-1996, 2011.

CHAPTER I

COMBINED IMPACT OF SAMPLE SIZE AND MODELING TECHNIQUES FOR
PREDICTING VOLUME IN *Eucalyptus* spp. PLANTATIONS FROM LiDAR DATA

SILVA, VANESSA SOUSA DA. Combined impact of sample size and modeling techniques for predicting volume in *Eucalyptus* spp. plantations from LiDAR. 2019. Advisor: Emanuel Araújo Silva. Co-Advisors: Carlos Alberto Silva e Gabrielle Hambrecht Loureiro.

ABSTRACT

Current forest growing stock inventory methods used in *Eucalyptus* spp. plantations in Brazil are based on statistical methods using field measurements of trees on sample plots. Such measurements are carried out with traditional methods and equipment. Nowadays, Light Detection and Ranging (LiDAR) remote sensing has been established as one of the promising and primary tools for large-scale forest characterization and mapping. The analysis of LiDAR remote sensing information combined with field data has been used by several authors to support forest management. Continuous advances in computational techniques, such as machine learning algorithms, have been increasingly used to model biological data attaining highly accurate forest attributes estimations. While there have been previous studies exploring the use of LiDAR and machine learning algorithm for forest inventory modeling no studies yet have demonstrated the combined impact of sample size and different modeling techniques for predicting and mapping stem total volume in industrial *Eucalyptus* spp. plantations. This study aimed to compare the effects of ten parametric and nonparametric modeling methods for estimating volume in Eucalyptus forest plantation using airborne LiDAR data while varying the reference data (sample) size. The study was conducted at the municipalities of Pilar do Sul and São Miguel Arcanjo, southeast region of the state of São Paulo, Brazil, based on LiDAR survey and field inventory. The modeling techniques were compared in terms of RMSE, Bias and R^2 with 500 simulations. The best performance was verified for the OLS method, which was able to provide comparable results to the traditional forest inventory approaches using only 40% of the total field plots, followed by the random forest (RF) algorithm with identical sample size value.

Key-words:— LiDAR, Eucalyptus, Volume, Machine learning, Remote Sensing.

SILVA, VANESSA SOUSA DA. Impacto combinado do tamanho da amostra e técnicas de modelagem para predição volumétrica em plantios de *Eucalyptus* spp. a partir de dados LiDAR. 2018. Orientador: Emanuel Araújo Silva. Co-orientadores: Carlos Alberto Silva e Gabrielle Hambrecht Loureiro.

RESUMO

Os atuais métodos de inventário de estoque florestal usados em plantações de *Eucalyptus* spp. no Brasil são baseados em métodos estatísticos usando medições de campo de árvores em amostras. Tais medições são realizadas com métodos e equipamentos tradicionais. Atualmente, o sensoriamento remoto com sensor LiDAR (Light Detection and Ranging) foi estabelecido como uma das ferramentas promissoras e primárias para a caracterização e mapeamento de florestas em larga escala. A análise de informações de sensoriamento remoto LiDAR combinada com dados de campo tem sido usada por vários autores para apoiar o manejo florestal. Avanços contínuos em técnicas computacionais, como algoritmos de aprendizado de máquina (machine learning), têm sido cada vez mais usados para modelar dados biológicos, obtendo estimativas de atributos florestais altamente precisos. Embora tenha havido estudos anteriores explorando o uso de LiDAR e algoritmos de aprendizado de máquina para modelagem de inventário florestal, nenhum estudo demonstrou o impacto combinado do tamanho da amostra e diferentes técnicas de modelagem para prever e mapear o volume total em plantios de *Eucalyptus* spp. O objetivo deste estudo foi comparar os efeitos de dez métodos de modelagem paramétrica e não-paramétrica para estimar o volume em plantio de florestas de eucalipto utilizando dados de LiDAR aerotransportado, variando o tamanho dos dados de referência (amostra). O estudo foi realizado nos municípios de Pilar do Sul e São Miguel Arcanjo, região sudeste do estado de São Paulo, Brasil, com base em levantamento LiDAR e inventário de campo. As técnicas de modelagem foram comparadas em termos de RMSE, Bias e R^2 com 500 simulações. O melhor desempenho foi verificado para o método OLS, que foi capaz de fornecer resultados comparáveis às abordagens tradicionais de inventário florestal usando apenas 40% do total de parcelas de campo, seguido pelo algoritmo Random Forest (RF) com valor de tamanho de amostra idêntico.

Palavras-chave:— LiDAR, *Eucalyptus*, Volume, Aprendizado de máquina, Sensoriamento Remoto.

1. INTRODUCTION

The area of land covered with planted forests is growing worldwide. According to FAO (2015), since 1990, tropical and subtropical regions have been experiencing particularly rapid rates of forest plantation expansion, mostly in countries in Asia and South America by 4.3 million ha/year. Planted forests correspond to an estimated 7% of the global forest area and cover an area of 264 million ha (BROTTO et al., 2016). Timber production is the main ecosystem service of planted forests and the main management objective for these plantations (GAO et al., 2016).

Members of eucalyptus are now among the most valuable and widely planted hardwoods (ROCKWOOD et al., 2008). In Brazil, the area of Eucalyptus plantations has dramatically risen in the last few decades. Because of its high growth rate, *Eucalyptus* spp. became the major short fiber source of raw material primarily to supply the pulp and paper industries in southeast Brazil. Currently, eucalyptus plantations occupy around 5.7 million hectares (71.9% of the total planted forest area in Brazil) and represent 17% of the harvested wood in the world (IBÁ, 2018).

The correct determination of stand productivity is essential to support forest management planning strategies (GONZÁLEZ-GARCÍA et al., 2015; RETSLAFF et al., 2015). Traditionally, productivity assessments and optimal harvesting time predictions are carried out based on field measurements of the diameter at breast height (DBH) and tree height via forest inventory. However, in fast-growing plantations, field-based inventory is an expensive, extremely time consuming and labor-intensive task, which may not even be sufficient to identify problematic conditions, such as those arising from losses due to pest and disease attacks or from climatic anomalies (GONZÁLEZ-GARCÍA et al., 2015, SCOLFORO et al., 2016).

In the past decade, advances in remote sensing have provided new tools, techniques, and technologies to support forest management. Thus, low-cost and accurate forest productivity assessment can be made, as well as allowing the collection of information in areas not sampled by forest inventory (MORGENROTH; VISSER, 2013). Light Detection and Ranging (LiDAR) remote sensing has been established as one of the promising and primary tools for broad-scale forest characterization (MONTAGHI et al., 2013). LiDAR data can be used to characterize local to regional spatial extents with high enough resolution to quantify the three-dimensional information of vertical and horizontal forest structures and the underlying topography with the support of efficiently collected field data and several statistical methods (NÆSSET, 2004; WHITE et al., 2013; SILVA et al., 2016).

The analysis of LiDAR remote sensing information combined with field data has been used by several authors to produce highly accurate retrievals of tree density, stem total and assortment volumes, basal area, aboveground carbon, and leaf area index, and thereby can be an effective way to predict and map forest attributes at unsampled locations (LEFSKY et al., 2005; LATIFI et al., 2010; SILVA et al., 2014, 2016, 2018; DOS REIS et al., 2018). (SILVA et al., 2016) estimated the volume of a Eucalyptus plantation under different relief conditions in the southern region of Brazil from LiDAR data. The results found by these authors corroborate the potential use of data collected by LiDAR remote sensing to estimate the productivity of Eucalyptus plantations.

In order to predict forest attributes aiming improved management practices for wood and pulp production, it is often necessary to model height, basal area and stem total volumes of Eucalyptus plantations in operational and experimental scenarios (GÖRGENS et al, 2015). Current predictive modeling methods include parametric (i.e., multiple linear regression) and non-parametric (i.e., machine learning algorithms) approaches (SHIN et al., 2016). Multiple linear regression has usually been the main tool for the estimation of parameters regressed from LiDAR statistics. The main advantage of using this methodology is the simplicity and clarity of the resulting model. However, the method also has some drawbacks: it results in a set of highly correlated predictors with little physical justification and, as a parametric technique, it is only recommended when assumptions such as normality, homocedasticity, independence and linearity are met (WERE et al., 2015).

The advances in computational techniques, such as machine learning algorithms, have been increasingly used to model biological data. These techniques are able to overcome some of the abovementioned difficulties of classical statistical methods. In addition, these algorithms allow the use of categorical data, with statistical noise and incomplete data, and thus can address needs under different dataset scenarios (BREIMAN, 2001). Nonparametric machine learning modeling techniques have proven higher ability to identify complex relationships between predictor and dependent variables showing therefore its superiority or promising level of performance over more classical statistics methods for estimating forest parameters for inventory modeling from LiDAR data at either plot or stand-levels (ZHAO et al., 2009; FALKOWSKI et al., 2010; ZHAO et al., 2011; HUDAK et al., 2014; RACINE et al., 2014; SILVA et al., 2016).

Ahmed et al. (2015) modelled a Landsat time-series data structure in conjunction with LiDAR data and found that the random forest algorithm achieved better results than multiple

regression. In another study, García-Gutiérrez et al. (2015) found that machine learning algorithms (mainly support vector machine) were superior for modelling LiDAR data of a range of forest variables (i.e., aboveground biomass, basal area, dominant height, mean height, and volume) compared with multiple linear regression. These studies highlight the benefits of applying more robust techniques in solving problems previously resolved by traditional statistical modelling.

While there have been previous studies exploring the use of LiDAR and non-parametric machine learning algorithm for forest inventory modeling (LATIFI et al., 2010; PENNER et al., 2013; VALBUENA et al., 2017; DOS REIS et al., 2018), no studies yet have demonstrated the combined impact of sample size and different modeling techniques for predicting and mapping stem total volume in industrial *Eucalyptus* spp. plantations. Identifying the effective sample size of field plots is an important issue in LiDAR-based forest inventory. However, it is unclear how the combined effect of sample size and data modeling (parametric and non-parametric approach) may impact the accuracy of the stem total volume estimation from LiDAR.

Accurate forest inventory is of foremost importance to make operational, tactical, and strategic management decisions efficiently. Furthermore, the optimization of the entire supply chain management in pulp and paper companies maximizes its sustainability from both economic and environmental perspectives (FALKOWSKI et al., 2008, SILVA et al., 2016, NAKAJIMA et al., 2017). Therefore, to improve plantation management there is a need to develop and implement more accurate, repeatable, and robust frameworks for modeling and mapping forest inventory attributes. Moreover, efficient frameworks also play a key role in helping LiDAR technology move from research to operational modes, especially in industrial forest plantation settings where lidar applications are relatively new.

In this context, the aim of this study was, through the integration of field-based forest inventory and LiDAR data, to compare the performance of parametric and nonparametric methods in the estimation of stem total volume in industrial *Eucalyptus* spp. plantations while assessing how the combined effect of sample size and different modeling techniques may impact the accuracy of the predictions. This investigation was based on the hypothesis that LiDAR technology and machine learning algorithms can facilitate accurate and precise volumetric inferences in *Eucalyptus* spp. plantations in southeast Brazil.

2. MATERIALS AND METHODS

2.1. Study Area

The study area consisted of three farms located in the municipalities of Pilar do Sul and São Miguel Arcanjo, southeast region of the state of São Paulo, Brazil (Fig. 1). The climate of the region is characterized as humid subtropical, with wet and hot summers and dry and cold winters. Mean annual precipitation is ~1700 mm; mean annual temperature is 18.8 °C (ALVARES et al., 2013). The topography in the selected plantations is complex, ranging from mildly to very hilly with an elevation ranging from 659 m to 1210 m. The soils of the region are predominantly red and yellow red latosol, all are clayey or very clayey.

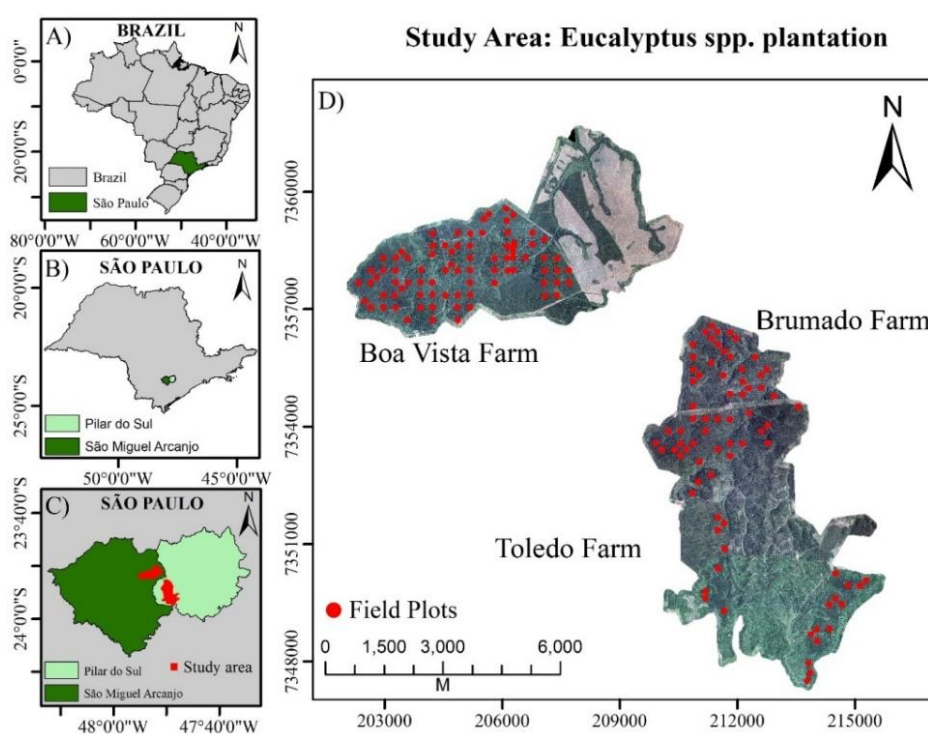


Figure 1. Location map of study area and plots. (A) Brazil and São Paulo State; (B) São Paulo State and the municipalities of Pilar do Sul and São Miguel Arcanjo; (C) Study area within the municipalities of Pilar do Sul and São Miguel Arcanjo; (D) Field plots in the study area.

The farms contained industrial Eucalyptus plantations managed by Suzano S.A., a pulp and paper company located in São Paulo state, Brazil. The plantations were composed of hybrid clones of two Eucalyptus species, *Eucalyptus grandis* W. Hill ex Maid and *Eucalyptus urophylla* S.T. Blake. All the trees were planted predominantly in a 3m x 2m grid configuration, resulting

in an average density of 1,667 trees ha⁻¹. Stand age across the farms was variable and ranged from 2 to 6 years.

2.2. Field Data

The study was based on data collected in a set of temporary and permanent sample plots installed for the purpose of annual forest inventory by the Suzano S.A. company. A total of 158 circular plots of 400 m² each were randomly established across the three farms. Measurements were carried out during the months of april to november of 2013. All the sample plots were georeferenced in the field using a geodetic GPS unit with differential correction capability (Trimble Pro-XR). The projected coordinate system used was UTM SIRGAS 2000, zone 23 S.

In each sample plot, individual trees were measured for diameter at breast height (DBH; cm) at 1.30 m, and a random subsample (15%) of trees for tree heights (Ht; m). Heights of unmeasured trees were estimated using locally adjusted hypsometric models, which use DBH as the predictor of Ht, following the model below:

$$\ln(Ht) = \beta_0 + \beta_1 \times \left(\frac{1}{dbh} \right) + \varepsilon \quad (1)$$

where $\ln(Ht)$ is the natural logarithm of tree height (m); β_0 and β_1 are the intercept and the slope of the model; dbh is the diameter at breast height (1.30 m) and ε is the random error of the model.

Field measurements were used to estimate two additional variables in each plot: tree basal area (BA; m²) and stem total volume (VOL; m³). Tree-level volumes (m³ tree⁻¹) were predicted by applying the respective diameter and height into the Schumacher-Hall allometric model adjusted for each region, rotation and genetic material, following the model below:

$$\ln(V) = \beta_0 + \beta_1 \ln(DBH) + \beta_2 \ln(Ht) + \varepsilon \quad (2)$$

where $\ln(V)$ = the natural logarithm of stem total volume (m³); β_i = model's parameters to be estimated ($i = 0, 1, 2$); DBH = diameter (cm) at breast height (1.30 m); Ht = total height and ε = model's random error.

All the field measurements, and predictions calculations from hypsometric and allometric models were provided by the inventory team of Suzano S.A. The coefficients of the models are

under the company's intellectual property rights and not made available to the public, however, the coefficients of determination (R^2) and standard errors of the estimate (SEE%) for the VOL models used in this study ranged from 0.96 to 0.98 and 8.3 to 12.7, respectively. The total of each variable of all individuals were summed at plot-level and scaled to a hectare. A summary of plot-level forest attributes including BA ($\text{m}^2 \text{ha}^{-1}$) and VOL ($\text{m}^3 \text{ha}^{-1}$) calculations for each class of stand ages is presented in Table 1.

Table 1. Summary statistics of forest attributes from ground measurement at the sample plots.

Ages	DBH (cm)		Ht (m)		BA ($\text{m}^2 \text{ha}^{-1}$)		VOL ($\text{m}^3 \text{ha}^{-1}$)		N plots
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
2.2	10.27	1.19	14.34	1.23	10.56	2.90	58.34	20.30	6
3.2	12.75	0.88	21.83	1.05	19.36	1.90	160.35	24.77	5
3.8	14.09	0.56	22.35	1.49	21.69	1.89	189.15	23.87	10
4.5	15.55	1.32	25.90	0.86	26.97	2.60	280.63	39.35	5
4.8	15.82	0.87	29.34	1.38	28.46	3.10	329.44	42.14	37
5.1	15.36	1.06	28.62	1.67	29.48	3.33	333.35	55.23	38
6	16.51	1.56	29.13	2.81	29.39	5.43	349.53	87.25	57

DBH= diameter at breast height (1.30 m); Ht= total height; BA= basal area; VOL= volume; Nplots= Number of plots; SD standard deviation

2.3. LiDAR data collection specifications and processing

An airborne LiDAR survey was conducted in the study area on December 5th, 2013 using a Harrier 68i sensor (Trimble, Sunnyvale, CA, USA) mounted on a CESSNA 206 aircraft. The characteristics of the LiDAR data acquisition are listed in Table 2. LiDAR data processing steps were performed using FUSION/LDV 3.7 software (US Forest Service, Washington, DC, USA) (MCGAUGHEY, 2016) which provided three major outputs: the digital terrain model (DTM), the Digital Surface Model (DSM), and the LiDAR-derived canopy height model (CHM).

In order to differentiate between ground and vegetation points, the original point cloud data were initially filtered using a classification algorithm available in the *groundfilter* function in FUSION/LDV. The *gridsurfacecreate* function was used to generate the 1-meter resolution Digital Terrain Models (DTMs), using the classified ground returns. The *canopymodel* tool was then used to interpolate vegetation points and to generate the Digital Surface Models (DSMs).

The *clipdata* function was applied to obtain normalized heights by subtracting the DTMs elevations from each LiDAR return. Normalized point clouds were subset within the field sample plots of interest using the *polyclipdata* function. Canopy height models (CHM) were derived by subtracting the LiDAR DTM from the DSM. Structure metrics were then computed at plot and stand levels by the *cloudmetrics* and *gridmetrics* functions respectively, using all returns above 1.30 m, at a grid cell resolution of 25 m.

Table 2. Airborne LiDAR survey specifications.

Parameter Value	Parameter Value
Scan angle (°)	±45°
Footprint	0.33 m
Flying altitude	438 m
Swath width	363.11 m
Overlap	100% (50% side-lap)
Scan frequency	300 kHz
Average point density	10 pts m ⁻²

From the point cloud, it is possible to compute many LiDAR metrics, however, it was generated only those metrics that have been often used as candidate predictors for forest attribute modeling in other studies (HUDAK et al., 2006; GARCÍA-GUTIÉRREZ et al., 2015; GÖRGENS et al., 2015; SHIN et al., 2016; SILVA et al., 2016). Therefore, a total of 26 LiDAR metrics calculated from all returns were considered as candidate predictor variables (Table 3).

Table 3. LiDAR-derived structure metrics considered as candidate predictor variables.

Variable	Description	Variable	Description
HMAX	Height maximum	HP25	Height 25th percentile
HMEAN	Height mean	HP30	Height 30th percentile
HMODE	Height mode	HP40	Height 40th percentile
HSD	Height standard deviation	HP50	Height 50th percentile
HVAR	Height variance	HP60	Height 60th percentile
HCV	Height coefficient of variation	HP70	Height 70th percentile
HIQ	Height interquartile distance	HP75	Height 75th percentile
HSKEW	Height skewness	HP80	Height 80th percentile
HKURT	Height kurtosis	HP90	Height 90th percentile
HP01	Height 01th percentile	HP95	Height 95th percentile
HP05	Height 05th percentile	HP99	Height 99th percentile
HP10	Height 10th percentile	CR	Canopy relief ratio (HMEAN – HMIN)/(HMAX – HMIN)
HP20	Height 20th percentile	COV	Canopy cover (% of first returns above 1.30 m)

2.4. Statistical Modeling

2.4.1. Variable Selection

The modeling techniques evaluated in this study to estimate the statistical relationship between plot-level stem total volume and LiDAR metrics fall into two different approaches: Parametric methods (i.e., multivariate linear regression) and non-parametric methods (i.e., machine learning regression). Parametric and non-parametric models have been proven to be useful for developing predictions from LiDAR derived metrics and field estimated forest structural attributes (HUDAK et al., 2006, 2008; LATIFI et al., 2010; SOKAL; ROHLF, 2012; GARCÍA-GUTIÉRREZ et al., 2015; SHIN et al., 2016; SILVA et al., 2017; CAO et al., 2018).

Even though, machine learning algorithms are usually not sensible to collinearity, normality or linearity, in order to obtain a set of predictor variables that could be commonly applied to all the selected modeling methods, it was used two variable selection approaches. First, Pearson's correlation (r) analysis was carried out to identify highly correlated metrics and to exclude redundant predictors ($r > 0.9$) (HUDAK et al., 2012; SILVA et al., 2017). Second, it was implemented Principal component analysis (PCA) to the most relevant LiDAR-derived candidate metrics, to achieve a final set of predictor variables. PCA describes the variation of a set of multivariate data in terms of a set of uncorrelated variables, each of which is a particular linear combination of the original variables. Using PCA, a subset of variables that explain the majority of variation can be selected from a large set of (possibly highly correlated) predictor variables (LI et al., 2008).

PCA was applied over the selected LiDAR metrics for each of the 158 sample plots by the *prcomp* function in R statistical package (R Development Core Team, 2015). A correlation matrix derived from the LiDAR metrics provided the basis for the eigenvalue and eigenvector calculations and for the subsequent determination of the 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, it could be established 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 (SILVA et al., 2016). The first five metrics that were most likely to contribute to the model development were identified by inspecting the eigenvectors in each PC. We then used the metrics with highest loading on the PCs as input variables for every modeling method.

2.4.2. Modeling Development and Assessment

To explore the effect of sample size on prediction accuracies a classical data-splitting approach was adopted. We sampled subsets with different sample sizes from the full dataset where two-thirds of the data was reserved as the training set (estimation) and the remaining one-third as the test set (validation). Subset sample sizes were therefore chosen in increments of 10 from 10% to 100% of the dataset. This procedure resulted in 10 sample sizes, for each sample size we carried out resampling simulations to approximate the sampling distributions of estimators for combinations of modeling methods and sample sizes. The use of simulations is well established in the statistical literature, and theoretical bases for their use can be found in a variety of textbooks (WOLTER, 2007). Five hundred simulations were used to approximate the sampling distributions for each of the examined combinations. For convenience we defined our population as our complete sample, all 158 observations collected in the field. This approach is similar to the approach used by Strunk et al. (2012) to contrast several estimation approaches with LiDAR to demonstrate general trends and issues and is considered indicative of behavior for similar areas and sampling designs. For each simulation we computed and saved its corresponding performance measures. The modeling methods used in this study were:

- i) Ordinary least-squares (OLS) multiple regression. The OLS regression algorithm fits a linear model by minimizing the residual sum of squares between the observed values in the training dataset and the predicted values by the linear model (CUI; GONG, 2018). The set of prospective multiple linear models was calculated using the *lm* linear model function in R environment.
- ii) Random Forest (RF) algorithm. RF is an ensemble classifier that generates a set of numerous individually trained decision trees and combines their results for classification and regression (GLEASON; IM, 2012). The algorithm was implemented through R package *randomForest* (LIAW; WIENER, 2002). RF was adjusted to 1000 ntree, and for the number of variables randomly sampled (mtry) as candidates at each split, we used the default value, which for regression is defined as $p/3$, where p is the number of covariates. For the remaining parameters, we used the default values.
- iii) k -nearest neighbors (k -NN) imputation. k -NN methods work by direct substitution (imputation) of measured values from sample locations (references) for locations for which we desire a prediction (targets). In this strategy, key considerations include the

- distance metric that is used to identify suitable references and the number of references (k) that are used in a single imputation (SHIN et al., 2016). In this study, we examined $k = 1$ neighbors for each of the mentioned distance metrics, in order to keep the original variation in the data (HUDAK et al., 2008). Many imputation methods can be used for associating target and reference observations, we decided to evaluate six different distance metrics for k-NN based approach: raw, euclidean (EU), mahalanobis (MA), most similar neighbor (MSN), most similar neighbor 2 (MSN2), and random forest (RF). Imputations were performed in R using the *yalImpute* package (CROOKSTON; FINLEY, 2008), in combination with the *Random Forest* package (LIAW; WIENER, 2002).
- iv) Support Vector Machine (SVM). The SVM algorithm operate by assuming that each set of inputs will have a unique relation to the response variable and that the grouping and the relation of these predictors to one another is sufficient to identify rules that can be used to predict the response variable from new input sets (REIS et al., 2018). SVM was loaded using the *e1071* package (DIMITRIADOU et al., 2008). Radial Base Function for the Kernel function was selected.
 - v) Artificial neural network (NNT). NNTs are a parallel-distributed information processing system that simulates the work of neurons in the human brain, being able to learn from examples. NNTs are widely used to model complex and non-linear relations between inputs and outputs or to determine patterns in data (DIAMANTOPOULOU, 2012). The neural network in this study was set up with 7 neurons in the input layer (number of variables), 1 neuron in the hidden layer, and 1 neuron in the output layer, corresponding to the estimated volume. The initial weights were set randomly, and the decay parameter was set to 0.1. During the NN learning process, the weights were adjusted to return a result as similar as possible to the training set and to indicate the relative influence of the variables. The NN was implemented in R using the *nnt* package (VENABLES; RIPLEY, 2002).

The performance assessment of the modeling methods along with the effect of the sample size for each simulation run, was computed via three performance measures: the root mean square error (RMSE) (3), which represents the estimate of the standard deviation and sample variance,

thus the lower the RMSE value, the better the model adjustment; the coefficient of determination (R^2) (4), which allows to measure the model degree of explanation by measuring the total proportion of variation to the average; and the Bias (both absolute and relative) (5), a value representing the average of the residuals. The closer to zero the less biased and preferable the adjusted model (SCHNEIDER et al., 2009).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

$$BIAS = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \quad (5)$$

where y_i is the observed value for plot i , \hat{y}_i is the estimated value for plot i and n is the number of plots. Relative RMSE and biases (RMSEr / BIASr) were calculated by dividing the absolute values (Eqs. 1, 3) by the mean of the observed response parameters. It was defined acceptable model precision and accuracy as a relative RMSE and Bias of $\leq 15\%$ to have a model precision and accuracy higher than or equal to the conventional forest inventory standard in fast-growing Eucalyptus plantations in Brazil (SILVA et al., 2017). Wilcoxon–Mann–Whitney test was performed to determine if the differences between the methods and sample sizes were statistically significant (at $p = 0.05$).

2.5. Predictive Maps

Predictive maps of stem total volume at 25 m of spatial resolution were generated based on the combination of the best modeling technique and sample size containing the selected LiDAR metrics according to PCA. Because we have a large number of stands in this study, the stem total volume at the stand level were then presented herein by stand ages of 2–3, 4–5 and 6–7 years.

3. RESULTS

3.1. Predictor Variable Selection

A total of 19 of the 26 lidar metrics showed a very strong correlation ($r > 0.9$). It was retained one of the highly correlated metrics (H99TH), which along with six other remaining metrics not highly correlated ($r \leq 0.9$) were included in PCA analysis. LiDAR metrics that were retained after correlation analysis included HMEAN, HMODE, HCV, HKUR, H25TH, H99TH and COV. The correlation structure of these seven metrics is shown in Table 4. Among these, HMEAN, HMODE, HCV, HKUR, and COV exhibited the highest PC eigenvector loadings (Table 5), which represented the contribution of each LiDAR metric toward the component, and therefore, were used for model development.

Table 4. Pearson correlations among selected LiDAR metrics

<i>r</i>	HMEAN	HMODE	HCV	HKUR	H25TH	H99TH	COV
HMODE	0.66 ***						
HCV	-0.10	-0.02					
HKUR	0.23	0	-0.79 ***				
H25TH	0.67 ***	0.39 **	-0.69 ***	0.54 ***			
H99TH	0.76 ***	0.52 ***	0.53 ***	-0.32 *	0.10		
COV	-0.27	-0.24	-0.07	0.22	-0.23	-0.26	

“***”: p-value < 0.001; “**”: p-value < 0.01; “*”: p-value < 0.05; If there is no *: p-value \geq 0.05.

Table 5. Loadings and eigenvectors for the first five PCs

PCs	Ev	Eigenvectors (Eg)						
		HMEAN	HMODE	HCV	HKUR	H25TH	H99TH	COV
PC1	2.80	0.54	0.42	-0.26	0.26	0.52	0.29	-0.20
PC2	2.47	-0.19	-0.23	-0.55	0.50	0.23	-0.51	0.23
PC3	0.91	0.19	0.14	0.14	0.19	-0.13	0.25	0.90
PC4	0.48	-0.29	0.84	-0.13	-0.21	-0.12	-0.36	0.08
PC5	0.27	0.01	-0.17	-0.14	-0.71	0.58	-0.07	0.31

PC is the given Principal Component; Ev is the eigenvalues for each PC. Bold values indicate the largest contributing LiDAR metric for a given PC.

The first five PCs accounted for 98.9% of the total variance contained in the selected set of seven LiDAR metrics. PC1, PC2, PC3, PC4 and PC5 accounted for 40.0, 35.3, 12.9, 6.8 and 3.8 per cent of the total variance, respectively (Figure 2a). We opted to use the first five PCs to select the best LiDAR metrics because PCs 6–7 explained a less than significant percentage (<2.5 per cent) of the remaining variance. From Figure 2b it was found that three major groups are highlighted for the first two PCs. The first group representing the first principal component (PC1) highlights canopy height variation. Therefore, it has highly correlations with distributional metrics, showing positive loadings by metrics of percentile height (i.e., HMEAN and H25TH) and negative loading of metrics of HCV and COV. While, the second group (PC2) was mainly influenced by the density metrics, and PC3 highlights canopy cover.

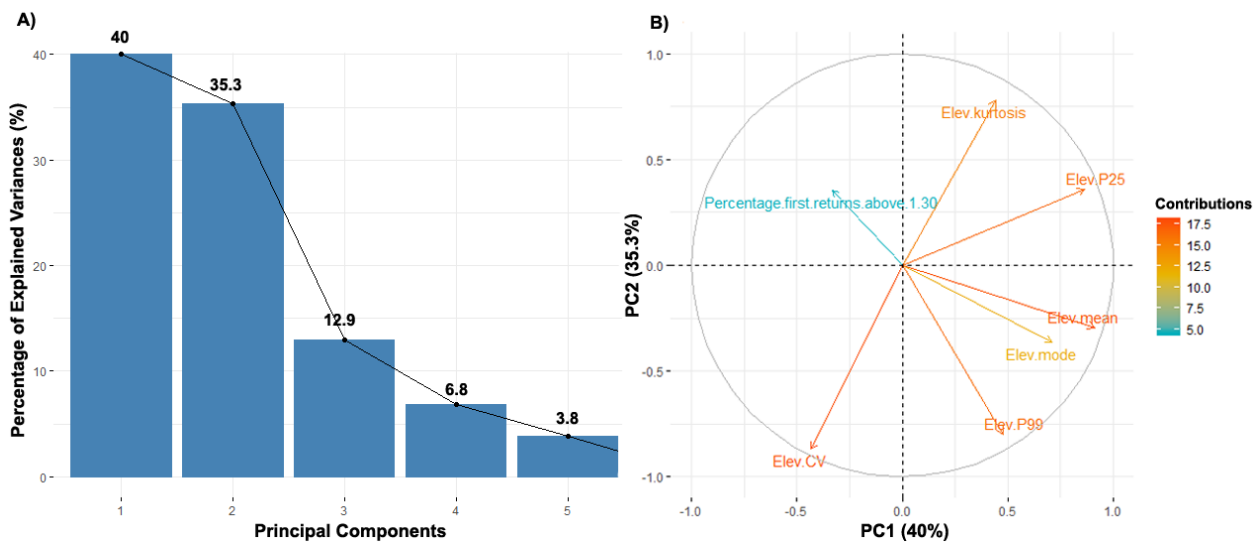


Figure 2. (A) The percentage of variance explained by the five PCs. (B) Projection of the first two PC scores from the selected LiDAR metrics. Different colors represent the visual LiDAR metrics different groups and levels of contribution for each PC.

3.2. Combined Impact of sample size and data modeling

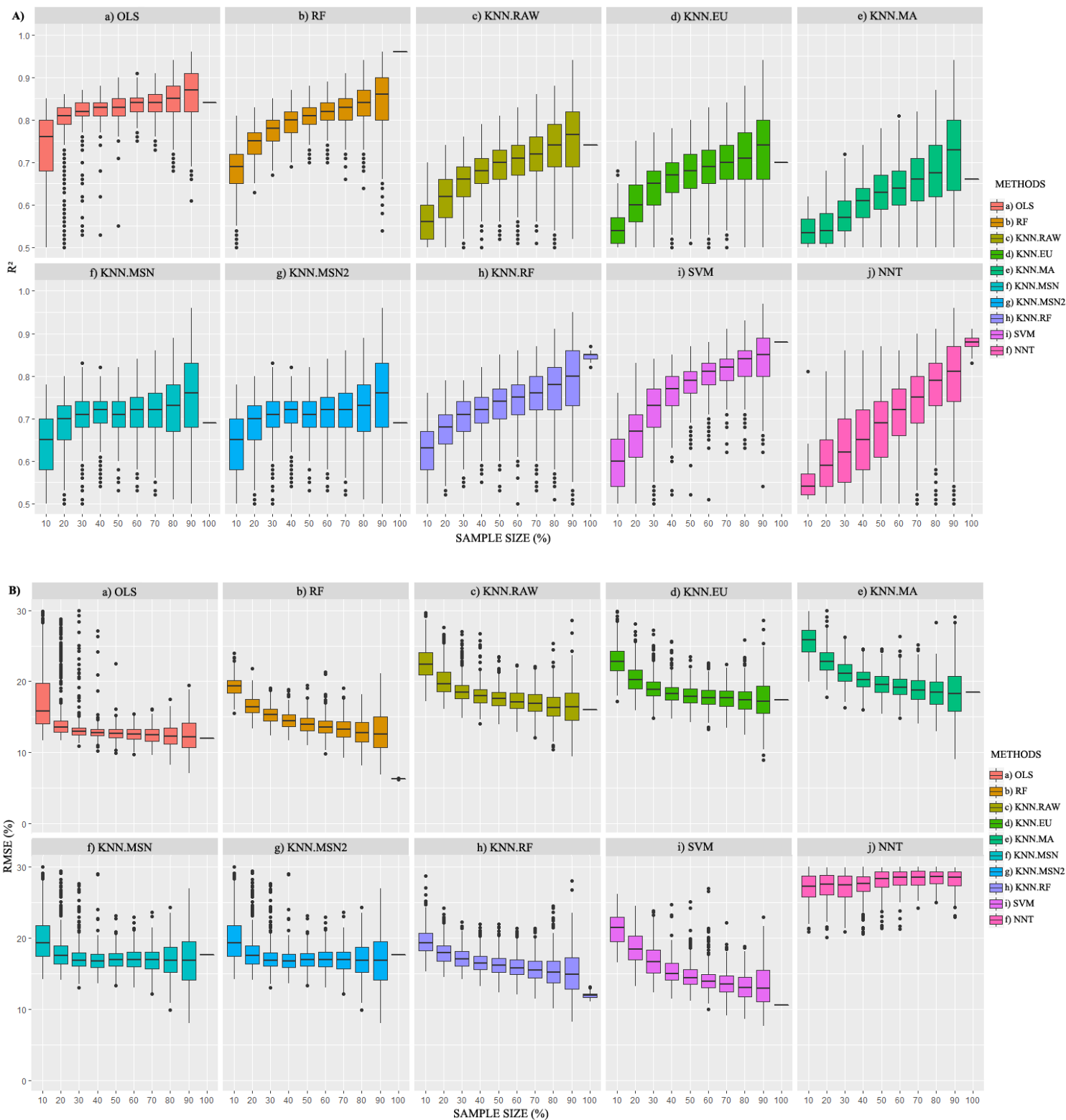
The evaluation of the modeling methods accuracy throughout the sample size was carried out by three performance measures indicators contemplated in Table 6. Different ranges of BIAS, RMSE, and R^2 were obtained according to both the modeling method and number of sample plots. Comparisons across the ten prediction methods indicated that OLS outperformed all other tested

methods, especially when more sample units were available. A relatively stable increase in accuracy and decrease in RMSEr were observed along with increasing sample size in all methods, but only OLS method was able to meet the acceptable model precision criteria (RMSEr and BIASr of $\leq 15\%$) from 20% of sample size. OLS presented R^2 values ranging from 0.77 - 0.85 for 20% to 90% of sample size. RF algorithm was the second best method, reaching the acceptable precision criteria from 40% of sample size. RF method presented R^2 values in the range of 0.80 - 0.84 for 40% to 90% of sample size.

SVM algorithm scored as the third best performance. This algorithm presented similar performance to RF algorithm, showing however lower values in all parameters evaluated. The algorithm was able to meet the acceptable model precision criteria (RMSEr and BIASr of $\leq 15\%$) from 50% of sample size, presenting R^2 values ranging from 0.78 - 0.84 for 50% to 90% of sample size. From the six derivations of the k -NN algorithm tested, none were able to reach the acceptable model precision criteria using less than the full dataset. Only the RF based k -NN approach was able to meet the criteria while using 100% of sample size, presenting a 11.95% in RMSEr and -1.12% in BIASr. The poorer performance in this group was found to be for the k -NN MA algorithm, which presented a RMSEr of 18.48%, BIASr of 1.13% and R^2 of 0.66. Among all the machine learning algorithms, NNT presented the worst performance. The statistics found for NNT demonstrated that this algorithm was not considered applicable to model stem volume using the metrics selected ($R^2= 0.78$; RMSEr= 30.68%; BIASr= 28.06). Slight underestimation was found to be more frequent for most of the modeling methods and sample sizes resulting in negative bias. Only KNN.MA and NNT algorithms always overestimated the forest attribute of interest.

In summary, the best method and sample size combination (minimum sample size) to provide better R^2 values and relatively low number of outliers was found to be OLS method with 40% of the sample size. The use of only 40% of the full dataset combined with the OLS method was able to provide an average of 0.82 for R^2 , 12.95% and -0.07% for RMSEr and BIASr respectively. No significant improvement was found by increasing sample size from 40% to 50%, neither the amount of outliers were much different. Wilcoxon test comparing 40% with full (100%) dataset showed a p value > 0.05 ; hence, 40% and 100% had similar distributions and mean, evidencing no significant difference between them. Although the OLS and RF methods compared presented slightly different performances (OLS reduced RMSEr by 11% over RF), when sample size is 40%, both methods were found to provide satisfactory results (RF40%: $R^2= 0.80$, RMSEr=

14.54%, $\text{BIASr} = -0.13$). Boxplots of the performance measures based on the modeling methods from the 500 prediction simulations are shown in Figure 3.



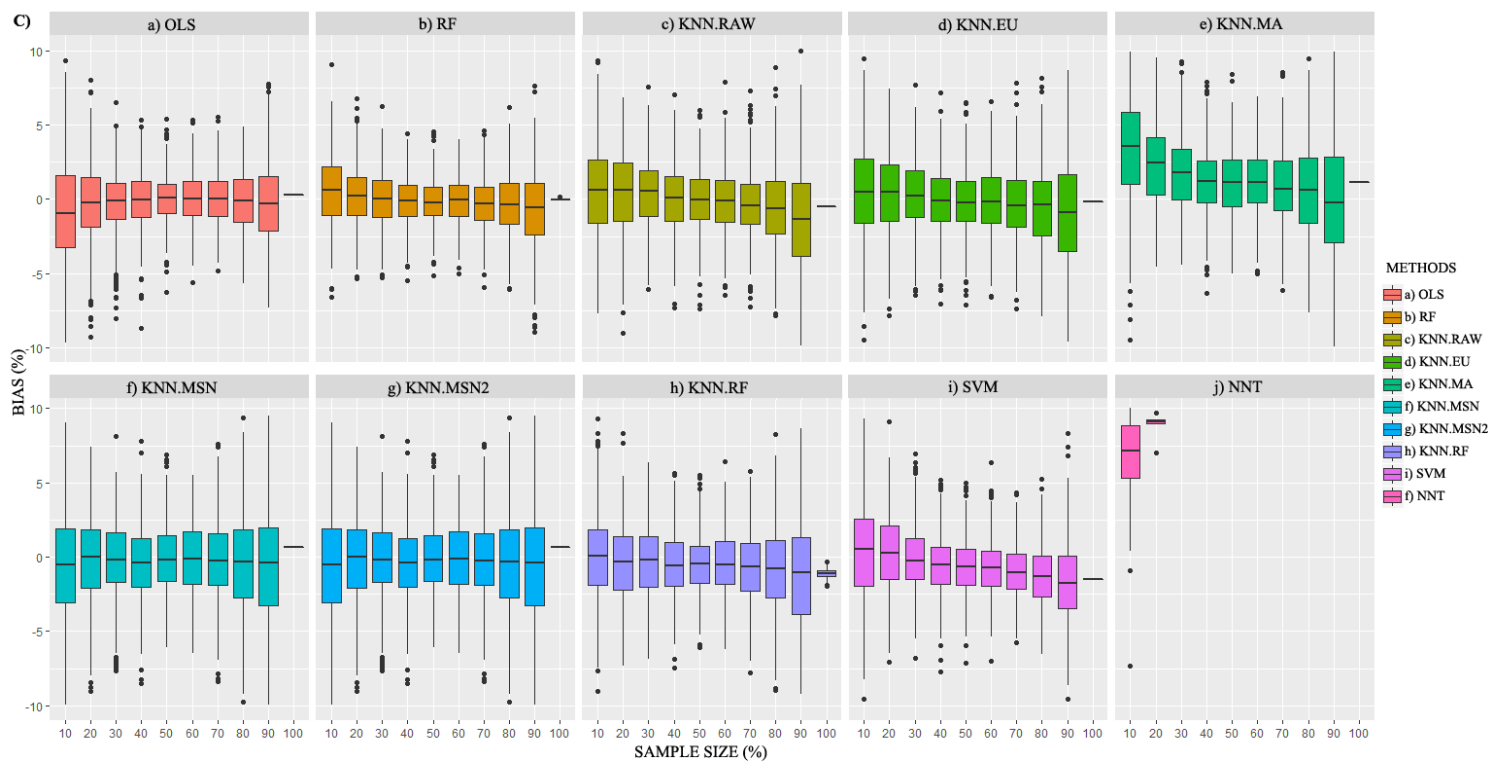


Figure 3. Boxplots of modeling methods performance measures in terms of coefficient of determination - R^2 (A), Relative Root Mean Square Error - RMSE% (B) and Relative Bias - BIAS% (C) derived from the 500 simulations for each different sample size.

Table 6. Modeling methods performance measures (PM) averages and standard deviations in terms of R², Root Mean Square Error (RMSE) and Bias (absolute and relative) derived from the 500 simulations for each different sample size are displayed.

SAMPLE SIZE (%)	PM	LM	RF	KNN.RAW	KNN.EU	KNN.MA	KNN.MSN	KNN.MSN2	KNN.RF	SVM	NNT
10	R ²	0.7 ± 0.13	0.68 ± 0.06	0.51 ± 0.08	0.48 ± 0.08	0.35 ± 0.10	0.61 ± 0.10	0.61 ± 0.10	0.62 ± 0.07	0.55 ± 0.10	0.22 ± 0.15
	RMSE	55.46 ± 15.78	60.15 ± 3.92	70.17 ± 7.34	71.83 ± 7.05	82.1 ± 9.27	63.08 ± 11.98	63.08 ± 11.98	60.84 ± 5.97	66.46 ± 6.49	112.44 ± 29.63
	RMSE(%)	17.8 ± 5.06	19.31 ± 1.27	22.52 ± 2.36	23.05 ± 2.26	26.35 ± 2.97	20.25 ± 3.84	20.25 ± 3.84	19.53 ± 1.92	21.33 ± 2.11	35.63 ± 9.52
	BIAS	-3.1 ± 13.11	1.76 ± 7.19	1.64 ± 9.77	1.55 ± 10.10	11.37 ± 11.97	-2.45 ± 13.55	-2.45 ± 13.55	-0.289 ± 9.05	1.02 ± 9.99	58.82 ± 26.83
	BIAS (%)	-0.99 ± 4.21	0.57 ± 2.31	0.53 ± 3.14	0.5 ± 3.24	3.65 ± 3.84	-0.78 ± 4.35	-0.78 ± 4.35	-0.09 ± 2.91	0.33 ± 3.21	18.66 ± 8.58
	R ²	0.77 ± 0.10	0.75 ± 0.04	0.6 ± 0.07	0.59 ± 0.07	0.49 ± 0.08	0.67 ± 0.08	0.67 ± 0.08	0.68 ± 0.05	0.65 ± 0.09	0.40 ± 0.18
20	RMSE	46.17 ± 11.49	51.62 ± 3.92	62.63 ± 6.83	63.7 ± 6.16	71.85 ± 6.72	56.67 ± 8.66	56.67 ± 8.66	55.98 ± 5.00	57.97 ± 7.41	103.88 ± 21.75
	RMSE(%)	14.83 ± 3.69	16.58 ± 1.28	20.11 ± 2.21	20.45 ± 1.98	23.07 ± 2.15	18.2 ± 2.79	18.2 ± 2.79	17.98 ± 1.62	18.62 ± 2.41	33.06 ± 7.05
	BIAS	-1.12 ± 8.65	0.93 ± 6.10	1.38 ± 8.48	0.99 ± 8.53	6.96 ± 8.79	-0.58 ± 9.35	-0.58 ± 9.35	-1.33 ± 8.25	0.75 ± 8.16	69.51 ± 18.57
	BIAS (%)	-0.36 ± 2.78	0.3 ± 1.96	0.44 ± 2.72	0.32 ± 2.74	2.24 ± 2.82	-0.18 ± 3.00	-0.18 ± 3.00	-0.43 ± 2.65	0.25 ± 2.62	22.13 ± 5.99
	R ²	0.81 ± 0.06	0.78 ± 0.03	0.65 ± 0.06	0.64 ± 0.06	0.55 ± 0.07	0.7 ± 0.06	0.7 ± 0.06	0.7 ± 0.05	0.71 ± 0.08	0.51 ± 0.17
	RMSE	41.67 ± 6.98	47.88 ± 4.11	58.45 ± 5.53	59.48 ± 5.37	66.49 ± 5.84	53.82 ± 6.19	53.82 ± 6.19	53.55 ± 4.45	52.61 ± 7.26	100.91 ± 18.41
30	RMSE(%)	13.38 ± 2.25	15.38 ± 1.35	18.77 ± 1.78	19.1 ± 1.72	21.35 ± 1.86	17.28 ± 2.00	17.28 ± 2.00	17.2 ± 1.44	16.9 ± 2.37	32.17 ± 5.95
	BIAS	-0.49 ± 6.39	0.05 ± 5.53	1.24 ± 7.40	0.73 ± 7.50	5.46 ± 7.84	-0.66 ± 7.75	-0.66 ± 7.75	-1.06 ± 7.41	-0.41 ± 6.93	74.85 ± 16.20
	BIAS (%)	-0.16 ± 2.05	0.02 ± 1.78	0.4 ± 2.38	0.23 ± 2.41	1.75 ± 2.52	-0.21 ± 2.49	-0.21 ± 2.49	-0.34 ± 2.38	-0.13 ± 2.22	23.87 ± 5.23
	R ²	0.82 ± 0.04	0.8 ± 0.03	0.67 ± 0.05	0.66 ± 0.05	0.59 ± 0.07	0.71 ± 0.05	0.71 ± 0.05	0.72 ± 0.05	0.76 ± 0.05	0.59 ± 0.14
	RMSE	40.34 ± 4.91	45.3 ± 3.91	56.35 ± 5.04	57.41 ± 4.88	63.48 ± 5.85	53.01 ± 5.48	53.01 ± 5.48	51.85 ± 4.49	48.32 ± 6.03	97.67 ± 12.51
	RMSE(%)	12.95 ± 1.56	14.54 ± 1.27	18.08 ± 1.62	18.43 ± 1.56	20.37 ± 1.86	17.01 ± 1.76	17.01 ± 1.76	16.64 ± 1.44	15.51 ± 1.97	31.15 ± 4.07
40	BIAS	-0.23 ± 5.70	-0.42 ± 4.98	0.04 ± 6.92	-0.32 ± 6.77	3.84 ± 7.39	-1.12 ± 7.43	-1.12 ± 7.43	-1.63 ± 6.68	-1.67 ± 6.01	76.78 ± 12.61
	BIAS (%)	-0.07 ± 1.83	-0.13 ± 1.60	0.01 ± 2.22	-0.1 ± 2.17	1.23 ± 2.37	-0.36 ± 2.38	-0.36 ± 2.38	-0.52 ± 2.14	-0.53 ± 1.93	24.49 ± 4.08
	R ²	0.83 ± 0.03	0.81 ± 0.03	0.69 ± 0.05	0.68 ± 0.06	0.62 ± 0.07	0.71 ± 0.05	0.71 ± 0.05	0.73 ± 0.05	0.78 ± 0.05	0.64 ± 0.13
	RMSE	39.51 ± 3.15	43.75 ± 4.08	54.83 ± 4.57	56.11 ± 4.53	61.22 ± 5.23	53.05 ± 4.48	53.05 ± 4.48	50.71 ± 4.79	46.07 ± 5.79	98.2 ± 10.27
	RMSE(%)	12.69 ± 0.99	14.05 ± 1.34	17.61 ± 1.46	18.02 ± 1.44	19.66 ± 1.65	17.04 ± 1.43	17.04 ± 1.43	16.29 ± 1.53	14.8 ± 1.90	31.35 ± 3.37
	BIAS	0.23 ± 4.84	-0.28 ± 4.60	-0.14 ± 6.55	-0.56 ± 6.58	3.4 ± 7.17	-0.21 ± 6.99	-0.21 ± 6.99	-1.37 ± 6.28	-2.05 ± 5.56	79.82 ± 10.57
50	BIAS (%)	0.07 ± 1.55	-0.09 ± 1.48	-0.04 ± 2.10	-0.18 ± 2.11	1.09 ± 2.30	-0.07 ± 2.24	-0.07 ± 2.24	-0.44 ± 2.02	-0.65 ± 1.79	25.48 ± 3.44
	R ²	0.84 ± 0.03	0.82 ± 0.03	0.7 ± 0.06	0.68 ± 0.06	0.63 ± 0.07	0.71 ± 0.05	0.71 ± 0.05	0.74 ± 0.05	0.8 ± 0.04	0.69 ± 0.12
	RMSE	39.13 ± 3.02	42.52 ± 4.31	53.97 ± 4.92	55.55 ± 4.79	59.96 ± 5.24	53.19 ± 4.87	53.19 ± 4.87	49.8 ± 4.98	44.09 ± 5.60	98.1 ± 8.99
	RMSE(%)	12.57 ± 0.94	13.66 ± 1.40	17.33 ± 1.55	17.84 ± 1.50	19.25 ± 1.64	17.08 ± 1.56	17.08 ± 1.56	15.99 ± 1.57	14.16 ± 1.84	31.35 ± 2.99
	BIAS	0.17 ± 5.13	-0.22 ± 4.82	-0.19 ± 6.74	-0.34 ± 6.95	3.59 ± 7.04	-0.5 ± 7.59	-0.5 ± 7.59	-1.31 ± 6.49	-2.35 ± 5.54	81.74 ± 9.48
	BIAS (%)	0.05 ± 1.65	-0.07 ± 1.55	-0.06 ± 2.16	-0.11 ± 2.23	1.15 ± 2.26	-0.16 ± 2.43	-0.16 ± 2.43	-0.42 ± 2.08	-0.75 ± 1.78	26.12 ± 3.13
60	R ²	0.84 ± 0.03	0.82 ± 0.04	0.71 ± 0.06	0.69 ± 0.07	0.65 ± 0.08	0.72 ± 0.06	0.72 ± 0.06	0.75 ± 0.06	0.81 ± 0.04	0.72 ± 0.10
	RMSE	38.81 ± 3.92	41.38 ± 4.86	52.98 ± 5.65	54.97 ± 5.36	58.74 ± 6.13	52.74 ± 5.69	52.74 ± 5.69	48.78 ± 5.87	42.39 ± 5.11	98.41 ± 9.63
	RMSE(%)	12.46 ± 1.23	13.28 ± 1.57	17 ± 1.77	17.64 ± 1.67	18.85 ± 1.92	16.93 ± 1.82	16.93 ± 1.82	15.66 ± 1.85	13.61 ± 1.65	31.45 ± 3.18
	BIAS	0.01 ± 5.46	-1.01 ± 5.37	-1.11 ± 7.12	-1.05 ± 7.61	2.62 ± 7.74	-0.98 ± 8.49	-0.98 ± 8.49	-1.92 ± 7.31	-3.14 ± 5.54	83.62 ± 8.84
	BIAS (%)	0 ± 1.75	-0.32 ± 1.72	-0.36 ± 2.29	-0.34 ± 2.44	0.84 ± 2.48	-0.31 ± 2.72	-0.31 ± 2.72	-0.61 ± 2.35	-1 ± 1.78	26.73 ± 2.93
	R ²	0.84 ± 0.04	0.84 ± 0.05	0.73 ± 0.07	0.71 ± 0.08	0.66 ± 0.09	0.72 ± 0.08	0.72 ± 0.08	0.77 ± 0.08	0.83 ± 0.05	0.76 ± 0.11
70	RMSE	38.42 ± 5.20	40.19 ± 6.02	51.39 ± 6.63	54.16 ± 6.38	57.44 ± 7.05	52.86 ± 7.68	52.86 ± 7.68	47.82 ± 7.44	41.22 ± 6.09	97.87 ± 8.29
	RMSE(%)	12.35 ± 1.64	12.93 ± 1.93	16.52 ± 2.08	17.41 ± 1.98	18.47 ± 2.22	17 ± 2.47	17 ± 2.47	15.38 ± 2.36	13.26 ± 1.96	31.33 ± 2.78
	BIAS	-0.35 ± 6.37	-1.03 ± 6.30	-1.83 ± 8.41	-1.55 ± 8.91	1.78 ± 9.28	-1.16 ± 9.90	-1.16 ± 9.90	-2.38 ± 8.93	-3.93 ± 6.60	84.39 ± 8.67
	BIAS (%)	-0.11 ± 2.05	-0.32 ± 2.03	-0.59 ± 2.71	-0.5 ± 2.87	0.57 ± 2.98	-0.37 ± 3.19	-0.37 ± 3.19	-0.76 ± 2.87	-1.26 ± 2.12	27.02 ± 2.91
	R ²	0.85 ± 0.06	0.84 ± 0.08	0.74 ± 0.11	0.72 ± 0.11	0.68 ± 0.15	0.73 ± 0.12	0.73 ± 0.12	0.78 ± 0.11	0.84 ± 0.07	0.78 ± 0.13
	RMSE	38.77 ± 7.97	40.42 ± 9.04	51.3 ± 9.01	54.41 ± 9.30	57.12 ± 11.44	52.96 ± 11.95	52.96 ± 11.95	47.3 ± 10.16	41.6 ± 9.25	98.41 ± 9.68
90	RMSE(%)	12.45 ± 2.44	12.98 ± 2.83	16.48 ± 2.77	17.48 ± 2.87	18.35 ± 3.74	17.01 ± 3.74	17.01 ± 3.74	15.19 ± 3.16	13.37 ± 2.91	31.53 ± 3.36
	BIAS	-1.01 ± 8.32	-2.2 ± 8.22	-4.13 ± 11.49	-3.06 ± 12.42	-0.25 ± 13.37	-1.79 ± 12.69	-1.79 ± 12.69	-3.57 ± 11.41	-5.13 ± 8.74	84.49 ± 10.15
	BIAS (%)	-0.32 ± 2.67	-0.68 ± 2.63	-1.31 ± 3.69	-0.96 ± 3.99	-0.07 ± 4.30	-0.56 ± 4.07	-0.56 ± 4.07	-1.14 ± 3.67	-1.62 ± 2.80	27.08 ± 3.52

3.3 Predictive Maps

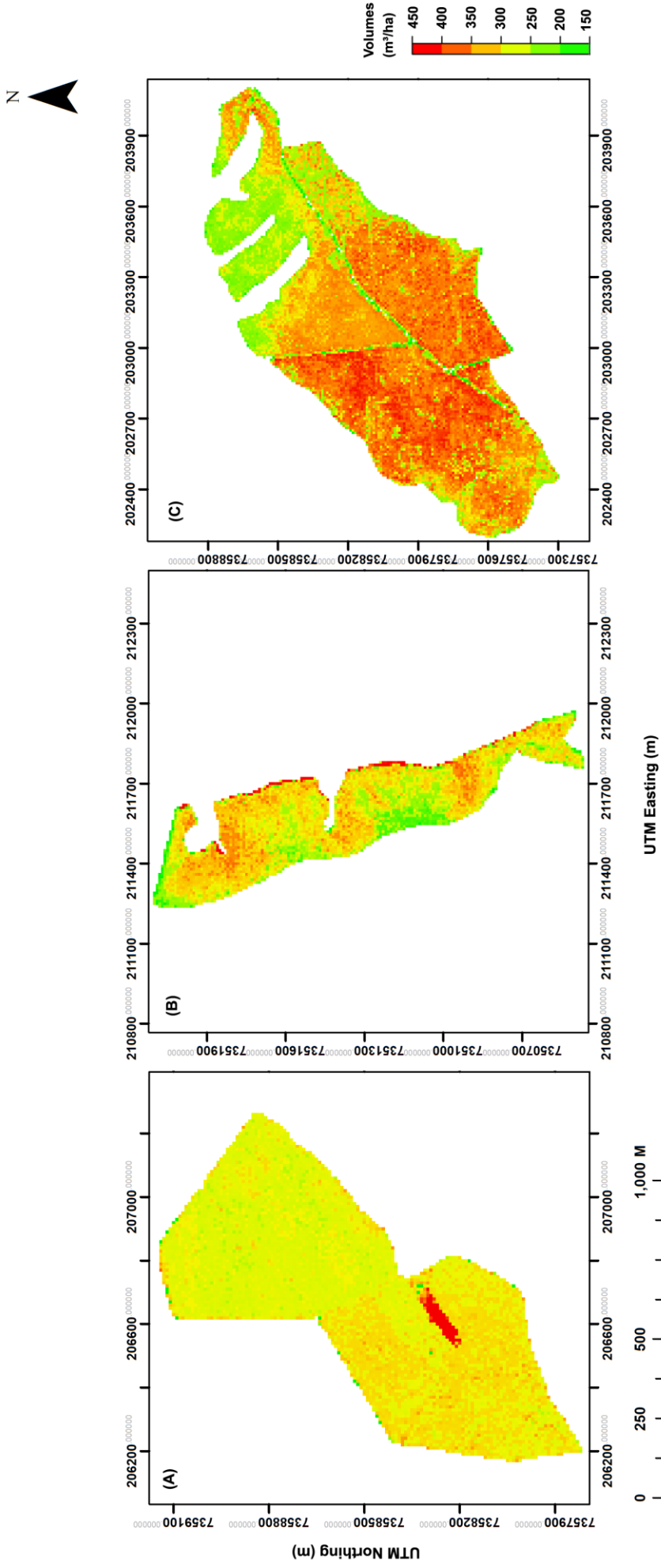


Figure 4. Predictive volume maps of *Eucalyptus* spp. plantations from LiDAR data at the stand-level obtained from the selected LiDAR metrics according to PCA, and the combination of the OLS modeling method using 40% of the sample size with the representative stand of early (i.e., 2–3 years) (A), intermediate (i.e., 4–5 years) (B) and advanced-stages of development (i.e., 6–7 years) (C).

4. DISCUSSION

LiDAR has shown to be a powerful technology for forest inventory around the world, yet studies exploring the application of airborne LiDAR technology for Brazilian Eucalyptus plantations is relatively new (CARVALHO et al., 2015; SILVA et al., 2017). On examination of the trends observed in previous studies, that have employed a wide range of modeling methods for forest attribute estimation and reported results representing varying accuracies, it is clear that appropriate selection of methods is paramount for attaining best prediction results (FASSNACHT et al., 2014; GÖRGENS et al., 2015; GARCÍA-GUTIÉRREZ et al., 2015; SHIN et al., 2016; XU et al., 2018). Taking this research one step further, we did an investigation on how combined influence of sample size and different modeling techniques affect the overall prediction accuracy of forest attributes and demonstrated the potential of reduced sample size in procuring accurate prediction results. This way, we offer expository insights and recommendations to forest managers and modelers for enhancing their model selection, data collection and decision-making strategies and thereby, assist them in optimizing cost, energy, labor and overall efficiency of the forest inventory operations.

For reducing model complexity and boosting overall prediction accuracy it is imperative to select fitting, yet minimal number of, parameters for developing predictive models by means of variable selection approaches (SILVA et al., 2016; GREGORUTTI al., 2017); this task, however, gets more challenging when highly correlated predictors are present. Application of dual variable selection approaches - Pearson's correlation analysis and PCA - proved beneficial in our case and allowed us to shortlist the five major variables - HMEAN, HCV, HMODE, HKUR, and COV – from a total of 26 LiDAR metrics. These 5 variables, which were used for model development, accounted for 98.9% of the total variation contained in the pre-selected set of LIDAR metrics; recent studies done on Eucalyptus plantations which had applied only PCA for variable selection, found similar total variance contained in the selected set of LiDAR metrics (97.7%) (SILVA et al., 2016).

There was a significant relationship between field-based volume estimates and LiDAR-derived metrics selected from the PCA analysis. The selected metrics from the PCA analysis was consistent with previous studies that also found that mean height had the largest absolute correlation with the first principle component, coefficient variation of height had the largest absolute correlation with the second principle component, and canopy cover had the largest

absolute correlation with the third principal component (LI et al., 2008). Models using these three first principal components likely capture the fundamental allometric relationships between volumes and heights, as seen in results from large-footprint SLICER data (LEFSKY et al., 2005), in which mean height, canopy cover, and height variability were found to explain the most of variability in forest physical characteristics. Tesfamichael et al. (2010) and Packalen et al. (2011), also found that metrics such as HMEAN and HCV have shown to be effective predictors of forest attributes, such as stem volume, height, basal area, and aboveground carbon in *Eucalyptus* spp. plantations. The biological basis behind these results is due to the ecological and biomechanical links between canopy vertical structure and forest stand structure parameters. From the perspective of tree form and function development, there is usually a connection between the differences in vertical canopy structure and differences in forest volume both through forest succession and across areas with contrasting environmental conditions (LI et al., 2008).

From our results, it was evident that algorithm performance was sensitive to sampling size and the level of influence varied from one algorithm to another. On placing constraints (<15%) for RMSE values, only 4 models – SVM, k-NN, RF and OLS – were found to be feasible for making predictions. Among these, k-NN was further marked as impracticable, as in this case only a sample size greater than 90% satisfied the requirement. In case of OLS, a sample size greater than 20% fell within the RMSE threshold; this might be because of the low level of multicollinearity within the model. Whereas, for RF and SVM, the ideal sample sizes were above 40% and 50% respectively. In terms of bias, we noticed all the models to fall within the maximum set limit, which was 15%. With respect to R^2 values, OLS proved to be the best among the given modeling methods, followed by RF and SVM respectively. Range of R^2 values was also comparatively higher for OLS – 0.77 to 0.85 – with reference to RF – 0.75 to 0.84 - and SVM – 0.65 to 0.84 - methods. Also, we noticed that the increase in R^2 values with increasing sample size was more evident in case of RF and SVM. However, this pattern was expected, considering the fact that the non-parametric models learn their functional form from the training data (FASSNACHT et al., 2014; XU et al., 2018) that means, the higher the sample size, the better their prediction accuracy will be. This dependence on sample size, might be the reason several other non-parametric algorithms failed to provide satisfactory results in our case, where field plots considered were limited (SHIN et al., 2016; NOI; KAPPAS, 2017).

Even though, for OLS, sample size above 20% met the chosen criteria, high levels of outliers were observed in this case as well as when the sample size was 30%. A former study done by Gobakken and Næsset (2008) had come up with the generalization that average standard deviation tends to increase with reduction of sample size, which very well matches with our findings. However, sample size on reaching 40% showed significant reduction in amount of outliers and a notable increase in R^2 values. On further increase to 50% sample size, there occurred not much difference in the outliers count or the R^2 values. Additionally, by performing wilcox test ($P > 0.05$), we confirmed that 40% and 100% were not significantly different in terms of distribution and mean. When sample size was 40%, RF also gave satisfactory results, even though R^2 value was slightly lower (0.8) compared to OLS (0.82). Based on our results and core objective – which was to find the minimum sample size required for attribute estimation – we inferred the best combination to be linear regression (OLS) model with sample size of 40%, followed by random forests (RF) method with identical sample size value.

Since, there existed no extensive studies that accounted for the combined influence of modeling methods and sample size, evaluating the accuracy of our model in regard to established and identical workflows were near-impossible. Nevertheless, on comparison with studies that have evaluated the influence of sample size and modeling methods on a discrete basis, we noticed our obtained trends and accuracies of the high performing models to be quite comparable with the inferences made by other studies. A recent study by Sterenczak et al. (2018) had investigated the influence of number and size of sample plots, as well effect of a single selection, on modeling growing stock volume (GSV) - of a Scots pine (*Pinus sylvestris* L.) dominated forest in Poland with 900 available study plots - using airborne LiDAR data. Based on their three major findings - i) influence of number of sample plots on the accuracy of GSV estimation above 400 sample plots was nominal, ii) number of sample plot size and estimation accuracy revealed an inverse relation, irrespective of the number of plots considered and iii) single selection doesn't have any impact when plots considered were above 400 - the authors concluded that it is possible to reduce the number of ground sample plots by almost one-third and still retain reasonable accuracy and precision levels, even when the sample plot area are relatively small. This was highly evident in our case as well – for sample size less than 40%. Another study done by Shin et al. (2016) compared the performance of 7 modeling methods - most similar neighbor (MSN) imputation, gradient nearest neighbor (GNN) imputation, Random Forest (RF)-based imputation, BestNN

imputation, ordinary least square (OLS) regression, spatial linear model (SLM), and geographically weighted regression (GWR) – for predicting 5 forest attributes, namely basal area (BA), stem volume (VOL), Lorey's height (LOR), quadratic mean diameter (QMD), and tree density (DEN), from airborne LiDAR metrics. Contrary to our results, in this case, the authors were not able to come up with a single modeling method that always performed superior to the others in prediction of the forest attributes; nonetheless, OLS and SLM gave best results in terms of RMSE values in maximum number of cases. The authors also stressed on the importance of selecting a modeling technique for forest attributes based on the objectives, conditions, and scales that is considered; the optimal ranges for training data was found to be in the range of 100-150 for point prediction and 200-250 for the prediction of the whole population set.

The major takeaway from our study is that with LiDAR data of only 40% of the total field plots, we are able to make accurate predictions, given that the right modeling technique in employed. This, when translated into large-scale area projects, means savings of huge amount of money and faster processing with high accuracy. With the same amount of time, we can get more things done or maybe even utilize the available budgets for performing surveys at increased frequency. Future studies can even narrow these results by taking reducing the intervals in sample size (that is instead of the 10% used here, perhaps use 5% or even 1%) and repeating the same process. Results also highlight that multiple modeling methods work well on predictions and depending on the level of data in hand these methods can be selected. However, it is incumbent on the modelers to keep in mind the limitations of each algorithms before applying them. For example, for applying linear regression models, assumptions of linear relationship, homoscedasticity, etc. needs to be met and this is not always true in case of several plantation data; in a lot of cases, since data is collected from a copious amount of sources and often has data of same location for multiple dates, a data hierarchy tend to exists and in this case, a mixed effects model need to be used to account for the random effects happening within the models (CRECENTE-CAMPO et al., 2010; WANG et al., 2007; HAO et al., 2015; DE SOUZA VISMARA et al., 2015). Therefore, a minimum knowledge of the study site and data exploratory analysis is recommended before making the method selection; one should also acknowledge the errors associated with field measurements, ALS data acquisition and data processing steps while interpreting the model results. Previous studies have reported minimum sample size required to vary with respect to the attribute and tree species under consideration. For instance, a study

undertaken by Kallio et al. (2010) observed the accuracy of estimated *Picea abies* volumes at the forest stand level to show no decrease until the number of plots was reduced to below 200. Whereas, for the case of other deciduous tree species, the volume estimation accuracy plummeted with gradual decrease in number of sample plots. Also, more often than not, limited field data and/or acquired LiDAR data quality place additional constraints on complementing studies that intend to evaluate the minimum sample size required for estimating accuracy of forest attributes using ALS metrics (STERENCZAK et al., 2018).

Based on our results it is seen that different algorithms perform differently to varying sample sizes. However, why this was the case was outside the scope of our study. Future studies focusing on this aspect would be able to throw more light on ideas presented. Apart from the 10 methods used here, other less-prominent methods, yet advancing methods like deep learning can also be tried in future works (GUAN et al., 2015; ZHANG et al., 2016). Here, we tested the combined influence of only sample size and machine learning algorithms, nonetheless, influence of additional features – such as plot size, LiDAR pulse density, GPS location errors, etc. – would also be interesting and helpful to the research community (STRUNK et al., 2012; HERNÁNDEZ-STEFANONI et al., 2018; FASSNACHT et al., 2014). Another possible extension of this study is by evaluating how our approach performs in case of study sites having different area sizes – i.e., very low sample size to very high – and to see if the combined influence value remains stable throughout or is having any relation with the area considered. This can be also tested for other forest plantation species as well as trees of different age groups and identify patterns existing between tree characteristics and algorithm performances. Another thing to keep in mind, is the cost associated with LiDAR, which makes this approach economically feasible for only large study areas (TILLEY et al., 2004; SILVA et al., 2017). Also, updating data over time using LiDAR can be perceived as a hurdle for the same reason. However, if we are willing to adopt a different perspective, that views the potential reduction of field work cost as a compensation for ALS data acquisition, then multiplying the ALS data collection frequency can be treated as a reasonable initiative. If not, data fusion techniques – that integrate LiDAR with other more affordable methods such as UAV remote sensing or other freely available cutting-edge technologies – can be deemed as alternative strategies (SANKEY et al., 2018; YANG; CHEN, 2015; HUO et al., 2018). For translating this framework from the research to the operational arena requires additional work, especially that tests its applicability on multiple sites and verifying stability in results, which needs

more investment in terms of field work and analysis. Even so, the expected benefits, that comes in form of reduced inventory cost, will be a huge leap for the forest management sector.

5. CONCLUSIONS

Improving Eucalyptus industrial plantations productivity requires the development of strategies that are tailored to its unique characteristics. Which reinforces the importance of a framework with more robust and accurate techniques that consider auxiliary data in the process of estimating stem total volume. In this study, we evaluated the impacts of different modeling methods and sample size on the accuracy of volume estimates predicted from LiDAR data in a Eucalyptus forest plantation in Brazil.

Our results showed that the precision of LiDAR derived stem total volume estimates was considerably impacted by the prediction method while varying sample sizes. Higher levels of accuracy were attained employing linear regression model (OLS), which was able to provide comparable results to the traditional forest inventory approaches using only 40% of the total field plots, followed by the random forest (RF) algorithm with identical sample size value. The precision of the combined impact of sample size and modeling methods was demonstrated through a relative RMSE and bias less than 15%, which is equal to or less than the level of error that is traditionally accepted in a conventional field inventory.

The methods used in this study formulate a framework for integrating field and LiDAR data, highlighting the importance of sample size for volume estimates. The major takeaway from our study indicates that collecting larger field reference data is not necessarily the most effective option for improving the accuracy of volume estimates. Thus, this study should assist the selection of an optimal sample size that minimizes estimation errors, processing time, plot establishment costs, and consequently increases the monitoring and managing efficiency in Eucalyptus stands.

Future directions for this research include the use of a larger number of datasets that tests additional features (i.e. plot size, LiDAR pulse density, GPS location errors); integrating multi-sensor data fusion approaches (i.e. terrestrial or UAV LiDAR, radar); and estimating forest attributes at an individual tree level. Testing advancing methods like deep learning would also be a further possible extension of our work. Additionally, the development of further studies to increase our understanding of the statistical modelling methods set-up role in the volume estimation of this forest type would be able to throw more light on ideas presented herein.

We hope that the findings from our study give more credibility and encouragement for respective specialists to pursue research in directions that will ultimately results in development of site-independent ALS data based models for predicting a wide range of forest attributes.

6. REFERENCES

- ALVARES, C. A et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v.22, p.711-728, 2013.
- AHMED, O.S. et al. Characterizing stand-level forest canopy cover and height using landsat time series, samples of airborne lidar, and the random forest algorithm”. **ISPRS J. Photogramm. Remote Sensing**, v. 101, p. 89-101, 2015.
- BREIMAN, L., **Random forests**. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.
- BENNETT, B. et al. Characterising performance of environmental models. **Environmental Modelling & Software**, 40, pp. 1-20, 2013.
- BROTTO L. et al. **Planted forests in emerging economies: Best practices for sustainable and responsible investments**. CIFOR, 2016.
- CARVALHO S. et al. Predição do volume de árvores integrando LiDAR e Geoestatística. **Scien For** 43(107): 627-637, 2015.
- CAO L. et al. Integrating Airborne LiDAR and Optical Data to Estimate Forest Aboveground Biomass in Arid and Semi-Arid Regions of China. **Remote Sensing**, v. 30, n. 10, p. 532, 2018.
- CHAVE, J. et al. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. **Oecologia**, v. 145, p. 87–99, 2005.
- CRECENTE-CAMPO, F. et al. A generalized nonlinear mixed-effects height–diameter model for *Eucalyptus globulus* L. in northwestern Spain. **Forest Ecology and Management**, v. 259, n. 5, p. 943-952, 2010.
- CROOKSTON N.L.; FINLEY A.O. yaImpute: An R Package for kNN Imputation. **J Stat Soft** v. 23, n. 10, p. 1-16, 2008.
- DE SOUZA VISMARA, E. et al. Linear mixed-effects models and calibration applied to volume models in two rotations of *Eucalyptus grandis* plantations. **Canadian Journal of Forest Research**, v. 46, n. 1, p. 132-141, 2015.
- DOS REIS, A. A. et al. Spatial prediction of basal area and volume in *Eucalyptus* stands using Landsat TM data: an assessment of prediction methods. **New Zealand Journal of Forestry Science**, v. 48, n. 1, p. 1–17, 2018.
- FAYAD, I. et al. Aboveground biomass mapping in French Guiana by combining remote sensing, forest inventories and environmental data. **International Journal of Applied Earth Observation and Geoinformation**, v. 52, p.502-514, 2016.

FALKOWSKI M.J. et al. The influence of conifer forest canopy cover on the accuracy of two individual tree measurement algorithms using LiDAR data. **Canadian Journal of Remote Sensing**, v. 34, n. 2, p. 1–13, 2008.

FALKOWSKI M.J. et al. Landscape-scale parameterization of a tree-level forest growth model: a k-NN imputation approach incorporating LiDAR data. **Canadian Journal of Remote Sensing** v. 40, p. 184-199, 2010.

FASSNACHT, F. E. et al. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. **Remote Sensing of Environment**, v. 154, p. 102–114, 2014.

FAO (Food and Agriculture Organization of the United Nations). **Global Forest Resources Assessment 2015: how are the world's forests changing?** Rome, 2015. 9 p.

FRIEDMAN J. , et al. The elements of statistical learning Springer Series in Statistics, Springer, Berlin, 2001.

FRITSCH, S.; GUENTHER, F. **Package ‘neuralnet’**. The Comprehensive R Archive Network, 2016.

GAO T. et al. Timber production assessment of a plantation forest: An integrated framework with field-based inventory, multi-source remote sensing data and forest management history. **International Journal of Applied Earth Observation and Geoinformation**, v. 1, n. 52, p. 155-65, 2016.

GARCÍA-GUTIÉRREZ, J. et al. A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. **Neurocomputing**, v. 167, p. 24–31, 2015.

GREAVES, H.E. et al. High-resolution mapping of aboveground shrub biomass in Arctic tundra using airborne lidar and imagery. **Remote sensing of environment**, v. 184, p 361-373, 2016.

GREGORUTTI, B. M. et al. Correlation and variable importance in random forests. **Statistics and Computing**, v. 27, n. 3, p. 659-678, 2017.

GOBAKKEN, T.; NÆSSET, E.. Assessing effects of laser point density, ground sampling intensity, and field sample plot size on biophysical stand properties derived from airborne laser scanner data. **Canadian Journal of Forest Research**, v. 38, n. 5, p. 1095-1109, 2008.

GONZÁLEZ-GARCÍA et al. Dynamic growth and yield model including environmental factors for Eucalyptus nitens (Deane & Maiden) Maiden short rotation woody crops in Northwest Spain. **New Forests**, v. 46, n. 3 ,p. 387-407, 2015.

GÖRGENS, E. B. et al. A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. **Computers and Electronics in Agriculture**, v. 116, p. 221–227, 2015.

GUAN, H. et al. Deep learning-based tree classification using mobile LiDAR data. **Remote Sensing**. Lett. v. 6, p. 864–873, 2015.

HAYASHI R. et al. Evaluation of alternative methods for using LiDAR to predict aboveground biomass in mixed species and structurally complex forests in northeastern North America.

Mathematical and Computational Forestry and Natural-Resource Sciences, v. 7, n. 2, p. 49-62, 2015.

HAO, P. et al. Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA. **Remote Sensing**, v. 7, n. 5, p. 5347-5369, 2015.

HERNÁNDEZ-STEFANONI, J. et al. Effects of Sample Plot Size and GPS Location Errors on Aboveground Biomass Estimates from LiDAR in Tropical Dry Forests. **Remote Sensing**, v. 10, n. 10, p. 1586, 2018.

HUDAK, A. T. et al. Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. **Canadian Journal of Remote Sensing**, v. 32, n. 2, p. 126–138, 2006.

HUDAK, A.T. et al. Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. **Remote Sensing Environment**. 123, 25–40, 2012.

HUDAK, A.T. et al. Imputing forest structure attributes from stand inventory and remotely sensed data in Western Oregon, USA. **Forest Science**, v. 60, n. 2, p. 253–269, 2014.

HUO, L.Z. et al. Supervised spatial classification of multispectral LiDAR data in urban areas. **PLoS ONE** 13, 2018.

KALLIO E. et al. Effect of sampling intensity on the accuracy of species-specific volume estimates derived with aerial data: A case study on five privately owned forest holdings. Proceedings of: **10th International Conference on LiDAR Applications for Assessing Forest Ecosystems**, Freiburg, Germany: 169-178, 2010.

KUHN, M. Caret package. **Journal of statistical software**, v. 28, n. 5, p.1-26, 2008.

KÖPPEN, W.; GEIGER, R. *Klimate der Erde*. Gotha: VerlagJustus Perthes. Wall-map 150cmx200cm. 1928

LAURIN, G.V. et al. Biodiversity mapping in a tropical West African forest with airborne hyperspectral data. **PLoS One**, v. 9, n. 6, 2014.

LATIFI, H. et al. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: Application of multiple optical/LiDAR-derived predictors. **Forestry**, v. 83, n. 4, p. 395–407, 2010.

LI, Y. et al. A comparison of statistical methods for estimating forest biomass from light detection and ranging data. **Western Journal of Applied Forestry**, v. 23, n. 4, pp. 223-231, 2008.

LEFSKY, M. A. et al. Patterns of covariance between forest stand and canopy structure in the Pacific Northwest. **Remote Sensing of Environment**, v. 95, n. 4, p. 517–531, 2005.

MCGAUGHEY R. J. M. FUSION/LDV: Software for LIDAR data analysis and visualization, Version 3.6. USDA Forest Service, Pacific Northwest Research Station (Seattle, WA), 2016.

MONTAGHI, A. et al. Airborne laser scanning of forest resources: an overview of research in Italy as a commentary case study”. **International Journal of Applied Earth Observation and Geoinformation**, v. 23, p. 288-300, 2013.

- MONTGOMERY, D. C. et al. **Introduction to linear regression analysis**. Hoboken: John Wiley & Sons, Inc, 2006.
- MORGENROTH, J.; VISSER, R. Uptake and barriers to the use of geospatial technologies in forest management. **New Zealand Journal of Forestry Science**, v. 43, p. 1–16, 2013.
- NAKAJIMA T. et al. A method to maximise forest profitability through optimal rotation period selection under various economic, site and silvicultural conditions. **New Zealand Journal of Forestry Science**. v. 47, n. 1, p. 4, 2017.
- NÆSSET, E. et al. Laser scanning of forest resources: the Nordic experience. **Scandinavian Journal of Forest Research**, v. 19, p. 482–499, 2014.
- NOI, T. P.; KAPPAS, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. **Sensors**, v. 18, n. 1, p. 18, 2018.
- PENNER, M. et al. Parametric vs. nonparametric LiDAR models for operational forest inventory in boreal Ontario. **Canadian Journal of Remote Sensing**, v. 39, n. 5, p. 426–443, 2013.
- R Core Team. R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**: Vienna, Austria, 2015. Available online: <https://www.r-project.org/> (access 15 Feb 2018).
- RACINE, E.B. et al. Estimating forest stand age from LiDAR-derived predictors and nearest neighbour imputation. **Forest Science**, v. 60, n. 1, p. 128–136, 2014.
- RETSLAFF, F. A et al. Curvas de sítio e relações hipsométricas para *Eucalyptus grandis* na região dos Campos Gerais, Paraná. **Cerne**, v. 21, n. 2, p. 219-225, 2015.
- SANKEY, T.T., et al. UAV hyperspectral and lidar data and their fusion for arid and semi-arid land vegetation monitoring. **Remote Sensing in Ecology and Conservation**, v. 4, n. 1, p.20-33, 2018.
- ROCKWOOD D.L et al. Energy product options for Eucalyptus species grown as short rotation woody crops. **International Journal of Molecular Sciences**. v. 9, n. 8, p. 1361-78, 2008.
- SCOLFORO, J. R. S.; MELLO, J. M. **Inventário florestal**. Lavras: FAEPE, 2006. 561 p.
- SCHNEIDER, P. R.; SCHNEIDER, P. S. P.; SOUZA, C. A. **Análise de regressão aplicada à engenharia florestal**. Santa Maria: Facos, 2009, 294 p.
- SHIN, J. et al. Comparing Modeling Methods for Predicting Forest Attributes Using LiDAR Metrics and Ground Measurements. **Canadian Journal of Remote Sensing**, v. 42, n. 6, p. 739–765, 2016.
- SILVA, C. A. et al. Mapping aboveground carbon stocks using LiDAR data in *Eucalyptus* spp . plantations in the state of São Paulo. **Scientia Forestalis**, v. 42, p. 591–604, 2014.
- SILVA, C. A. et al. Imputation of individual Longleaf Pine (*Pinus palustris* Mill.) tree attributes from field and LiDAR data. **Canadian Journal of Remote Sensing**, v. 42, n. 5, p. 554-573, 2016.

SILVA, C. A. et al. Predicting stem total and assortment volumes in an industrial *Pinus taeda* L. Forest plantation using airborne laser scanning data and random forest. **Forests**, v. 8, n. 7, p. 254-267, 2017

SILVA, C. A. et al. Estimating stand height and tree density in pinus taeda plantations using in-situ data, airborne LiDAR and k-nearest neighbor imputation. **Anais da Academia Brasileira de Ciencias**, v. 90, n. 1, p. 295–309, 2018.

SOKAL, R.; ROHLF, F. **Biometry** (4th edn). WH Freeman, NY, USA, 2012. 937 p.

STEREŃCZAK, K. et al. The influence of number and size of sample plots on modelling growing stock volume based on Airborne Laser Scanning. **Drewno: prace naukowe, doniesienia, komunikaty**, v. 61, 2018.

STRUNK, J. et al. Effects of lidar pulse density and sample size on a model-assisted approach to estimate forest inventory variables. **Canadian Journal of Remote Sensing**, v. 38, n. 5, p. 644–654, 2012.

YANG, B.; CHEN, C. Automatic registration of UAV-borne sequent images and LiDAR data. **ISPRS J. Photogrammetry and Remote Sensing**, v. 101, p. 262–274, 2015.

PACKALÉN, P. et al. 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. **Annals of Forest Science**, v. 68, n. 6, p. 1085–1092, 2011.

TESFAMICHAEL, S. G. et al. Estimating plot-level tree height and volume of *Eucalyptus grandis* plantations using small-footprint, discrete return lidar data. **Progress in Physical Geography**, v. 34, n.4, pp. 515-540, 2010.

TILLEY, B.K. et al. Cost Considerations of Using Lidar for Timber Inventory. 2004. Available online: <http://sofew.cfr.msstate.edu/papers/0504tilley.pdf/> (accessed on 21 November 2018)

VALBUENA, R. et al. Most similar neighbor imputation of forest attributes using metrics derived from combined airborne LIDAR and multispectral sensors. **International Journal of Digital Earth**, p. 1–14, 2017.

WANG, Z. et al. Color and LIDAR data fusion: application to automatic forest boundary delineation in aerial images. **International Archives of Photogrammetry Remote Sensing and Spatial Information Science [CD]**, v. 36, n. 1, p. 51, 2007.

WHITE, J.C. et al. The utility of image-based point clouds for forest inventory: a comparison with airborne laser scanning. **Forests**, Basel, v. 4, n. 3, p. 518-536, 2013.

WERE, K. et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, 52, 394–403, 2015.

WOLTER, K.M. **Introduction to variance estimation**. Springer Verlag Wykoff, 2007.

XU, C. et al. Evaluation of modelling approaches in predicting forest volume and stand age for small-scale plantation forests in New Zealand with RapidEye and LiDAR. **International Journal of Applied Earth Observation and Geoinformation**, v. 73, p. 386-396, 2018.

ZHAO, K. et al. Lidar remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers. **Remote Sensing of Environment**, Amsterdam, v. 113, n. 1, p. 182-196, 2009.

ZHANG, Z. et al. Estimating Forest Structural Parameters Using Canopy Metrics Derived from Airborne LiDAR Data in Subtropical Forests. **Remote Sensing**, v. 9, p. 940, 2017.

ZHAO, K. et al. Characterizing forest canopy structure with lidar composite metrics and machine learning. **Remote Sensing of Environment**, Amsterdam, v. 115, n. 8, p. 1978-1996, 2011.

CHAPTER II

COMPARISON OF ALGORITHMS FOR INDIVIDUAL TREE DETECTION IN *Eucalyptus*
spp. PLANTATIONS FROM LiDAR DATA

SILVA, VANESSA SOUSA DA. Comparison of algorithms for individual tree detection in *Eucalyptus* spp. plantations from LiDAR data. 2019. Advisor: Emanuel Araújo Silva. Co-Advisors: Carlos Alberto Silva e Gabrielle Hambrech Loureiro.

ABSTRACT

Light Detection and Ranging (LiDAR) has emerged as a well-suited technology for accurate estimates of structural parameters both in natural and industrial plantation forest ecosystems. However, studies of LiDAR-derived images to retrieve forest attributes at individual tree level are still not as widely developed as plot or stand level approaches. A variety of approaches can be used to detect and delineate individual trees, but due to inadequate tree finding methods, significant omission and commission errors occur frequently in the segmentation results. Aiming errors reduction and accuracy refinement, this study evaluates a novel framework to automatically detect individual *Eucalyptus* trees. The study was conducted at three farms located in the state of São Paulo, in southeast Brazil. Data from field inventory and LiDAR of the year 2013 were used. For the analysis, a total of 15 circular plots of 400 m² each were used. In order to access the more accurate tree detection method the following algorithms were tested: Dalponte, Silva and Watershed. Results showed that Dalponte and Silva algorithms presented more accurate results with a total difference of 101 trees (15.19%) each from the reference field data. The performance of individual-tree detection was better using Silva and Dalponte algorithms with errors ranging from 2.22% to 26.19%. When evaluating the tree detection quality, it was observed that the use of Silva algorithm presented slightly better results, mainly due to the lower number of comission and omission errors, resulting in better F-scores in most of the sample plots.

Key-words — LiDAR, Forest Inventory, 3D tree segmentation, individual tree detection, Remote Sensing.

SILVA, VANESSA SOUSA DA. Comparação de algoritmos para detecção individual de árvores em plantios de *Eucalyptus* spp. a partir de dados LiDAR. 2019. Orientador: Emanuel Araújo Silva. Co-orientadores: Carlos Alberto Silva e Gabrielle Hambrecht Loureiro.

RESUMO

Light Detection and Ranging (LiDAR) surgiu como uma tecnologia adequada para estimativas precisas de parâmetros estruturais em ecossistemas florestais de plantações naturais e industriais. No entanto, estudos de imagens derivadas de LiDAR para obtenção de atributos florestais em nível de árvore individual ainda não são tão amplamente desenvolvidos quanto as abordagens em nível de parcela ou talhão. Uma variedade de abordagens pode ser usada para detectar e delinear árvores individuais, mas devido a métodos inadequados de identificação de árvores, erros significativos de omissão e comissão ocorrem com frequência nos resultados da segmentação. Com o objetivo de redução de erros e refinamento da precisão, este estudo avalia um novo framework para detecção automática de árvores de Eucalipto individuais. O estudo foi realizado em três fazendas localizadas no estado de São Paulo, no sudeste do Brasil. Dados do inventário de campo e de LiDAR do ano de 2013 foram utilizados. Para a análise, foram utilizadas 15 parcelas circulares de 400 m² cada. A fim de acessar o método mais preciso de detecção de árvores, foram testados os seguintes algoritmos: Dalponte, Silva e Watershed. Os resultados mostraram que os algoritmos Dalponte e Silva apresentaram resultados mais precisos com uma diferença total de 101 árvores (15.19%) cada um a partir dos dados de campo de referência. Ambos algoritmos apresentaram erros variando de 2,22% a 26,19%. Ao avaliar a qualidade da detecção arbórea, observou-se que o uso do algoritmo Silva apresentou resultados ligeiramente melhores, principalmente pelo menor número de erros de comissão e omissão, resultando em melhores F-scores na maioria das parcelas amostradas.

Palavras-chave — LiDAR, Inventário Florestal, Segmentação de árvores 3D, Detecção Individual de Árvores, Sensoriamento Remoto.

1. 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. Forest inventory in plantation of *Eucalyptus* hybrid clones is usually conducted annually to monitor growth, identify problematic conditions during initial growth stages, and determine optimal harvest time later in the growth cycle (IBÁ, 2018). The most usual method for estimating attributes such as tree density (tree/ha), and tree characteristics such as height (Ht), basal area (BA), and stem volume (V) is to physically sample them in the field. However, individual tree field measurements over large areas can become uneconomical, time and effort consuming, and hence are not ideal for studies dealing with periodic data collection (GARDNER et al., 2008; SILVA et al., 2016).

Individual tree information is important in many forestry-related activities, such as selective cuts, silviculture treatment, and tree growth modelling (LICHSTEIN et al., 2010). The density of trees in a given locality is also considered a relevant information due to the fact it is highly associated to biometric variables of a forest stand, such as basal area and volume. (SILVA et al., 2017) According to the aforementioned authors, the detection of individual trees automatically is a fundamental point in studies that aim to extract biometric information at the tree level. Individual tree detection (ITD) refers to partitioning the raw LiDAR data into objects representing individual trees based on the arrangement of the returns in space (BREIDENBACH et al., 2010). In this context, approaches for deriving more efficient, less expensive and time-consuming forest inventory information based on remotely sensed data have become of great utility and interest (GAMA et al., 2010).

Airborne Light Detection and Ranging (LiDAR), is now considered an important remote sensing technique for plot- and stand-level forest inventory, mainly because this technology can quickly provide highly accurate and spatially detailed information about forest attributes across entire forested landscapes (SILVA et al., 2014). Key LiDAR applications include high accuracy retrieval of tree density, stem volume, above ground carbon, leaf area index and basal area (ANDERSEN et al., 2005; ROBERTS et al., 2005; HUDAK et al., 2006; COOPS et al., 2007; SILVA et al., 2014).

The use of airborne LiDAR to retrieve forest attributes at the tree level is promising, however, not as widely studied as plot- or stand-level approaches (SILVA et al., 2016). A LiDAR-

derived Canopy Height Model (CHM) can be used for detecting individual trees, delineating tree crowns, and subsequently estimating biophysical attributes such as biomass and stem volume (POPESCU, 2007; FALKOWSKI et al., 2009; HU et al., 2014; DUNCANSON et al., 2015). Individual-tree attributes are predicted following the individual tree detection and metrics extraction, however, the accurate prediction of tree-level attributes is highly dependent on the methods used to detect and extract individual-tree and forest structure as well (KANKARE et al., 2015).

A variety of approaches can be used to detect and delineate individual trees from LiDAR-derived CHMs. These include identifying local maxima (POPESCU et al., 2003; WEINACKER et al., 2004; FALKOWSKI et al., 2008; FALKOWSKI et al., 2009) for tree detection, as well as region growth (HYYPPIA et al., 2008; SOLBERG et al., 2006; PANG et al., 2008), valley following (LECKIE et al., 2003), and watershed (CHEN et al., 2007; JING et al., 2012) for tree crown delineation. As LiDAR remote sensing techniques are undergoing rapid improvement along with the availability of high spatial resolution remotely sensed imagery there is potential for conducting and automating high accuracy forest inventory and analysis in a cost-effective manner. In this context this study aims to evaluate the ability of automated tree identification algorithms to accurately perform a LiDAR-based individual-tree detection in *Eucalyptus* spp. plantations in southeast Brazil.

2. MATERIALS AND METHODS

2.1. Study Area

The study area consisted of three farms located in the municipalities of Pilar do Sul and São Miguel Arcanjo, southeast region of the state of São Paulo, Brazil (Figure 1). According to the Köppen classification, the climate of the region is characterized as humid subtropical, with wet and hot summers and dry and cold winters. Mean annual precipitation is ~1700 mm; mean annual temperature is 18.8 °C (ALVARES et al., 2013). The topography in the selected plantations ranges from mildly to very hilly with an elevation ranging from 659 m to 1210 m. The soils of the region are predominantly red and yellow red latosol, all classified as clayey or very clayey.

The farms contained industrial eucalyptus plantations managed by Suzano S.A., a pulp and paper company located in São Paulo state, Brazil. The plantations were composed of hybrid clones of two *Eucalyptus* species, *Eucalyptus grandis* W. Hill ex Maid and *Eucalyptus urophylla*

S.T. Blake. All the trees were planted predominantly in a 3m x 2m grid configuration, resulting in an average density of 1,667 trees ha⁻¹. Stand age across the farms was variable and ranged from 2 to 6 years.

2.2 Field Data

The study was based on data collected in a set of temporary and permanent sample plots installed for the purpose of annual forest inventory by the company. A total of 15 circular plots of 400 m² each were randomly established across the three farms. In each sample plot, individual trees were measured for diameter at breast height (DBH; cm) at 1.30 m, and a random subsample (15%) of trees for tree heights (Ht; m). Measurements were carried out during the months of April to November of 2013. All the sample plots were georeferenced in the field using a geodetic GPS unit with differential correction capability (Trimble Pro-XR). The projected coordinate system used was UTM SIRGAS 2000, zone 23 S. All the field measurements were provided by the inventory team of Suzano S.A.

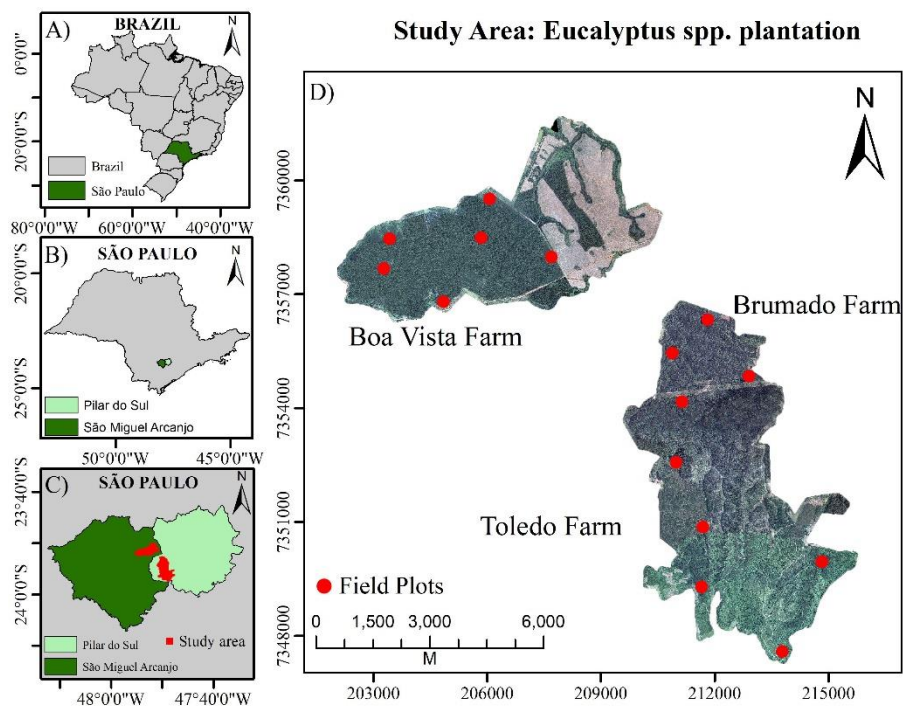


Figure 1. Location map of study area and plots. (A) Brazil and São Paulo State; (B) São Paulo State and the municipalities of Pilar do Sul and São Miguel Arcanjo; (C) Study area within the municipalities of Pilar do Sul and São Miguel Arcanjo; (D) Field plots in the study area.

2.2. LiDAR data collection specifications and processing

An airborne LiDAR survey was conducted in the study area on December 5th, 2013 using a Harrier 68i sensor (Trimble, Sunnyvale, CA, USA) mounted on a CESSNA 206 aircraft. The characteristics of the LiDAR data acquisition are listed in Table 1. LiDAR data processing steps were performed using FUSION/LDV 3.7 software (US Forest Service, Washington, DC, USA) (MCGAUGHEY, 2016) which provided three major outputs: the Digital Terrain Model (DTM), the Digital Surface Model (DSM), and the LiDAR-derived canopy structure metrics (CHM).

Table 1. Airborne LiDAR survey specifications.

Parameter	Value
Scan angle (°)	±45°
Footprint	0.33 m
Flying altitude	438 m
Swath width	363.11 m
Overlap	100% (50% side-lap)
Scan frequency	300 kHz
Average point density	10 pts.m ⁻²

In order to differentiate between ground and vegetation points, the original point cloud data were initially filtered using a classification algorithm available in the *groundfilter* function in FUSION/LDV. The *gridsurfacecreate* function was used to generate the 1-meter resolution Digital Terrain Models (DTMs), using the classified ground returns. The *canopymodel* tool was then used to interpolate vegetation points and to generate the Digital Surface Models (DSMs) and Canopy height models (CHM). The *clipdata* function was applied to obtain normalized heights by subtracting the DTMs elevations from each LiDAR return. Normalized point clouds were subset within the field sample plots of interest using the *polyclipdata* function. The *cloudmetrics* tool with a height and cover thresholds of 1.37 m were used to compute the canopy cover (COV,%), within sample plots. COV was calculated as the number of lidar first returns above 1.37 m, divided by the total number of first returns. LiDAR-derived CHM often contain height irregularities within individual-tree crowns (data pits) which reduce accuracy in tree detection and subsequent extraction of biophysical parameters (SHAMSODDINI et al., 2013). Therefore, the pit-free algorithm, developed by Khosravipour et al., (2014) was used to generate a pit-free CHM though a workflow implemented in LAsTools (ISENBURG, 2017).

2.4. Individual Tree Detection and HMAX Extraction

Individual tree segmentation was performed in R (R Development Core Team 2015) using the `lastrees` function from the `lidR` package (ROUSSEL et al., 2017). The `lastrees` function uses several possible algorithms to search for treetops in the CHM based on local maxima, matching lidar and field trees automatically. This method identifies the locations of maximum brightness intensity of the image in individual bands, using a mask (WULDER et al., 2000). In this case, the brightest areas refer to the highest areas of the canopy. A 5 x 5 moving window with a fixed tree-top window size (TWS) was used on a CHM smoothed by a mean smooth filter with fixed smoothing window size (SWS) of 3 x 3, which was chosen based on the best results obtained in the work of Silva et al. (2016).

In order to access the more accurate tree detection method 3 algorithms (Dalponte, Watershed, and Silva) were tested. The number of trees detected (NTD) per plot from LiDAR were manually compared with field-based data and an orthomosaic, and then evaluated in terms of true positive (TP, correct detection), false negative (FN, omission error) and false positive (FP, commission error). The accuracy of the detection was further evaluated for recall (1), precision (2) and F-score (3) according to Li et al. (2012), using the following equations (GOUTTE; GAUSSIER 2005; SOKOLOVA et al. 2006):

$$r = TP / (TP + FN) \quad (1)$$

$$p = TP / (TP + FP) \quad (2)$$

$$F = 2 * ((r * p) / (r + p)) \quad (3)$$

Recall is inversely related to omission error and represents the tree detection rate. Precision is inversely related to commission error and describes the rate of correct detections. F-score is used to represent the harmonic mean of recall and precision, which takes both commission and omission errors into consideration. Hence, a higher F-score indicates that both commission and omission errors are lower (Li et al., 2012). Recall, precision and F-score ranges from 0 to 1, and the F-score will become higher with higher p and r values.

The results of each method detection were also evaluated by the Chi-square test, considering that the expected frequency was the real value of trees in the plot (the visible trees

observed in the field) and the value observed was the total of trees detected TP and FP). Chi-square values calculated for each method was compared with the tabulated value of the test, considering the probability of 95%, applying the equation 4:

$$X_k^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (4)$$

were O_{ij} = observed number of trees; and E_{ij} = expected number of trees.

The total number of trees observed in each plot was also compared to the total number of trees detected in each method, in order to evaluate the overall quality of the detection of each method.

3. RESULTS

The total number of trees detected in the 15 Eucalyptus plots with the different methods is shown in Table 2. All methods tended to underestimate the results in most plots, with errors ranging from 2.22% to 56.86%. Dalponte and Silva algorithms presented more accurate results with a total difference of 101 trees (15.19%) each from the field data. The average number of trees detected by both algorithms did not differ significantly from the values observed in the field. Watershed was the method that underestimated the number of trees the most with a total difference of 213 trees (32.03%), presenting a significant difference by the chi-square test compared to the field data. The analysis of the tree detection quality in the plots is presented in Table 3.

It was observed that for most of the cases the best F-scores was obtained with the Silva algorithm. The overall F-score of the Silva algorithm was 0.87, followed by the detections with Dalponte algorithm (F-score 0.87) and Watershed algorithm (F-score 0.77). Both Silva and Dalponte algorithms stands out for the higher values of r and p, which reached maximum value (1) in 3 plots.

The results of the detection with the applied methods is illustrated in Figure 2 for plots 7 (F-score 0.76), and 12 (F-score 0.95), which presented respectively the lowest and the highest F-scores with the use of Silva and Dalponte algorithms. In plot 12 it is observed that the detection was the most accurate, while the other methods tended to not identify some trees. The highest FP numbers were observed in the Watershed followed by the Dalponte methods. In plot 7, it was

observed a high occurrence of FN, especially of trees with smaller crowns, some not detected in most of the methods. Errors of the FP type were observed mainly in larger crowns, by the identification of two points in the same crown. The use of Dalponte algorithm tended to present more FP type errors, detecting more than one point in some cups. FN errors were also more frequent using Dalponte algorithm, especially in smaller cups. Although Silva and Dalponte methods compared presented slightly different performances both methods were found to provide comparable results.

Table 2. Results of tree detection by the different algorithms.

Plots	Census	Watershed			Dalponte			Silva		
		N	Dif	%	N	Dif	%	N	Dif	%
1	47	40	7	14.89	51	4	8.51	51	4	8.51
2	33	27	6	18.18	36	3	9.09	36	3	9.09
3	44	37	7	15.91	50	6	13.64	50	6	13.64
4	51	32	19	37.25	44	7	13.73	44	7	13.73
5	51	22	29	56.86	39	12	23.53	39	12	23.53
6	47	36	11	23.40	43	4	8.51	43	4	8.51
7	37	28	9	24.32	31	6	16.22	31	6	16.22
8	41	37	4	9.76	46	5	12.20	46	5	12.20
9	38	19	19	50.00	30	8	21.05	30	8	21.05
10	46	27	19	41.30	36	10	21.74	36	10	21.74
11	42	23	19	45.24	31	11	26.19	31	11	26.19
12	39	30	9	23.08	35	4	10.26	35	4	10.26
13	47	24	23	48.94	37	10	21.28	37	10	21.28
14	45	29	16	35.56	44	1	2.22	44	1	2.22
15	57	41	16	28.07	47	10	17.54	47	10	17.54
ΣPlots	665	452	213*	32.03	600	101 ^{ns}	15.19	600	101 ^{ns}	15.19

N: number of trees detected by the different methods; Dif: difference between census and detection; %: percentage difference; * and ns: respectively, significant and non-significant by the chi-square test at 95% probability.

Table 3. Analysis of the tree detection quality by the different algorithms.

TP: true positive; FP: false positive; FN: false negative; r: recall; p: precision; F: F-score. The best F-scores for each plot are in bold.

Plots	COV (%)	Census	Watershed						Dalponte						Silva					
			TP	FP	FN	r	p	F	TP	FP	FN	r	p	F	TP	FP	FN	r	p	F
1	88.02	47	38	2	9	0.81	0.95	0.87	45	6	2	0.96	0.88	0.92	44	7	3	0.94	0.86	0.90
2	82.51	33	24	3	9	0.73	0.89	0.80	30	6	3	0.91	0.83	0.87	32	4	1	0.97	0.89	0.93
3	81.79	44	37	0	7	0.84	1.00	0.91	42	8	2	0.95	0.84	0.89	43	7	1	0.98	0.86	0.91
4	82.61	51	30	2	21	0.59	0.94	0.72	43	1	8	0.84	0.98	0.91	43	1	8	0.84	0.98	0.91
5	89.95	51	22	0	29	0.43	1.00	0.60	39	0	12	0.76	1.00	0.87	39	0	12	0.76	1.00	0.87
6	81.73	47	35	1	12	0.74	0.97	0.84	41	2	6	0.87	0.95	0.91	40	3	7	0.85	0.93	0.89
7	92.34	37	25	3	12	0.68	0.89	0.77	26	5	11	0.70	0.84	0.76	26	5	11	0.70	0.84	0.76
8	87.22	41	36	1	5	0.88	0.97	0.92	40	6	1	0.98	0.87	0.92	39	7	2	0.95	0.85	0.90
9	86.30	38	19	0	19	0.50	1.00	0.67	27	3	11	0.71	0.90	0.79	28	2	10	0.74	0.93	0.82
10	85.56	46	27	0	19	0.59	1.00	0.74	34	2	12	0.74	0.94	0.83	34	2	12	0.74	0.94	0.83
11	92.53	42	21	2	21	0.50	0.91	0.65	31	0	11	0.74	1.00	0.85	28	3	14	0.67	0.90	0.77
12	82.36	39	30	0	9	0.77	1.00	0.87	35	0	4	0.90	1.00	0.95	35	0	4	0.90	1.00	0.95
13	89.72	47	22	2	25	0.47	0.92	0.62	35	2	12	0.74	0.95	0.83	37	0	10	0.79	1.00	0.88
14	83.32	45	27	2	18	0.60	0.93	0.73	40	4	5	0.89	0.91	0.90	40	4	5	0.89	0.91	0.90
15	84.55	57	38	3	19	0.67	0.93	0.78	44	3	13	0.77	0.94	0.85	45	2	12	0.79	0.96	0.87
Total		665	431	21	234	0.65	0.95	0.77	552	48	113	0.83	0.92	0.87	553	47	112	0.83	0.92	0.87

Figure 2. Detection of trees on plots by the algorithms: (A) Watershed (B) Dalponte (C) Silva.

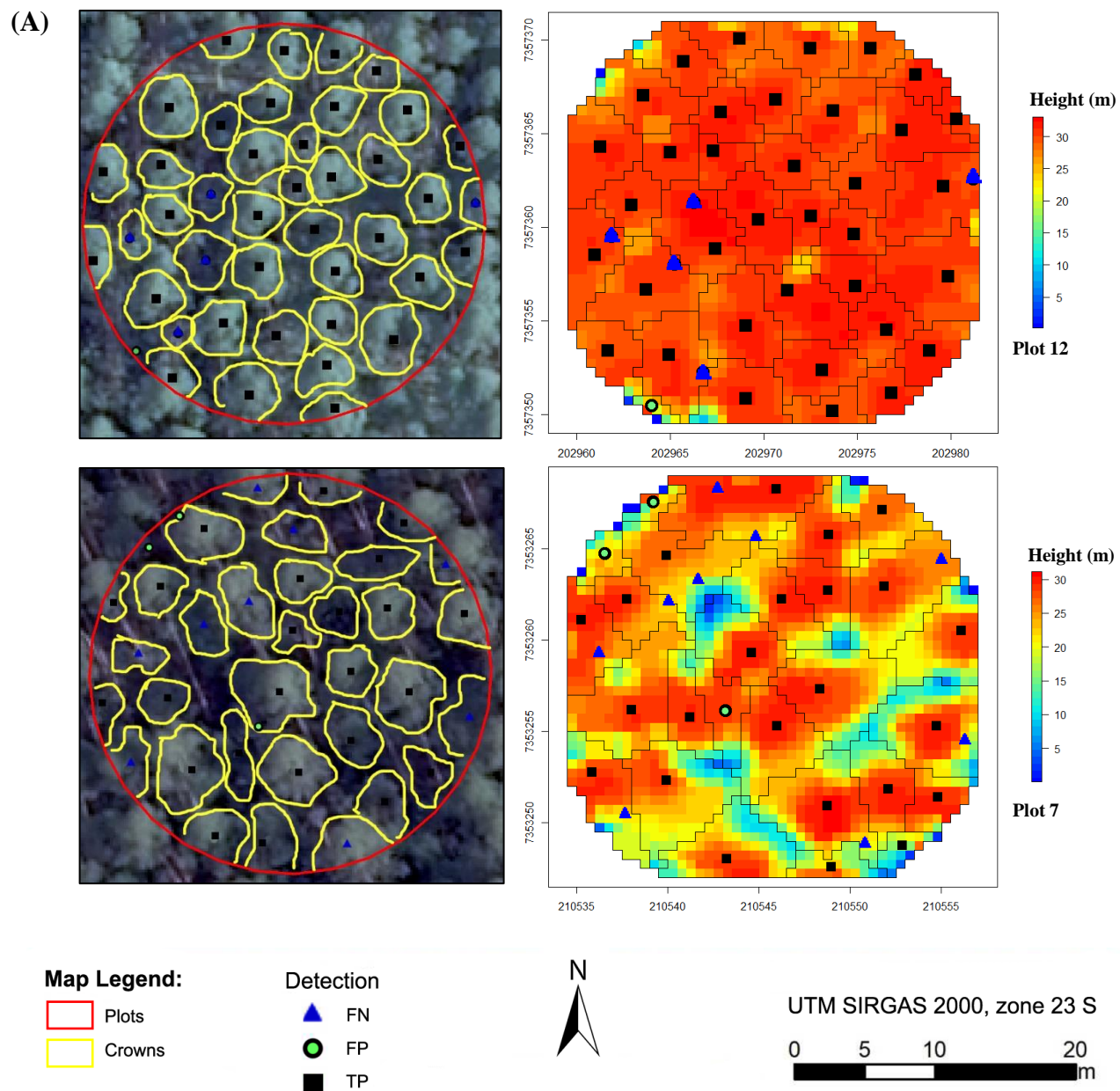


Figure 2. Detection of trees on plots by the algorithms: (A) Watershed (B) Dalponte (C) Silva.

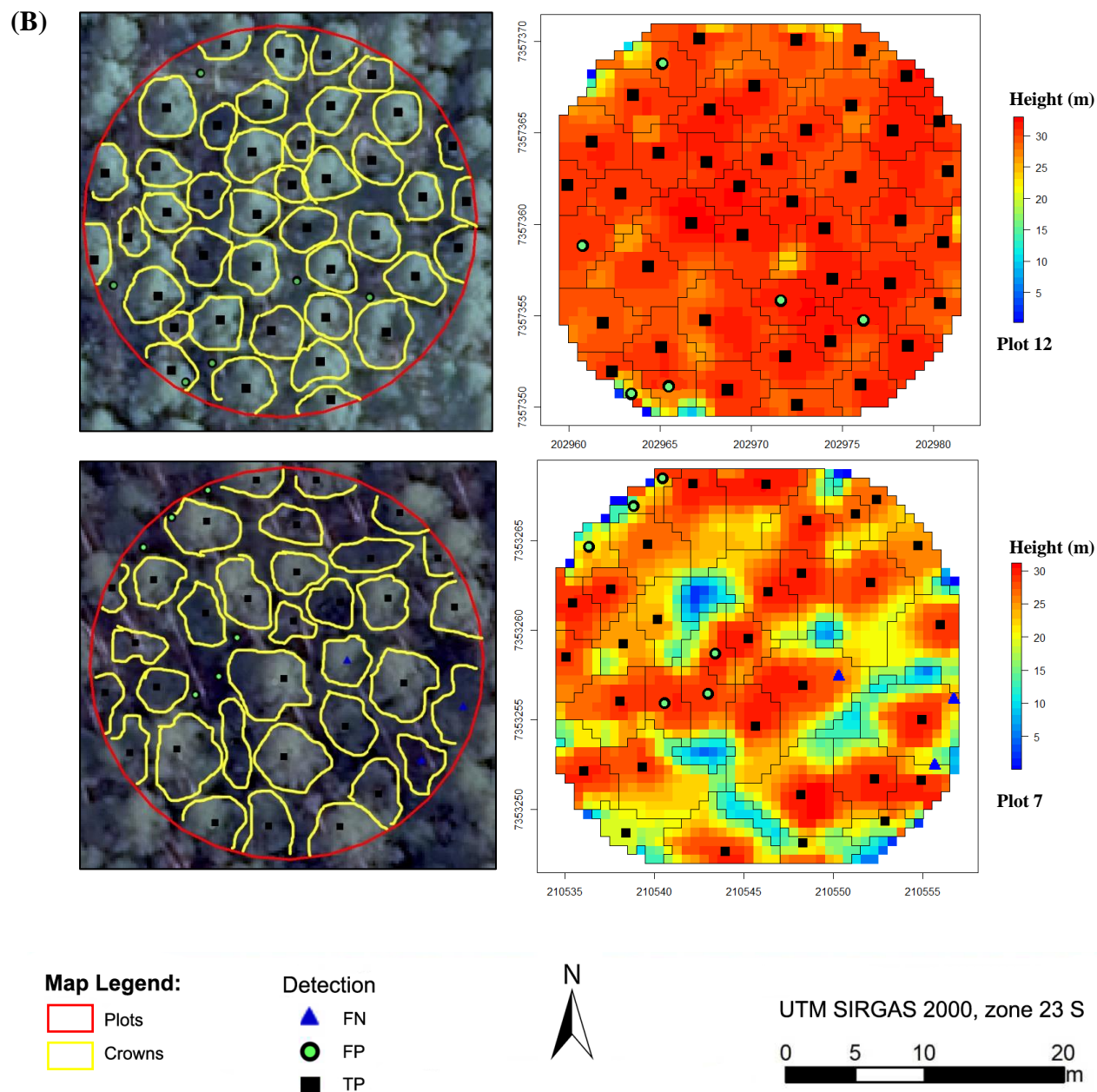
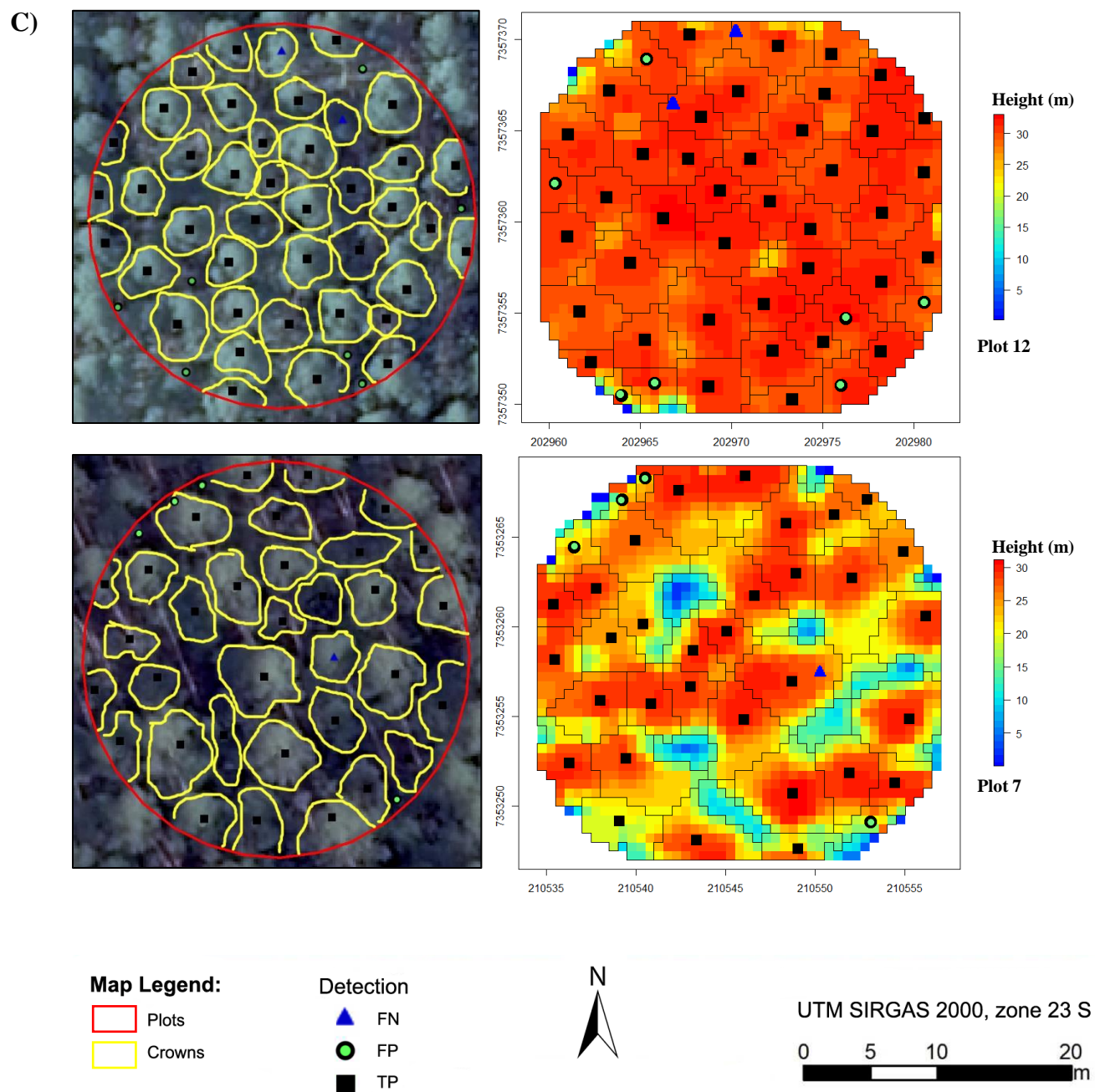


Figure 2. Detection of trees on plots by the algorithms: (A) Watershed (B) Dalponte (C) Silva.



The summary of the detection analysis related to the number of trees detected and the quality of the detection is presented in Table 4. It was observed that both methods (Silva and Dalponte) using the *lastrees* function from the *lidR* package did not present significant difference when comparing the experimental results with the real number of trees in all the plots. In the general ranking, was observed that the Silva method presented a slightly better result for the quality detection, being the first placed. Subsequently, Dalponte method come in second, followed by the Watershed method in third.

Table 4. Summary results of the tree detection with all methods tested in the 15 plots.

Method	TP	FP	FN	F-score	Rank
Silva	553	47	112	0.87	1
Dalponte	552	48	113	0.87	2
Watershed	431	21	234	0.77	3

TP: true positive; FP: false positive; FN: false negative; F-score.

4. DISCUSSION

Accurate information on forest attributes at individual-tree level has a decisive impact on decision-making processes in forest and timber management. The most accurate method of estimating these attributes is to physically sample them in the field. However, individual tree field measurements are limited by budgets and time, making them impractical over large areas. As a result of the need for individual tree-based attributes and finer scale descriptions of stands, airborne LiDAR technology has become the main remote sensing technique for individual tree detection (ITD). Since an accurate individual tree detection is highly dependent on the applied methods, this study presents an investigation of simplified automated frameworks for lidar-based individual-tree detection.

We found that the successful tree detection using the CHM based on local maxima technique was greatly affected by the different methods. The performance of individual-tree detection was better using Silva and Dalponte algorithms with errors ranging from 2.22% to 26.19%. When evaluating the tree detection quality, it was observed that the use of Silva algorithm presented slightly better results, mainly due to the lower number of false positives. The majority of the commission errors (FP) occurred in large, especially irregular crowns. False negatives

(omission errors) were also observed, generally related to smaller or forked trees (with two visible crowns, but only one detected). It was also observed that all the methods presented higher number of false negatives than false positives, confirming that most methods underestimated the real number of trees.

It was noted that even though Dalponte and Silva methods did not differ significantly between each other in total number of trees detected, they did not necessarily present the same results when the quality of detection was analyzed in the sample plots. The Silva algorithm attained better F-scores in most of the sample plots. According to Tanhuanpää et al. (2016), this is because the occurrence of both omission and commission errors may end up masking errors in the total tree count, which reinforces the importance of analyzing factors such as F-scores in order to evaluate the quality of the methods.

The tree-detection results from this study are comparable to the results obtained in other studies using local maxima approaches. Li et al. (2012), using an individual tree segmentation method from LiDAR data in a mixed conifer forest in the US, showed that the algorithm detected 86% of the trees (“recall”), and 94% of the trees were segmented correctly (“precision”), with an overall F-score of 0.90. Vega et al. (2014), when segmenting individual trees in a conifer plantation in France, reported overall recall, precision, and F-score of 0.93%, 0.98%, and 0.95, respectively. Khosravipour et al. (2014), in a mixed forest in France, achieved an overall accuracy of 70.6%. Mohan et al. (2017) obtained a F-score ranging between 0.73 and 0.95 in an open canopy forest area in the US, while Huang et al. (2018) found F-scores ranging from 0.74 to 0.90 to a similar typology with different densities. In Eucalyptus plantations using local maxima methods, Shinzato et al. (2017) obtained a detection success of only 58% of the trees. Guerra-Hernández et al. (2018), also in Eucalyptus plantations presented a detection success of about 79.6% with omission and commission errors of 20.8% and 6.5%.

Previous researches conducted by Falkowski et al. (2008) and Silva et al. (2016) has shown that tree-detection accuracy tends to decrease with increasing canopy cover. In this study was found the same trend, where the accuracy of individual-tree detection measured by the F-score was inversely proportional to forest COV. Overall, commission errors were more prevalent in less dense test plots, and omission errors were more common where crowns overlapped. The influence of the tree’s geolocation absence can also be considered an unquantifiable source of uncertainty in the current study. However, Popescu (2007) reported that treetop positions might be determined

with higher accuracy using a CHM image detected from lidar rather than error-prone measurements derived from differential GPS in the field, especially in high-canopy-cover conditions that can degrade field GPS accuracy (WING et al., 2008).

It is also important to highlight that for the methods based on local maxima algorithms, the results are highly influenced by the smoothing parameters of the CHM and the tree-top window size (MOHAN et al., 2017; PANAGIOTIDIS et al., 2016). In general, a CHM smoothing is recommended, but in some tests, it was observed that the original CHM was more effective in some of the plots (MOHAN et al., 2017). Despite this, a smoothing window was chosen in this study due to some canopy irregularity (i.e. broken tree tops, forking trees and others), when the original CHM was selected, the number of false positives was larger. Guerra-Hernández et al. (2018) also observed that the use of a LiDAR derived CHM requires the application of a smoothing window for the correct detection of Eucalyptus trees using a local maxima filter.

5. CONCLUSION

Predictions of plot-level tree attributes averages often do not provide a sufficient description of the stand for decision-making regarding management of forest resources. The capacity to make accurate predictions of not only the total stand parameters, but also of the frequency distribution of individual tree (tree lists), provides valuable information that can be used in forest inventories, especially in an operational context.

In this study, we investigated the ability of automated tree identification algorithms to accurately perform a LiDAR-based individual-tree detection in *Eucalyptus* spp. plantations. Two (Dalponte and Silva) of the three tested methods were able to detect individual trees with high accuracy in areas with < 70% COV. The precision and accuracy of LiDAR in detecting individual trees using the framework presented was demonstrated through detection quality parameters. Both Silva and Dalponte algorithms proved to be promising achieving comparable results, but the quality analysis showed a slightly superiority of the Silva tool, since it presented a non-significant difference between census and detection with lower commission and omission errors.

Future directions for this research include the test of additional features (i.e. GPS location errors, comparing other algorithms with different TWS and SWS); integration of a multi-sensor data fusion approaches (i.e. UAV LiDAR, spectral data, aerial photographs); and the estimative of dendrometric variables such as DAP, tree height, crown size and diameter, and thereby develop

predictive models for estimating aboveground biomass and stem volume at individual tree level. We hope that the promising results for individual-tree-level detection in this study will support and stimulate further research and applications not just in *Eucalyptus* spp. plantations management but other forest types for predicting a wide range of forest attributes.

6. REFERENCES

- ANDERSEN, H. et al. Estimating forest canopy fuel parameters using LIDAR data. **Remote Sensing Environment**, v. 94, p. 441–449, 2005.
- BREIDENBACH, J. et al. Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. **Remote Sensing of Environment**, v. 114, p. 911–924, 2010.
- BREIMAN, L., Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.
- CHEN, Q. et al. Estimating basal area and stem volume for individual trees from LiDAR data. **Photogrammetric Engineering & Remote Sensing**, v. 73, n. 12, p. 1355–1365, 2007.
- COOPS, N. C. et al. Estimating canopy structure of Douglas-Fir forest stands from discrete-return LiDAR, **Trees - Structure and Function**, v. 21, p. 295-310, 2007.
- CROOKSTON N.L.; FINLEY A.O. yaImpute: An R Package for kNN Imputation. **J Stat Soft**, v. 23, n. 10, p. 1-16, 2008.
- DUNCANSON, L.I. et al. The importance of spatial detail: Assessing the utility of individual crown information and scaling approaches for Lidar- based biomass density estimation. **Remote Sensing of Environment**, v. 168, p. 102–112, 2015.
- FALKOWSKI, M.J., et al.Characterizing forest succession with LiDAR data: An evaluation for the Inland Northwest, USA. **Remote Sensing of Environment**, v.113, n. 5, p. 946–956, 2009.
- FALKOWSKI, M.J. et al.The influence of conifer forest canopy cover on the accuracy of two individual tree measurement algorithms using LiDAR data. **Canadian Journal of Remote Sensing**, v. 34, n. 2, p. S1–S13, 2008.
- GAMA, F.F. et al. Eucalyptus biomass and volume estimation using interferometric and polarimetric SAR data. **Journal of Applied Remote Sensing** v. 2, p. 939–956, 2010.
- GARDNER, T.A. et al. The cost-effectiveness of biodiversity surveys in tropical forests. **Ecology Letters**, v. 11, p. 139–150, 2008.
- GOETZ, S. et al. Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an Eastern Temperate forest, USA. **Remote Sensing of Environment**, v. 108, n. 3, p. 254–263, 2007.
- GOUTTE, C.; GAUSSIÉ, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. **Advances in Information Retrieval**, v. 3408, p. 345– 359, 2005.
- GUERRA-HERNÁNDEZ, J. et al. Comparison of ALS- and UAV(SfM)-derived high density point clouds for individual tree detection in *Eucalyptus* plantations. **International Journal of Remote Sensing**, v. 39, n. 15-16, p. 5211-5235, 2018.
- HYYPPÄ, J. et al. Review of methods of small - footprint airborne laser scanning for extracting forest inventory data in boreal forests. **International Journal of Remote Sensing**, v. 29, n. 5, p. 37–41, 2008.

HU, B. et al. Improving the efficiency and accuracy of individual tree crown delineation from high-density LiDAR data. **International Journal of Applied Earth Observation and Geoinformation**, v. 26, p. 145–155, 2014.

HUANG, H et al. Individual tree crown detection and delineation from very-high-resolution UAV images based on bias field and marker-controlled watershed segmentation algorithms. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v.17, p. 1–10, 2018.

HUDAK, A. T. et al. Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. **Canadian Journal of Remote Sensing**, v. 32, n. 2, p. 126–138, 2006.

KÖPPEN, W.; GEIGER, R. *Klimate der Erde*. Gotha: VerlagJustus Perthes. Wall-map 150cmx200cm.1928

INDÚSTRIA BRASILEIRA DE ÁRVORES. **Relatório Iba 2018**, Brasília, p. 80, 2018. em: Available: https://iba.org/images/shared/Biblioteca/IBA_RelatorioAnual2017.pdf. Access in: June 2018.

ISENBURG, M. **LAStools—Efficient Tools for LiDAR Processing**, 2017. Version 1.0-1.3, Available: <http://rapidlasso.com/lastools> Access in: December 2017.

JING, L., et al. Automated delineation of individual tree crowns from LiDAR data by multi-scale analysis and segmentation. **Photogrammetric Engineering & Remote Sensing**, v. 78, n. 12, p. 1275–1284, 2012.

JENSEN, J. R. **Sensoriamento Remoto do Ambiente: Uma perspectiva em recursos terrestres**. Tradução de José Carlos N. Epiphanyo [et al.]. 1 Ed. São José dos Campos: Parêntese, 2011. 672 p.

KANKARE, V. et al. Diameter distribution estimation with laser scanning based multisource single tree inventory.” **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 108, p. 161–171, 2015.

KHOSRAVIPOUR, A., et al. Generating pit-free canopy height models from airborne LiDAR. **Photogrammetric Engineering & Remote Sensing**, v. 80, n. 9, p. 863–872, 2014.

LI, W., et al. A new method for segmenting individual trees from the LiDAR point cloud. **Photogrammetric Engineering & Remote Sensing**, v. 78, n. 1, p. 75–84, 2012.

LICHSTEIN, J.W. et al. Unlocking the forest inventory data: Relating individual tree performance to unmeasured environmental factors. **Ecology Applied**, v. 20, p. 684–699, 2010.

LECKIE, D. et al. Combined high-density LiDAR and multispectral imagery for individual tree crown analysis. **Canadian Journal of Remote Sensing**, v. 29, n. 5, p. 633–649, 2003.

MCGAUGHEY R. J. M. **FUSION/LDV: Software for LIDAR data analysis and visualization**, Version 3.6. USDA Forest Service, Pacific Northwest Research Station (Seattle, WA), 2016.

- MCROBERTS, R.E. et al. Optimizing the k-nearest neighbor technique for estimating forest aboveground biomass using airborne laser scanning data. **Remote Sensing of Environment**, v. 163, p. 13–22, 2015.
- MOHAN, M. et al. Individual tree detection from unmanned aerial vehicle (UAV) derived canopy height model in an open canopy mixed conifer forest. **Forests**, v. 8, n. 9, p. 340, 2017.
- PANAGIOTIDIS, D. et al. Determining tree height and crown diameter from high resolution UAV imagery. **International Journal of Remote Sensing**, v. 38, n. 8–10, p. 2392–2410, 2016.
- PANG, Y. et al. Validation of the ICESat vegetation product using crown-area- weighted mean height derived using crown delineation with discrete return LiDAR data. **Canadian Journal of Remote Sensing**, v. 34, n. 2, p. 471–484, 2008.
- POPESCU S.C., et al. Measuring individual tree crown diameter with LiDAR and assessing its influence on estimating forest volume and biomass. **Canadian Journal of Remote Sensing**, v. 29, p. 564-577, 2003.
- POPESCU, S.C. et al. Estimating biomass of individual pine trees using airborne LiDAR. **Biomass and Bioenergy**, v. 31, n. 9, p. 646–655, 2007.
- R Core Team. R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**, 2015. Available online: <https://www.r-project.org/> (accessed on 15 Feb 2018).
- RACINE, E.B. et al. Estimating forest stand age from LiDAR-derived predictors and nearest neighbor imputation. **Forest Science**, v. 60, n. 1, p. 128–136, 2014.
- ROBERTS, S.D. et al. Estimation individual tree leaf area in loblolly pine plantations using LiDAR derived measurements of height and crown dimensions. **Forest Ecology Management**, v.213, p. 54–70, 2005.
- ROBINSON, A.P. et al. A regression-based equivalence test for model validation: shifting the burden of proof. **Tree Physiology**, v. 25, p. 903–913, 2005.
- ROUSSEL, J.R, et al. **lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications**, version 1.4.1, 2018.
- SHAMSODDINI, A., et al. Improving LiDAR-based forest structure mapping with crown-level pit removal. **Journal of Spatial Science**, v. 58, p. 29–51, 2013.
- SHINZATO, E. T. et al. Integrating area-based and individual tree detection approaches for estimating tree volume in plantation inventory using aerial image and airborne laser scanning data. **iForest**, v. 10, n. 1, p. 296–302, 2017.
- SILVA, C. A. et al. Mapping aboveground carbon stocks using LiDAR data in *Eucalyptus* spp. plantations in the state of São Paulo. **Scientia Forestalis**, v. 42, p. 591–604, 2014.
- SILVA, J.A.A. da. Conceitos e princípios básicos de modelagem matemática em ciências florestais. **Anais da Academia Pernambucana de Ciência Agrônômica**, v. 11/12, p.195-215, 2015.

SILVA, C. A. et al. Predicting stem total and assortment volumes in an industrial *Pinus taeda* L. Forest plantation using airborne laser scanning data and random forest. **Forests**, v. 8, n. 7, p. 254-267, 2017

SILVA, C. A. et al. Imputation of individual Longleaf Pine (*Pinus palustris* Mill.) tree attributes from field and LiDAR data. **Canadian Journal of Remote Sensing**, v. 42, n. 5, p. 554-573, 2016.

SOKOLOVA, M. et al. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. **Advances in Artificial Intelligence**, p. 1015–102, 2006.

SOLBERG, S. et al. Single tree segmentation using airborne laser scanner data in a structurally heterogeneous spruce forest. **Photogrammetry Engineering and Remote Sensing**, v. 72, n. 12, p.1369–1378, 2006.

TANHUANPÄÄ, T., N. et al. Evaluating the Performance of High-Altitude Aerial Image-Based Digital Surface Models in Detecting Individual Tree Crowns in Mature Boreal Forests. **Forests** v. 7, n. 7, p. 143, 2016.

VAUHKONEN, J. et al. Comparative testing of single-tree detection algorithms under different types of forest. **Forestry**, v. 85, n. 1, p. 27–40, 2012.

VEGA, et al. PTrees: A point-based approach to forest tree extraction from lidar data. **International Journal of Applied Earth Observation and Geoinformation**, v. 33, p. 98–108, 2014.

WEINACKER, H. et al. Development of filtering, segmentation and modelling modules for LiDAR and multispectral data as a fundamental of an automatic forest inventory system. **Photogrammetry, Remote Sensing and Spatial Information Sciences**. v. 36, n. 8, p. 50–55, 2004.

WING, M.G. et al. Horizontal measurement performance of five mapping-grade GPS receiver configurations in several forested settings. **Western Journal of Applied Forestry**, v. 23, n. 3, p. 166–171, 2008.

WULDER, M. et al. Local maximum filtering for the extraction of tree locations and basal area from high spatial resolution imagery. **Remote Sensing of the Environment**, Saint Paul, v. 73, p. 103- 114, 2000.

GENERAL CONCLUSION AND RECOMMENDATIONS

The research presented in this dissertation aims to contribute to the understanding of how LiDAR remote sensing can be efficiently applied for predicting and mapping critical forest structural attributes, such as volume and individual-tree-level detection, at industrial forest plantations. Major findings, contributions of this dissertation and future research directions are summarized for each chapter and presented as follows:

Chapter 1 presented a framework to predict and map stem total volumes in industrial *Eucalyptus* spp. plantations from LiDAR through the comparison of ten parametric and nonparametric modeling methods combined with varying the reference data (sample) size. The results of this chapter demonstrated that LiDAR data combined with OLS method, using only 40% of the total field plots could provide reliable estimates of total volumes. When LiDAR-derived estimates of stem volume were compared to reference forest inventory data, the accuracy of plot-level total volumes were high, presenting an average of relative root mean square error (RMSEr) of only 12.95%. Accurate estimates of crown attributes at the highest attainable spatial resolution is desired to increase the efficiency of monitoring and managing *Eucalyptus* spp. plantations. Future research should focus on estimating other forest attributes and at the tree level as well. Crown estimates would be highly desired information to assist in common forestry tasks, such as in thinning operations. Also, crown attributes could be used in combination with field data to fit taper models and improve the accuracy of volume estimates.

In Chapter 2 a framework to automatically detect individual trees and evaluate the detection efficacy was developed. Individual tree locations were estimated with high accuracy (90.22%), especially in low-canopy-cover conditions. While the methodology developed here shows promising results, further work is necessary in order to refine aspects of the approach to increase accuracy when estimating the total tree count. Future directions for this research include the combined use of spatial data, airborne and terrestrial lidar to better describe the structure of individual trees. Besides crown height and crown projected area, additional crown metrics, such as crown volume and surface area, should be computed and tested as new predictors for estimating not only the number of trees but other important forest attributes at tree-level as well, such as basal area and volume. As Unmanned Aerial Vehicle (UAV) remote sensing technologies and methods improve, there is potential for combining airborne lidar-derived DTM from a previous acquisition with UAV photogrammetry and Structure from Motion (SfM) algorithms for effectively

monitoring and mapping forest attributes at the individual tree level in *Eucalyptus* spp. plantations in a cost-effective manner.

There are many challenges yet to be faced in order to use the full potential of information provided by LiDAR. Other studies should be conducted in order to evaluate this methodology for other plantations and forested environments, more tree spacing topographic conditions, other algorithms and parameters, and test the accuracy of estimating other characteristics such as DBH and crown area at landscape, plot and tree-level, which are important factors required for estimating biomass and stem volume. Formulating methods to increase stem volume estimates efficiency and developing strategies that optimize tree detection algorithms based on the characteristics of the point cloud can surely open new windows in LiDAR data analytics. We hope that the results presented and discussed here will stimulate further research and applications of LiDAR remote sensing not just in experimental scenarios but in operational modes as well.