

RORAI PEREIRA MARTINS NETO

**EXTRACTION OF STRUCTURAL VARIABLES USING LIDAR
DATA COMBINED WITH HYPERSPECTRAL IMAGES FOR
CLASSIFICATION OF UPPER CANOPY TREE SPECIES IN
BRAZILIAN ATLANTIC FOREST**



Presidente Prudente
2021



UNIVERSIDADE ESTADUAL PAULISTA
CAMPUS DE PRESIDENTE PRUDENTE
FACULDADE DE CIÊNCIAS E TECNOLOGIA
Programa de Pós-Graduação em Ciências Cartográficas

RORAI PEREIRA MARTINS NETO

**EXTRACTION OF STRUCTURAL VARIABLES USING LIDAR DATA
COMBINED WITH HYPERSPECTRAL IMAGES FOR
CLASSIFICATION OF UPPER CANOPY TREE SPECIES IN
BRAZILIAN ATLANTIC FOREST**

Doctoral thesis presented to the Graduate Program of Cartographic Sciences (PPGCC) at São Paulo State University (UNESP), School of Technology and Sciences, campus Presidente Prudente, SP, Brazil, for the partial fulfilment of the requirements for the grade of Doctor in Cartographic Sciences.

Supervisor: Dr. Antonio M. G. Tommaselli

Co-Supervisors: Dr. Nilton N. Imai

Dr. Hassan C. David

Dr. Milto Miltiadou

Presidente Prudente
2021

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação - Diretoria Técnica de Biblioteca e Documentação - UNESP, Campus de Presidente Prudente

M345e Martins Neto, Rorai Pereira.
Extraction of structural variables using LiDAR data combined with
hyperspectral images for classification of upper canopy tree species in Brazilian
Atlantic Forest / Rorai Pereira Martins Neto. - 2021
167 f. : il.

Orientador: Antonio Maria Garcia Tommaselli
Coorientadores: Nilton Nobuhiro Imai, Hassan Camil David, Milto Miltiadou
Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de
Ciências e Tecnologia, Presidente Prudente, 2021
Inclui bibliografia

1. Remote sensing. 2. Tropical forest. 3. Forest modelling. 4. Machine
learning. 5. Tree species identification. I. Tommaselli, Antonio Maria Garcia. II.
Imai, Nilton Nobuhiro. III. David, Hassan Camil. IV. Miltiadou, Milto. V.
Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia,
Presidente Prudente. VI. Título.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: Extraction of structural variables using LiDAR data combined with hyperspectral images for classification of upper canopy tree species in Brazilian Atlantic Forest

AUTOR: RORAI PEREIRA MARTINS NETO

ORIENTADOR: ANTONIO MARIA GARCIA TOMMASELLI

COORIENTADOR: HASSAN CAMIL DAVID

COORIENTADOR: NILTON NOBUHIRO IMAI

COORIENTADORA: MILTO MILTIADOU

Aprovado como parte das exigências para obtenção do Título de Doutor em CIÊNCIAS CARTOGRÁFICAS, área: Aquisição, Análise e Representação de Informações Espaciais pela Comissão Examinadora:

Prof. Dr. ANTONIO MARIA GARCIA TOMMASELLI (Participação Virtual)
Departamento de Cartografia / Faculdade de Ciências e Tecnologia de Presidente Prudente

Prof. Dr. EDSON APARECIDO MITISHITA (Participação Virtual)
Departamento de Geomática / Universidade Federal do Paraná

Prof. Dr. VERALDO LIESENBERG (Participação Virtual)
Centro de Ciências Agroveterinárias / Universidade do Estado de Santa Catarina

Dra. EIJA HONKAVAARA (Participação Virtual)
Finnish Geospatial Research Institute / Finnish Geospatial Research Institute

Prof. Dr. ALUIR PORFIRIO DAL POZ (Participação Virtual)
Departamento de Cartografia / Faculdade de Ciências e Tecnologia de Presidente Prudente

Presidente Prudente, 08 de novembro de 2021



*To my beloved ones
who always believed in me*

CURRICULUM VITAE

Rorai Pereira Martins Neto

Birth	18/01/1991 – Curitiba – Paraná – Brazil
1997 – 2004	Elementary School Colégio Padre João Bagozzi
2005 – 2007	High School Curso e Colégio Acesso
2009 – 2013	Bachelor in Forestry Universidade Federal do Paraná – UFPR
2014 – 2015	Master in Forestry Universidade do Estado de Santa Catarina – CAV/UDESC Dissertation title: “ <i>Extração de variáveis dendrométricas em árvores de Pinus taeda L. a partir de dados TLS e ALS</i> ”
2015 – 2016	Substitute Professor Universidade Federal de Santa Maria – UFSM/FW Substitute Professor in Geosciences and Statistics area for undergraduate courses of Forestry and Agronomy
2017 – 2021	PhD in Cartographic Sciences Universidade Estadual Paulista – FCT/UNESP

ACKNOWLEDGMENTS

I would like to sincerely thank the many people who made this work possible.

First, I would like to thank my supervisor, mentor, and role model, Prof Dr. Antonio Tommaselli, who helped me complete this work through his dedication, confidence, experience, exchange of knowledge, help, and friendship. It was a great professional and personal privilege to be your student.

Thank you to my co-supervisors Prof Dr Nilton Imai, Dr. Hassan David, and Dr. Milto Miltiadou. Prof Dr Nilton Imai, for all your constructive discussions, exchange of ideas, and new insights that have enriched my work. Dr. Hassan David from the Brazilian Forest Service, for his longstanding friendship, supervision, advice, and help with forest modelling and statistics. And, to Dr. Milto Miltiadou from the Cyprus University of Technology for all her assistance in understanding and processing full-waveform data in DASOS software, her feedback in writing the thesis papers, and her help in generating ideas for future directions of this work.

Dr. Eija Honkavaara from Finnish Geodetic Institute for her expertise in LiDAR and hyperspectral data and for always helping me to improve submitted papers.

To the company ENGEMAP for using their LiDAR aerial survey in the study area, and to Thiago Tiedtke dos Reis, Survey Engineer, for his willingness to solve any doubts regarding the data. To Baltazar and Valtinho for helping with field surveys and with tree species recognition.

I am grateful to the Graduate Program in Cartographic Sciences professors at PPGCC for their dedication to passing on their knowledge during my PhD, and to the many university employees for doing their utmost to keep the working environment possible despite COVID19 and for their assistance in resolving bureaucratic issues.

My personal thanks, to my family for their unconditional support in my achievements. Dealing with distance and absence was never easy, but we always knew it was for something bigger. Thank you for always believing in my potential and in my dreams.

Many thanks to the great friends that life has given me: André, Cadu, David, Emenize, Gabriella Aragão, Kessye, Luana, Luani, Luís Paulo, Magda Zanon, Nathália, Regina, Thiago Meyer, and Vinícius, who understood my absence from their lives but always received me with open arms during my sporadic appearances. Special thanks to my friends Tiago de Conto and Tainá Mamede for their help with programming in R and forest ecology respectively and for

all the summers we spent together. To Eric Glunt for always being willing to review the English to improve the manuscript. And to the group "Quarentenados" for the companionship, friendship, and exchange of experiences during the pandemic.

To Prof Dr Christel Lingnau from Federal University of Paraná, who, since my first degree, has always believed in me and in my capacity to go further, encouraging me to pursue both a master's and a doctorate degree. Part of who I am today I owe to your teachings and guidance. And to Prof. Veraldo Liesenberg from Santa Catarina State University, for his willingness to always answer my questions when they arose.

Last, but not least, I thank my colleagues, interns, and friends at the PPGCC, especially Marcela Machado, Erika Saito, Guilherme Nascimento, Lucas Dias, Lara Lopes, Mariana Batista, Tayná, Samara, Fabrício Prol, Daniel Arana, Fernando Baixo, Guilherme Santos and Luís Macanoni for their assistance in the scope of cartographic sciences, reception, advice, good times, conversations, and happy hours. In addition, I thank my friends from Presidente Prudente: Marcella Rocha and Letícia, and all those who indirectly was part of my trajectory during my doctorate.

FINANCIAL ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This study was also supported by: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP—Grant: 2013/50426-4) and Academy of Finland (Grant: 273806), which jointly funded the international project Unmanned Airborne Vehicle - based 4D Remote Sensing for Mapping Rain Forest Biodiversity and Its Change in Brazil (UAV_4D_Bio).



All photos were taken by the author. They are from different species of Lapacho (ipês) in the city of Presidente Prudente - SP, Brazil.

“Nature has fixed no limits on our hopes”

Björk

ABSTRACT

The main goal of this thesis was to explore LiDAR data to estimate stand and diversity variables in a remnant of Seasonal Semideciduous Forest, also known as inland Atlantic Forest, one of the most degraded phytophysiognomies of the Atlantic Forest. In addition, the fusion of LiDAR data with hyperspectral images for classification of upper canopy species was also approached. LiDAR data were delivered as peak returns (PR) and full-waveform (FWF) and at two flight heights (900 m and 2000 m), and the hyperspectral images were collected by a light-weight camera. Despite interventions over the years, the study area, the *Ponte Branca* forest remnant, has a good state of conservation, with the presence of pioneer species to climax. PR LiDAR metrics were extracted from the 900 m height flight. Variable selection methods and machine learning techniques were tested to estimate stand and diversity variables. The new variables generated by transforming the metrics with Principal Component Analysis (PCA) and the technique of artificial neural networks (ANN) were the combination that best estimated the variables. The comparison of the attributes extracted by a voxel-based approach, available in the DASOS software, was performed for PR and FWF data. The results with data from two flight heights in the estimation of forest variables were also compared. FWF data showed less biased estimates than PR data. Eight species present in the upper canopy were classified using isolated and combined data: PR, FWF LiDAR, and hyperspectral images. The LiDAR PR metrics transformed by PCAs combined with vegetation indices obtained by hyperspectral images provided the best accuracy in the classification of tree species, with an overall accuracy of 62.8%. LiDAR data, especially FWF LiDAR, presented a great potential for the characterization of tropical forests, especially using the voxel-based approach as implemented in software DASOS. Combining data from PR LiDAR with vegetation indices from hyperspectral images improved the accuracy of classification.

Keywords: tropical forests, forest modelling, forest inventory, diversity, airborne laser scanner, peak returns, full-waveform analysis, software DASOS, machine learning, R programming, hyperspectral remote sensing, tree species identification.

.

RESUMO

O objetivo principal desta tese foi explorar os dados LiDAR para estimar variáveis de inventário florestal e de diversidade em um remanescente de Floresta Estacional Semidecidual, também conhecida como Mata Atlântica de interior, uma das fitofisionomias mais degradadas da Mata Atlântica. Além disso, a fusão de dados LiDAR com imagens hiperespectrais para classificação de espécies do dossel superior também foi abordada. Os dados LiDAR foram entregues como retornos de pico (PR) e forma de onda completa (FWF) e em duas alturas de voo (900 m e 2000 m), e as imagens hiperespectrais foram coletadas por uma câmera leve de baixo custo. Apesar das intervenções ao longo dos anos a área de estudo, o remanescente florestal Ponte Branca apresenta bom estado de conservação, com presença de espécies pioneiras até clímax. Foram extraídas métricas LiDAR com os PR no voo de 900 m de altura. Métodos de seleção de variáveis e técnicas de aprendizado de máquina foram testados para estimar parâmetros de inventário florestal e de diversidade. As novas variáveis geradas pela transformação das métricas por meio da Análise de Componentes Principais (PCA) e a técnica das redes neurais artificiais (ANN) foi a combinação que melhor estimou os parâmetros florestais. A comparação dos atributos extraídos por uma abordagem baseada em voxel, disponível no *software* DASOS, foi para dados de PR e FWF. A comparação das estimativas de variáveis florestais com dados obtidos nas duas diferentes alturas de voo também foi realizada. Os dados FWF estimaram as variáveis com menos tendência que os dados de PR. Oito espécies presentes no dossel superior foram classificadas usando os seguintes dados: PR LiDAR, FWF LiDAR e imagens hiperespectrais, usados isoladamente e combinados entre si. As métricas PR LiDAR transformadas pelas PCA combinados com os índices de vegetação obtidos por imagens hiperespectrais, proporcionaram a melhor precisão na classificação das espécies arbóreas, com uma acurácia global de 62,8%. Os dados LiDAR, especialmente os FWF LiDAR, apresentaram um grande potencial para a caracterização de florestas tropicais, especialmente usando a abordagem baseada em voxel implementada no *software* DASOS. A combinação de dados PR LiDAR com os índices de vegetação das imagens hiperespectrais melhorou a precisão da classificação.

Palavras-chave: florestas tropicais, modelagem florestal, inventário florestal, diversidade, varredura laser aerotransportada, retornos de pico, análise da forma de onda completa, *software* DASOS, aprendizado de máquina, programação em R, sensoriamento remoto hiperespectral, identificação de espécies arbóreas.

LIST OF FIGURES

CHAPTER 1

Figure 1 – Natural occurrence of Brazilian Atlantic Forest. Adapted from SOS Mata Atlântica; Inpe, (2019)	5
Figure 2 – Remnant of Atlantic Forest in São Paulo state. Adapted from SOS Mata Atlântica; Inpe, (2019)	7

CHAPTER 2

Figure 1 – Location of the study area in the State of São Paulo and aerial image (RGB) of <i>Ponte Branca</i> forest remnant where T1 to T8 represent the transects and P1 to P7 the plots. Adapted from Berveglieri et al., (2018).	18
Figure 2 – Species accumulation curve with a boxplot of each plot. The dotted lines represent the confidence interval of 95%.	20
Figure 3 – Diameter distribution for all trees and for species with higher IVIs.	24
Figure 4 – (a) Canopy height model of <i>Ponte Branca</i> forest remnant. (b) Frequency distribution of pixels referring to tree heights.	25

CHAPTER 3

Figure 1 – Location of the study area and the canopy height model representing the tree heights inside the <i>Ponte Branca</i> forest remnant.	36
Figure 2 – Airborne laser scanner (ALS) point cloud representing the vertical structure of a plot of the <i>Ponte Branca</i> forest remnant. (a) Three-dimensional view of the plot. (b) Cross section of the same plot.	39
Figure 3 – Correlogram between the 54 extracted LiDAR metrics.	47
Figure 4 – (a) First five principal components (PCs) and the percentage of variance explained by each one. (b) Projection of Table 5. PCs and the LiDAR metrics.	49
Figure 5 – (a) RMSE in percentage for the regression methods tested for the seven estimated forest variables; (b) Bias in the percentage of each modelled variable for each regression method tested.	51
Figure 6 – Relative importance of PCs for each forest variable modelled by the best-selected regression method.	54

CHAPTER 4

Figure 1 – Representation of a modelled waveform. The light blue dots are the LiDAR pulses, and the dark blue line is the Gaussian fitted model. Adapted from Ducic et al. (2006).	78
---	----

Figure 2 – LiDAR pulses emitted at different scanning angles intercepting the voxelized space. Wang and Glennie, (2015).....	80
Figure 3 – Location of the study area. (a) Map of Brazil with emphasis on the state of São Paulo. (b) West of the state of São Paulo. (c) <i>Ponte Branca</i> forest remnant with the location of the surveyed plots in the different successional stages found. Adapted from Berveglieri et al. (2018); Martins-Neto et al. (2021).	81
Figure 4 – Representation of LiDAR datatypes. The blue dots are the discrete peak returns peaks, and the solid black line is the FWF data that is digitized at intervals of 1 ns as denoted by the red dotted lines.....	84
Figure 5 – Flowchart of LiDAR data processing to extract the forest variables.....	85
Figure 6 – Point clouds at different heights of flights and division of the area into voxels of equal size and voxels with different intensities.	87
Figure 7 – (a) Point cloud representation of a plot with an established center and a radius that contains the plot area. (b) Plot after the voxelization process and feature vector extraction from a cuboid three-dimensional window, with the movement of the top window to the first non-empty voxel of the middle column. Adapted from Miltiadou et al. (2020).	89
Figure 8 – Pixel value computed from the thickness information in a column of voxels. Adapted from Miltiadou et al., 2019.....	92
Figure 9 – Maps containing the distance between the first and the last non-empty voxel for PR and FWF data with different noise filtering thresholds at flight heights of 900 m and 2000 m. Details of waveform samples at different thresholds and flight heights are also shown.....	93
Figure 10 – Percentage of importance of the attributes extracted by DASOS in the best scenarios for the estimation of the four forest variables.....	97

CHAPTER 5

Figure 1 – Canopy height models of <i>Ponte Branca</i> forest remnant. (a) All tree heights. (b) Lower stratum. (c) Middle stratum. (d) Upper stratum.	118
Figure 2 – Targets located near the overflown area. In red the radiometric targets, and in blue the GCPs targets.	122
Figure 3 – PR LiDAR data processing flowchart.....	123
Figure 4 – FWF LiDAR data processing flowchart.	124
Figure 5 – Hyperspectral images flowchart. Adapted from Moriya et al. (2017) and Näsi et al. (2015).	125

Figure 6 – Top view of a voxelised tree, with the ITC (in green), the centroid calculated from the ITC (in blue) and the 3D window with 5 voxels in X and Y.....	129
Figure 7 – Individual tree crowns samples (in green) for each species identified in the field and delineated in hyperspectral orthomosaic (R: 780.49 nm; G: 650.96 nm; B: 535.09 nm). The yellow dots are the centroids calculated from the ITCs.	130
Figure 8 – (a) Percentage of variance explained by each PC. (b) Projection of PR LiDAR metrics in each PCs and their respective contribution.	133
Figure 9 – Mean spectra of each tree species with standard error bars.	134
Figure 10 – Importance of variables for classification of tree species in scenario "d".	137
Figure B1 – Spectra obtained from each tree for each species, from hyperspectral orthomosaics.	150

LIST OF TABLES

CHAPTER 2

Table 1 – Phytosociological parameters of the tree species sampled in the <i>Ponte Branca</i> forest remnant.	21
--	----

CHAPTER 3

Table 1 – Statistics about forest variables calculated from field data.	37
Table 2 – Light detection and ranging (LiDAR) metrics extracted from normalized point clouds.	42
Table 3 – Summary of the architectures adopted for modelling using artificial neural networks (ANNs).	44
Table 4 – Summary of the parameters adopted for modelling using epsilon regression type-support vector machine (ϵ -SVM).	46
Table 5 – Selection of the best model for each estimated forest variable.	50
Table A1 – Statistics of forest variables calculated from each of 15 surveyed plots.	70
Table A2 – The best multiple linear regression models fitted for the 7 forest variables using the 5 PCs and 15 uncorrelated LiDAR metrics as input data.	71

CHAPTER 4

Table 1 – Descriptive statistics of measured and derived forest variables.	82
Table 2 – Technical specifications of Riegl LMS-Q680i equipment (RIEGL, 2012).	84
Table 3 – Attributes extracted in DASOS.	90
Table 4 – Accuracy assessment of forest variables estimated with PR and FWF data from different flight heights. The underlined values are those with the smallest bias for each variable estimated in each of the tested scenarios.	96

CHAPTER 5

Table 1 – Characteristics of selected species in the field.	119
Table 2 – Wavelengths (λ) used in each of the Rikola camera bands and their respective FWHM.	121
Table 3 – Summary of flight campaigns for images acquisition.	122
Table 4 – Vegetation indices (VIs) calculated from hyperspectral orthomosaics.	130
Table 5 – Different scenarios tested with the three different datasets with the number of features used in each test.	131
Table 6 – Classification results for each tested scenario. The best results are highlighted.	136
Table 7 – Confusion matrix for the classification of the 8 tree species in the best scenario.	136

LIST OF SYMBOLS AND ABBREVIATIONS

AD	Absolute density
AGB	Aboveground biomass
AIC	Akaike information criterion
ALS	Airborne laser scanner
ANN	Artificial neural network
AnPe	<i>Anadenanthera peregrina</i>
APG	Angiosperm phylogeny group
ApLe	<i>Apuleia leiocarpa</i>
AsPo	<i>Aspidosperma polyneuron</i>
BA	Basal area
BBA	Bundle block adjustment
BRDF	Bidirectional reflectance distribution function
CHM	Canopy height model
cm	Centimeter
CoLa	<i>Copaifera langsdorffii</i>
D	Simpson diversity index
DBH	Diameter at breast height
DEN	Density
DN	Digital numbers
DR	Discrete return
DSM	Digital surface model
DTM	Digital terrain model
EOPs	Exterior orientation parameters
EVA	Ethylene vinyl acetate
FWF	Full-waveform
FWHM	Full width at half maximum
GCP	Ground control points
GNSS	Global navigation satellite system
GSD	Ground sample distance
H	Height
H'	Shannon-Weaver index
ha	Hectare
HeAp	<i>Helietta apiculata</i>
Hm	Average height
HyCo	<i>Hymenaea courbaril</i>
IOPs	Interior orientation parameters
ITC	Individual tree crowns
IVI	Importance value index
J	Pielou equability evenness index
km	Kilometers
LiDAR	Light detection and ranging
LOOCV	Leave-one-out cross-validation
m	Meters
MCARI	Modified Chlorophyll Absorption in Reflectance Index
MDBH	Mean diameter at breast height
Mg	Megagram
ms	Miliseconds

N	Number of trees
ND	Normalized Difference 682/553
NDVI	Normalized Difference Vegetation Index
NDVI _h	Normalized Difference 780/550 Green NDVI hyper
nl	Noise level
nm	Nanometers
ns	Nanoseconds
NTS	Number of tree species
OA	Overall accuracy
OLS	Ordinary least-squares multiple regression
PA	Producer's accuracy
PC	Principal component
PCA	Principal component analysis
PR	Peak returns
PRI	Photochemical Reflectance Index
PSRI	Plant Senescence Reflectance Index
PSSR	Pigment Specific Simple Ratio
PtPu	<i>Pterodon pubescens</i>
QMD	Quadratic mean diameter
R ²	Coefficient of determination
RD	Relative density
RDo	Relative dominance
RE	Red edge
RENDVI	Red Edge Normalized Difference Vegetation Index
REP	Red edge position
RF	Relative frequency
RF	Random forest
RMSE	Root-mean-square error
s	Seconds
SIPI	Structure Insensitive Pigment reflectance index
SVM	Support vector machine
SWIR	Short-wave infrared
SyRo	<i>Syagrus romanzoffiana</i>
TB&C	Tropical Biomass & Carbon
TIN	Triangular irregular network
UA	User's accuracy
UAV	Unmanned aerial vehicle
VI	Vegetation index
VNIR	Visible-near infrared
Voxel	Volume element
ε-SVM	Epsilon support vector machine
κ	Cohen's Kappa index
λ	Wavelengths
μm	Micrometers

CONTENTS

CHAPTER 1 – INTRODUCTION

1 OVERVIEW AND MOTIVATION.....	5
2 HYPOTHESIS AND OBJECTIVES	10
3 CONTENTS AND CONTRIBUTIONS.....	11
4 INTERNATIONAL COOPERATION	12
5 REFERENCES	13

CHAPTER 2 – STRUCTURE AND TREE DIVERSITY OF AN INLAND ATLANTIC FOREST – CASE STUDY OF *PONTE BRANCA* FOREST REMNANT, BRAZIL

1 SCOPE	17
2 MATERIALS AND METHODS.....	17
2.1 STUDY AREA.....	17
2.2 FOREST INVENTORY	18
2.3 VEGETATION ANALYSIS	19
3 RESULTS AND DISCUSSION.....	20
4 CONCLUSIONS.....	29
5 REFERENCES	30

CHAPTER 3 – IDENTIFICATION OF SIGNIFICATIVE LIDAR METRICS AND COMPARISON OF MACHINE LEARNING APPROACHES FOR ESTIMATING STAND AND DIVERSITY VARIABLES IN HETEROGENEOUS BRAZILIAN ATLANTIC FOREST

1 INTRODUCTION	32
2 MATERIALS AND METHODS.....	35
2.1 FIELD SURVEY	35
2.2 LIDAR DATA COLLECTION	38
2.3 LIDAR DATA PROCESSING	38
2.4 INPUT DATA SELECTION	41
2.5 REGRESSION TECHNIQUES SETTINGS	43
2.6 EVALUATION AND PERFORMANCE OF TESTED MODELS.....	46
3 RESULTS.....	47
3.1 CORRELATION ANALYSIS AND PCAs.....	47
3.2 MODEL PERFORMANCE AND EVALUATION	50
3.3 IMPORTANCE OF INPUT METRICS	53
4 DISCUSSION	55
5 CONCLUSIONS.....	60
6 REFERENCES	61

APPENDIX A – CHAPTER 3

CHAPTER 4 – EXTRACTING TROPICAL FOREST VARIABLES BY VOXEL APPROACH – HOW DO LIDAR DATA TYPE AND FLIGHT HEIGHT AFFECT ESTIMATES?

1 INTRODUCTION	73
2 BACKGROUND ON FULL-WAVEFORM LIDAR DATA	77
3 MATERIALS AND METHODS.....	80

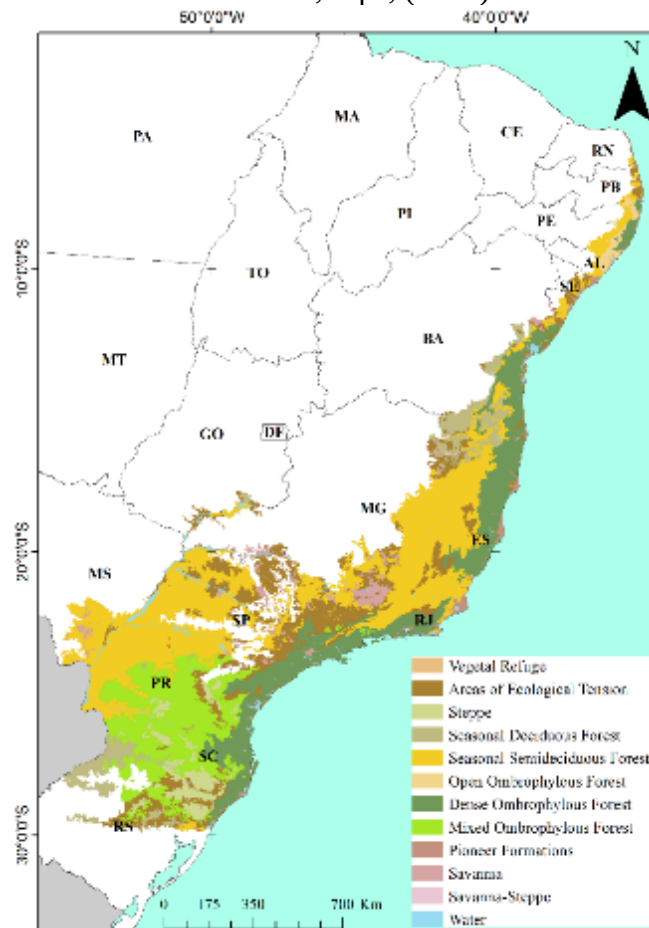
3.1 STUDY AREA.....	80
3.2 FIELD PLOT AND VARIABLES ACQUISITION.....	82
3.3 ABOVEGROUND BIOMASS ESTIMATION	82
3.4 LiDAR DATA SURVEY	83
3.5 LiDAR DATA PROCESSING	85
3.5.1 Voxelization	86
3.5.2 Extraction of feature vectors.....	88
3.6 MODELING FOREST VARIABLES	90
3.7 MODEL GOODNESS-OF-FIT	91
4 RESULTS.....	92
4.1 SELECTING NOISE LEVEL THRESHOLD.....	92
4.2 FOREST VARIABLES ESTIMATION	95
5 DISCUSSION	99
6 CONCLUSION.....	105
7 REFERENCES	107
CHAPTER 5 – CLASSIFICATION OF TREE SPECIES IN THE UPPER CANOPY OF ATLANTIC FOREST INTEGRATING HYPERSPECTRAL IMAGES AND LIDAR DATA	
1 INTRODUCTION.....	116
2 DATASETS ACQUISITION.....	117
2.1 STUDY SITE AND SAMPLES SELECTION.....	117
2.2 REMOTE SENSING DATA	120
2.2.1 LiDAR data	120
2.2.2 Hyperspectral imagery data	120
3 METHODOLOGY	122
3.1 LiDAR DATA PROCESSING	122
3.2 HYPERSPECTRAL IMAGES PROCESSING.....	124
3.3 FEATURES EXTRACTION	127
3.4 AUTOMATIC CLASSIFICATION.....	131
4 RESULTS.....	132
4.1 PRINCIPAL COMPONENT ANALYSIS AND SPECTRAL FEATURES	132
4.2 TREES SPECIES CLASSIFICATION	135
5 DISCUSSION	137
6 CONCLUSIONS.....	140
7 REFERENCES	142
APPENDIX B – CHAPTER 5	149
CHAPTER 6 – FINAL REMARKS.....	151

CHAPTER 1 – INTRODUCTION

1 OVERVIEW AND MOTIVATION

The Brazilian Atlantic Forest domain is the second largest tropical forest in America, with high biodiversity rates and a unique biota with a very rich endemism, which are second only to those of the Amazon forest (CARVALHO JÚNIOR et al., 2008; FIASCHI; PIRANI, 2009). The Atlantic domain occurs along the Brazilian coast (Figure 1), extending far inland in some areas of south and south-eastern Brazil (FIASCHI; PIRANI, 2009). It embraces a huge variety of formations and forest ecosystems with very different structures and floristic compositions, conditioned by the region's edaphoclimatic and topographic characteristics (SOS MATA ATLÂNTICA; INPE, 2019). More than 70% of the Brazilian population lives in the Atlantic forest domain; as a result, it is the most threatened biome in Brazil (SCARANO; CEOTTO, 2015).

Figure 1 – Natural occurrence of Brazilian Atlantic Forest. Adapted from SOS Mata Atlântica; Inpe, (2019)

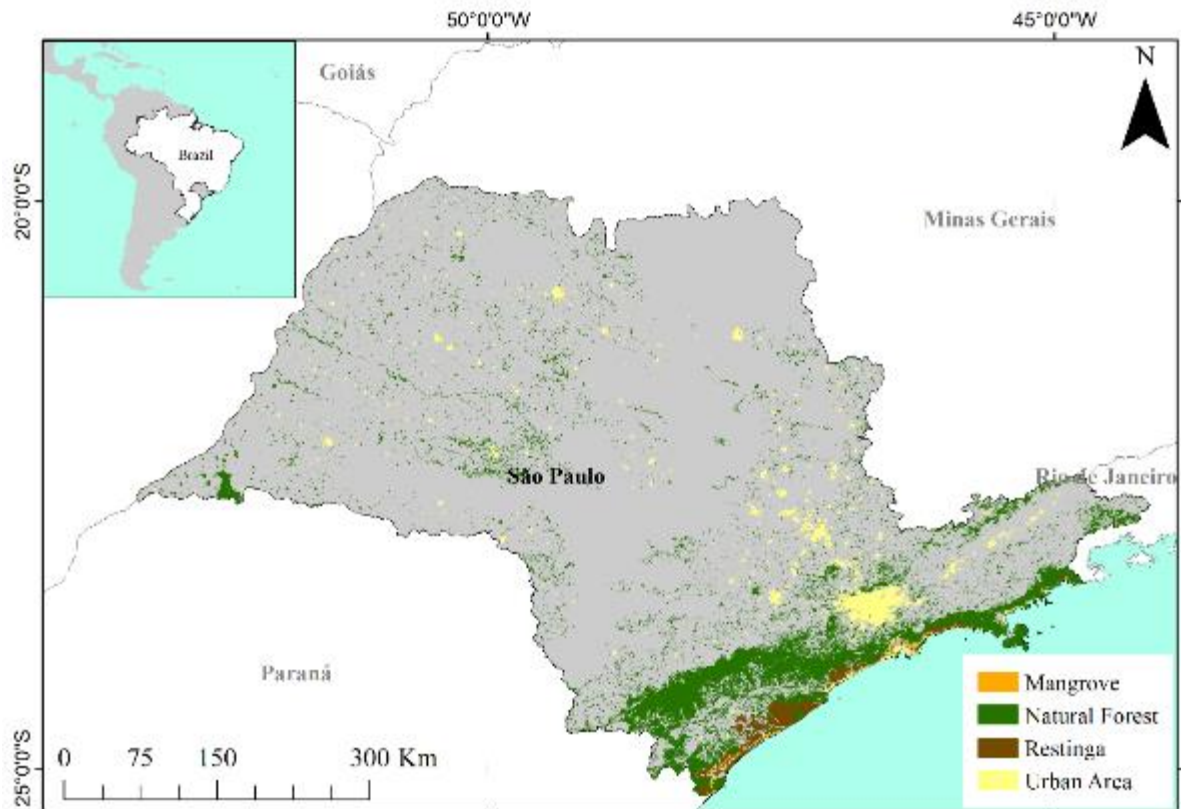


Logging, disordered urban growth, agricultural encroachment and industrialization have contributed to the deforestation and fragmentation of the Atlantic forest, reducing the original coverage to only 11.6% of the original natural environment. The remaining forest is distributed in small areas (DEAN, 1996; HADDAD et al., 2015; HARGREAVES, 2008; RIBEIRO et al., 2009) in the secondary successional stage that have been modified from their original vegetation cover (FONSECA; RIBEIRO; CARVALHO, 2013). There is biological evidence that mature forests with climax species are being replaced and dominated by pioneer species, in a process called “secondarization” (JOLY; METZGER; TABARELLI, 2014; LEAO et al., 2014; SCARANO; CEOTTO, 2015).

The current forest remnants are insufficient to maintain biodiversity, which has led to borderline situations, such as isolation of fauna and flora populations, genetic impoverishment, and growing edge effects (GUERRA et al., 2013; SCARANO; CEOTTO, 2015). Thus, the Atlantic forest is considered one of the world’s priorities for biodiversity conservation, with approximately 14000 vascular plant species distributed in 208 families, of which approximately 8000 are classified as endemic (WERNECK et al., 2011). Currently, an average of 170 unpublished species are identified each year (FIASCHI; PIRANI, 2009; MYERS et al., 2000). Therefore, this biome is a biodiversity hotspot because it has already lost more than 75% of its original coverage and still shows a high degree of endemism (MYERS et al., 2000; WILLIAMS et al., 2011).

Currently, the state of São Paulo, Brazil, has remnants of approximately 16,3% of the original area of the Atlantic Forest in the state (Figure 2). These remnants are located mainly along the coast and in the nearby areas, which are represented by dense ombrophilous forest and mixed ombrophilous forest, respectively (SOS MATA ATLÂNTICA; INPE, 2019). In the western region of the state, called *Pontal do Paranapanema*, there are a few remnants of the seasonal semideciduous forest: those in the *Morro do Diabo* state park and the *Mico-Leão-Preto* (Black Lion Tamarin - *Leontopithecus chrysopygus*) ecological station. Even though these remnants are small, they are recognized for their importance in the strategies for conservation of the Atlantic forest due to their extreme biological importance for endangered specimens of flora and fauna (MMA; IBAMA; ICMBIO, 2007).

Figure 2 – Remnant of Atlantic Forest in São Paulo state. Adapted from SOS Mata Atlântica; Inpe, (2019)



The seasonal semideciduous forest, which is also known as the inland Atlantic Forest, is influenced by dual climatic seasonality. In tropical regions, the dry and rainfall periods are well defined and have an average annual temperature of 21°C, but in the subtropical region, there is a short period of drought with a temperature that decreases to below 15°C in winter (DITT, 2002). Due to the decrease in precipitation and humidity in winter, the presence of epiphytes is not very noticeable, but these weather conditions significantly increase the number of lianas (RODERJAN et al., 2002). These climatic features affect the upper canopy layer by inducing dormancy; 20 to 50% of the trees in the forest are deciduous and lose their leaves in seasons with unfavorable conditions (CAMPANILI; SCHÄFFER, 2010).

For an understanding of these native forest remnants, structural analysis of these fragments is necessary. It includes various approaches such as floristic composition (estimates of similarity, richness, and diversity of species), determination of parameters related to horizontal and vertical structures of vegetation, and additional information related to the size and health of tree trunks, infestation of vines and identification of rare species (DE SOUZA; SOARES; SOARES, 2013).

Information about the structure of forests is obtained through forest inventories. This information is essential to infer about the development, degradation, and successional stage of a given forest fragment (VAUGHN; MOSKAL; TURNBLOM, 2012). However, field inventories are costly, expensive, and restricted to small areas or sampling techniques. Furthermore, in tropical forests, these activities are hampered by the high understory density, presence of lianas and vines, and aerial roots in these environments, resulting in difficulties for walking through the forest. Besides, measuring some tree variables like canopy area and tree heights is troublesome because of the variation in tree height and overlapping of tree crowns (MARTINS-NETO et al., 2021).

The scarcity of efficient methods for determining the forest structure limits the relationship between the forest structure and its functional characteristics at compatible spatial scales (LEFSKY et al., 2002; PARKER; RUSS, 2004; POPESCU; WYNNE; NELSON, 2002). Remote Sensing data offer the possibility of providing information about forests at different spatial and temporal scales with high geometric precision and are very useful for extracting information about forest cover, forest biomass, and forest structure (SHEN et al., 2018; TAUBERT et al., 2021).

Among the available remote sensing technologies, LiDAR (Light detection and ranging) data is a powerful tool for monitoring forest environments. The ability of the pulses to penetrate through small openings in the vegetation canopy allows obtaining three-dimensional information on the forest structure and terrain (LIM et al., 2003; SHAN; TOTH, 2018), which are used to estimate important biophysical parameters of forests (HYYPÄÄ et al., 2008; VAUHKONEN et al., 2012).

LiDAR systems can be classified according to their return type as discrete return (DR) and full waveform (FWF). Discrete return point clouds have been widely used in forestry studies, in which metrics are extracted representing the vertical structural, such as descriptive statistics of heights, intensity reflected by the object, and the number of returns for each emitted pulse. With these metrics it is possible to extract parameters of the vegetation, such as diameter at breast height, aboveground biomass, tree density and volume, and also for the classification of tree species (FALKOWSKI et al., 2009; MALTAMO; NÆSSET; VAUHKONEN, 2014; SHI et al., 2018). However, discrete return systems record information of the number of returns for each emitted pulse, limiting detection to surfaces sufficiently separated in space (CAO et al., 2016). With the technology advancement, FWF systems were developed to record and digitize the amount of energy returned to the sensor, after being backscattered by objects present

on the surface. The returned signal is sampled at equal time intervals (LIM et al., 2003; MALLET; BRETAR, 2009). More information is recorded in FWF than in DR systems, and the resulting waveform contains the properties of all elements that intersect the emitted beam path. This allows a better interpretation of the physical structure and geometric backscatter properties of the objects intercepted, presenting new opportunities for forestry studies that can improve the characterization of the forest structure (MALLET; BRETAR, 2009; PIROTTI, 2011; REITBERGER; KRZYTEK; STILLA, 2006; SHAN; TOTH, 2018).

Another remote sensing technology that has shown potential for studies in forests is hyperspectral imaging. Hyperspectral sensors offer large amounts of continuous-narrow spectral bands, which give detailed spectral signatures of trees directly related to the chemical, structural and physiological properties of foliage and canopies. As tropical forests are very complex, hyperspectral sensors provide valuable information, mainly for the identification of species and determination of forest succession (FERREIRA et al., 2016; LIESENBERG; BOEHM; GLOAGUEN, 2009; MIYOSHI et al., 2020; SANCHEZ-AZOFEIFA et al., 2017; SOTHE et al., 2019).

The recent availability of sensors at orbital level, such as the full wave form LiDAR sensor GEDI (Global Ecosystem Dynamics Investigation), and hyperspectral imaging sensors like EnMAP (Environmental Mapping and Analysis Program), PRISMA (Hyperspectral Precursor of the Application Mission) and HISUI (Hyperspectral Imager Suite), allow the knowledge of the vegetation structure at regional and continental scales. However, in fragmented landscapes, such as the Brazilian Atlantic Forest, where a large part of the remaining forests is small, it is necessary to use sensors with higher spatial resolutions that better capture the structure of these forest remnants. Thus, with data from sensors such as LiDAR and RGB cameras, multi and hyperspectral on-board on aircraft and unmanned aerial vehicles (UAV), it is possible to obtain information about the structure of these small forest remnants at scales compatible with their size.

In Brazil, the use of high-resolution remote sensing data such as LiDAR point cloud and UAV images are mainly focused on extracting information from homogeneous forest plantations. In the case of native forests, studies are mainly focused on the Amazon, due to the availability of LiDAR data provided by the Sustainable Landscapes Brazil program (<https://www.paisagenslidar.cnptia.embrapa.br/webgis/>) and the Amazon Fund (<http://www.fundoamazonia.gov.br/en/home/>). Thus, there is a lack of data that can describe the vegetation in other Brazilian forest typologies, mainly the more fragmented ones.

As mentioned before, the Atlantic Forest is the most degraded biome in Brazil, and few studies describe the structure and composition of the Atlantic Forest remnants in the interior of the state of São Paulo, since most of these conservation units are located on the coast (Figure 2). Thus, there is a potential use of LiDAR data and hyperspectral images for extracting variables to understand the composition and structure of forest environments. These data will be used to obtain information from fragmented areas, e.g., the remnants of the Brazilian Atlantic Forest, and will serve for a better knowledge of these degraded fragments and best practices of sustainable management and conservation purposes.

2 HYPOTHESIS AND OBJECTIVES

There are two hypothesis on this doctoral thesis: (i) the use of peak return and full waveform airborne LiDAR allows the estimation of variables that describe the structure and diversity of a small and heterogeneous Atlantic forest remnant, that has been disturbed over time; and (ii) the combination of LiDAR data with hyperspectral images improves the accuracy of tree species classification in complex forests. Based on these hypothesis, the following objectives were established:

- From field data, characterize the forest remnant in terms of structure, floristic composition, and diversity.
- Compare four machine learning approaches with different numbers of peak return LiDAR metrics as input data to estimate seven stand and diversity variables.
- Analyze the full waveform and verify if it is worth using it compared to the peak return data to estimate more complex variables.
- Check if different flight heights and different return types affect the estimations of forests variables.
- Test the automatic classification of tree species present in the upper canopy using LiDAR data and hyperspectral images isolated and combined to check improvements in classification accuracy.

3 CONTENTS AND CONTRIBUTIONS

This thesis is structured into six chapters, and each one has a specific objective to meet the main project issues. In **Chapter 1**, an introduction about the Atlantic Forest, its conservation status, and the need for further studies was shown. The potential of remote sensing to describe the structure of these forest remnants is discussed.

Chapter 2 describes the study area and the field forest inventories. The objective was to obtain the vertical and horizontal structure, floristic composition, and diversity of the studied forest remnant. The data obtained in this chapter served as parameters for comparison with the estimations using LiDAR data and classification using LiDAR data and hyperspectral images.

LiDAR data were explored as peak returns in **Chapter 3**. Several metrics related to the distribution of heights, intensity, and pulse returns were extracted. Input data selection methods were tested to choose the best input dataset to estimate stand and diversity variables of the studied remnant. In addition, several machine learning modeling techniques were tested to understand how they deal with the problem of low sample sufficiency and a large number of predictor variables.

Some variables were harder to estimate due to the difficulty of finding relationships with LiDAR metrics, especially in tropical forests which present a more complex structure. Thus, in **Chapter 4**, the full waveform LiDAR data were explored using a voxel-based processing technique. The objective was to compare whether full-waveform analysis improves variable estimations in tropical forests. The influence of flight height on the estimations was also analyzed and how the voxel-based analysis behaves with peak returns and full-waveform LiDAR data.

In **Chapter 5**, the classification of upper canopy species was performed, testing different datasets: hyperspectral images obtained with a lightweight camera, peak returns LiDAR data, and full-waveform LiDAR data. These datasets were used separately and combined to understand how the spectral information from the hyperspectral images with the geometric information of the LiDAR data influence the accuracy of tree species classification. Finally, in **Chapter 6**, a summary of thesis results, conclusions and future recommendations are presented.

There is an extensive literature about the application of LiDAR remote sensing in boreal and temperate forests, as well the use of hyperspectral sensors for species classification. Studies in Brazilian forests related to the themes of extracting inventory and diversity variables using

LiDAR are scarce for other biomes. Regarding tree species classification, studies are even scarcer, due to the great diversity of species that Brazilian forests have. Thus, in the thesis chapters, we always seek to base our results and discussions in tropical forests, Brazil, or elsewhere, whenever possible.

With this study we look forward to show the potential of using high-resolution remote sensing data from different sources for the characterization and extraction of structural information from small remnants of Brazilian native forests. Due to their size, these forests are often neglected when using sensors with medium and low spatial resolution.

Several well-established techniques in the literature were used to process peak return LiDAR data, and some were adapted to select the best set of attributes for estimating forest parameters. A new processing technique is presented using a voxel-based approach for analyzing the full waveform. We expect that FWF data will present new possibilities for the characterization of tropical forests.

Finally, we expect the combination of geometric data from LiDAR with spectral information from hyperspectral images will improve species classification in high complexity forests. The results of this thesis will serve for conservation decision-making and management plans. This data integration and extraction of the best set of attributes to estimate forest structural variables and species classification brings innovation and new processing perspectives for the structural understanding of small fragments of native forests.

4 INTERNATIONAL COOPERATION

This PhD thesis is a result of an international project entitled “Unmanned Airborne Vehicle-based 4D Remote Sensing for Mapping Rain Forest Biodiversity and its Change in Brazil” (UAV4D-BIO), jointly funded by São Paulo Research Foundation, Brazil (FAPESP – Grant: 2013/50426-4) and the Academy of Finland (AKA - Grant: 273806). The objective of UAV-4D-BIO was to develop technologies for biodiversity and its change mapping. The results obtained here served to understand better the studied forest remnant and the presentation of methodologies that can be tested in other forest remnants and forest typologies.

5 REFERENCES

- CAMPANILI, M.; SCHÄFFER, W. B. **Mata Atlântica: manual de adequação ambiental**. Brasília: MMA/SBF, 2010.
- CAO, L.; GAO, S.; LI, P.; YUN, T.; SHEN, X.; RUAN, H. Aboveground biomass estimation of individual trees in a coastal planted forest using full-waveform airborne laser scanning data. **Remote Sensing**, [s. l.], v. 8, n. 9, p. 729, 2016.
- CARVALHO JÚNIOR, O. A. De; COELHO, M. A. N.; MARTINS, É. de S.; GOMES, R. A. T.; COUTO JUNIOR, A. F.; OLIVEIRA, S. N. De; SANTANA, O. A. Mapeamento da vegetação na floresta atlântica usando o classificador de árvore de decisão para integrar dados de sensoriamento remoto e modelo digital de terreno. **Revista Brasileira de Geofísica**, [s. l.], v. 26, n. 3, p. 331–345, 2008.
- DE SOUZA, A. L.; SOARES, C. P. B.; SOARES, C. P. B. **Florestas nativas: estrutura, dinâmica e manejo**. [s.l: s.n.].
- DEAN, W. A ferro e fogo: a história e a devastação da Mata Atlântica brasileira. In: **A ferro e fogo: a história e a devastação da Mata Atlântica brasileira**. [s.l: s.n.]. p. 484–484.
- DITT, E. H. **Fragmentos florestais no Pontal do Paranapanema**. São Paulo: Annablume, 2002. . Acesso em: 12 nov. 2019.
- FALKOWSKI, M. J.; EVANS, J. S.; MARTINUZZI, S.; GESSLER, P. E.; HUDAK, A. T. Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. **Remote Sensing of Environment**, [s. l.], v. 113, n. 5, p. 946–956, 2009.
- FERREIRA, M. P.; ZORTEA, M.; ZANOTTA, D. C.; SHIMABUKURO, Y. E.; DE SOUZA FILHO, C. R. Mapping tree species in tropical seasonal semi-deciduous forests with hyperspectral and multispectral data. **Remote Sensing of Environment**, [s. l.], v. 179, p. 66–78, 2016.
- FIASCHI, P.; PIRANI, J. R. Review of plant biogeographic studies in Brazil. **Journal of systematics and evolution**, [s. l.], v. 47, n. 5, p. 477–496, 2009.
- FONSECA, S. N.; RIBEIRO, J. H. C.; CARVALHO, F. A. Estrutura e diversidade da regeneração arbórea em uma floresta secundária urbana (Juiz de Fora, MG, Brasil). **Floresta e Ambiente**, [s. l.], v. 20, n. 3, p. 307–315, 2013.
- GUERRA, T. N. F.; RODAL, M. J. N.; E SILVA, A. C. B. L.; ALVES, M.; SILVA, M. A. M.; DE ARAÚJO MENDES, P. G. Influence of edge and topography on the vegetation in an Atlantic Forest remnant in northeastern Brazil. **Journal of forest research**, [s. l.], v. 18, n. 2, p. 200–208, 2013.
- HADDAD, N. M.; BRUDVIG, L. A.; CLOBERT, J.; DAVIES, K. F.; GONZALEZ, A.; HOLT, R. D.; LOVEJOY, T. E.; SEXTON, J. O.; AUSTIN, M. P.; COLLINS, C. D. Habitat fragmentation and its lasting impact on Earth's ecosystems. **Science advances**, [s. l.], v. 1, n. 2, p. e1500052, 2015.

HARGREAVES, P. Phytosociology in Brazil. **J Plant Sci Biotechnol**, [s. l.], v. 2, p. 12–20, 2008.

HYYPÄ, J.; HYYPÄ, H.; LECKIE, D.; GOUGEON, F.; YU, X.; MALTAMO, M. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. **International Journal of Remote Sensing**, [s. l.], v. 29, n. 5, p. 1339–1366, 2008.

JOLY, C. A.; METZGER, J. P.; TABARELLI, M. Experiences from the Brazilian Atlantic Forest: ecological findings and conservation initiatives. **New Phytologist**, [s. l.], v. 204, n. 3, p. 459–473, 2014.

LEAO, T. C.; FONSECA, C. R.; PERES, C. A.; TABARELLI, M. Predicting extinction risk of Brazilian Atlantic Forest angiosperms. **Conservation Biology**, [s. l.], v. 28, n. 5, p. 1349–1359, 2014.

LEFSKY, M. A.; COHEN, W. B.; HARDING, D. J.; PARKER, G. G.; ACKER, S. A.; GOWER, S. T. Lidar remote sensing of above-ground biomass in three biomes. **Global ecology and biogeography**, [s. l.], v. 11, n. 5, p. 393–399, 2002.

LIESENBERG, V.; BOEHM, H.-D. V.; GLOAGUEN, R. The contribution of CHRIS/PROBA data for tropical peat swamp landscape discrimination purposes. In: 2009 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM 2009, **Anais...**: IEEE, 2009.

LIM, K.; TREITZ, P.; WULDER, M.; ST-ONGE, B.; FLOOD, M. LiDAR remote sensing of forest structure. **Progress in physical geography**, [s. l.], v. 27, n. 1, p. 88–106, 2003.

MALLET, C.; BRETAR, F. Full-waveform topographic lidar: State-of-the-art. **ISPRS Journal of photogrammetry and remote sensing**, [s. l.], v. 64, n. 1, p. 1–16, 2009.

MALTAMO, M.; NÆSSET, E.; VAUHKONEN, J. Forestry applications of airborne laser scanning. **Concepts and case studies. Manag For Ecosys**, [s. l.], v. 27, p. 2014, 2014.

MARTINS-NETO, R. P.; TOMMASELLI, A. M. G.; IMAI, N. N.; DAVID, H. C.; MILTIADOU, M.; HONKAVAARA, E. Identification of Significant LiDAR Metrics and Comparison of Machine Learning Approaches for Estimating Stand and Diversity Variables in Heterogeneous Brazilian Atlantic Forest. **Remote Sensing**, [s. l.], v. 13, n. 13, p. 2444, 2021.

MIYOSHI, G. T.; IMAI, N. N.; GARCIA TOMMASELLI, A. M.; ANTUNES DE MORAES, M. V.; HONKAVAARA, E. Evaluation of hyperspectral multitemporal information to improve tree species identification in the highly diverse atlantic forest. **Remote Sensing**, [s. l.], v. 12, n. 2, p. 244, 2020.

MMA; IBAMA; ICMBIO. **Plano de manejo da Estação Ecológica Mico - Leão - Preto.**, 2007. Disponível em: <http://www.icmbio.gov.br/portal/images/stories/imgs-unidades-coservacao/ese_mico_leao_preto.pdf>. Acesso em: 6 nov. 2019.

MYERS, N.; MITTERMEIER, R. A.; MITTERMEIER, C. G.; DA FONSECA, G. A.; KENT, J. Biodiversity hotspots for conservation priorities. **Nature**, [s. l.], v. 403, n. 6772, p. 853, 2000.

PARKER, G. G.; RUSS, M. E. The canopy surface and stand development: assessing forest canopy structure and complexity with near-surface altimetry. **Forest Ecology and Management**, [s. l.], v. 189, n. 1–3, p. 307–315, 2004.

PIROTTI, F. Analysis of full-waveform LiDAR data for forestry applications: a review of investigations and methods. **iForest-Biogeosciences and Forestry**, [s. l.], v. 4, n. 3, p. 100, 2011.

POPESCU, S. C.; WYNNE, R. H.; NELSON, R. F. Estimating plot-level tree heights with lidar: local filtering with a canopy-height based variable window size. **Computers and electronics in agriculture**, [s. l.], v. 37, n. 1–3, p. 71–95, 2002.

REITBERGER, J.; KRZYTEK, P.; STILLA, U. Analysis of full waveform lidar data for tree species classification. **International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 36, n. Part 3, p. 228–233, 2006.

RIBEIRO, M. C.; METZGER, J. P.; MARTENSEN, A. C.; PONZONI, F. J.; HIROTA, M. M. The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. **Biological conservation**, [s. l.], v. 142, n. 6, p. 1141–1153, 2009.

RODERJAN, C. V.; GALVÃO, F.; KUNIYOSHI, Y. S.; HATSCHBACH, G. G. As unidades fitogeográficas do estado do Paraná, Brasil. **Ciência & Ambiente**, [s. l.], v. 24, n. 1, p. 75–92, 2002.

SANCHEZ-AZOFEIFA, A.; ANTONIO GUZMÁN, J.; CAMPOS, C. A.; CASTRO, S.; GARCIA-MILLAN, V.; NIGHTINGALE, J.; RANKINE, C. Twenty-first century remote sensing technologies are revolutionizing the study of tropical forests. **Biotropica**, [s. l.], v. 49, n. 5, p. 604–619, 2017.

SCARANO, F. R.; CEOTTO, P. Brazilian Atlantic forest: impact, vulnerability, and adaptation to climate change. **Biodiversity and Conservation**, [s. l.], v. 24, n. 9, p. 2319–2331, 2015.

SHAN, J.; TOTH, C. K. **Topographic laser ranging and scanning: principles and processing**. [s.l.] : CRC press, 2018.

SHEN, X.; CAO, L.; CHEN, D.; SUN, Y.; WANG, G.; RUAN, H. Prediction of forest structural parameters using airborne full-waveform LiDAR and hyperspectral data in subtropical forests. **Remote Sensing**, [s. l.], v. 10, n. 11, p. 1729, 2018.

SHI, Y.; WANG, T.; SKIDMORE, A. K.; HEURICH, M. Important LiDAR metrics for discriminating forest tree species in Central Europe. **ISPRS journal of photogrammetry and remote sensing**, [s. l.], v. 137, p. 163–174, 2018.

SOS MATA ATLÂNTICA; INPE. **Atlas dos Remanescentes Florestais da Mata Atlântica: Relatório técnico período 2017 - 2018.**, 2019. Disponível em: <https://www.sosma.org.br/wp-content/uploads/2019/05/Atlas-mata-atlantica_17-18.pdf>. Acesso em: 6 nov. 2019.

SOTHE, C.; DALPONTE, M.; ALMEIDA, C. M. De; SCHIMALSKI, M. B.; LIMA, C. L.; LIESENBERG, V.; MIYOSHI, G. T.; TOMMASELLI, A. M. G. Tree Species Classification in a Highly Diverse Subtropical Forest Integrating UAV-Based Photogrammetric Point Cloud and Hyperspectral Data. **Remote Sensing**, [s. l.], v. 11, n. 11, p. 1338, 2019.

TAUBERT, F.; FISCHER, R.; KNAPP, N.; HUTH, A. Deriving tree size distributions of tropical forests from LiDAR. **Remote Sensing**, [s. l.], v. 13, n. 1, p. 131, 2021.

VAUGHN, N. R.; MOSKAL, L. M.; TURNBLOM, E. C. Tree species detection accuracies using discrete point lidar and airborne waveform lidar. **Remote sensing**, [s. l.], v. 4, n. 2, p. 377–403, 2012.

VAUHKONEN, J.; ENE, L.; GUPTA, S.; HEINZEL, J.; HOLMGREN, J.; PITKÄNEN, J.; SOLBERG, S.; WANG, Y.; WEINACKER, H.; HAUGLIN, K. M. Comparative testing of single-tree detection algorithms under different types of forest. **Forestry**, [s. l.], v. 85, n. 1, p. 27–40, 2012.

WERNECK, M. de S.; SOBRAL, M. E. G.; ROCHA, C. T. V.; LANDAU, E. C.; STEHMANN, J. R. Distribution and endemism of angiosperms in the Atlantic Forest. **Natureza & Conservação**, [s. l.], v. 9, p. 188–193, 2011.

WILLIAMS, K. J.; FORD, A.; ROSAUER, D. F.; DE SILVA, N.; MITTERMEIER, R.; BRUCE, C.; LARSEN, F. W.; MARGULES, C. Forests of East Australia: the 35th biodiversity hotspot. In: **Biodiversity hotspots**. [s.l.] : Springer, 2011. p. 295–310.

CHAPTER 2 – STRUCTURE AND TREE DIVERSITY OF AN INLAND ATLANTIC FOREST – CASE STUDY OF *PONTE BRANCA* FOREST REMNANT, BRAZIL

This chapter was submitted to “*Indonesian Journal of Geography*”.

<https://journal.ugm.ac.id/ijg>

1 SCOPE

This chapter describes the forest inventories realized in the *Ponte Branca* forest remnant. The objective was to characterize the composition, floristic diversity and forest structure of the tree component in the study. Data collected and forest variables extracted from field surveys served as validation of estimated forest parameters using LiDAR metrics (Chapter 3 and Chapter 4) and for tree species classification combining LiDAR data and hyperspectral imagery (Chapter 5).

2 MATERIALS AND METHODS

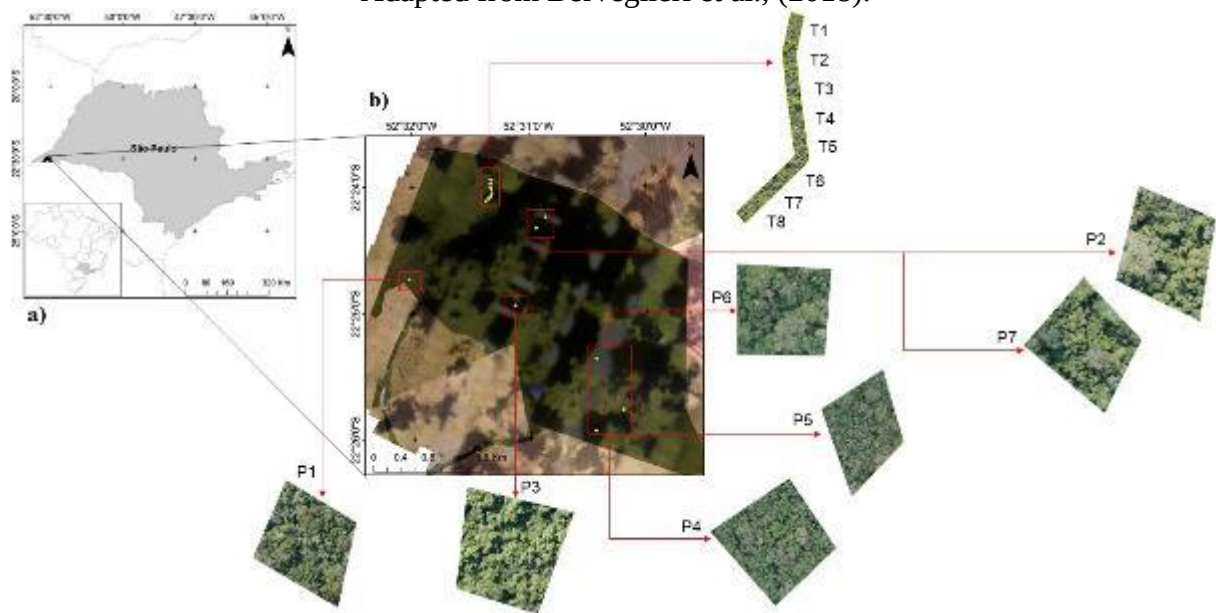
2.1 STUDY AREA

The *Mico-Leão-Preto* ecological station is a federal protected reserve located in the county of Euclides da Cunha Paulista, which is in western São Paulo state, Brazil (Figure 1), and consists of four remnants: *Santa Maria*, *Água Sumida*, *Ponte Branca* and *Tucano*. The study was conducted in the *Ponte Branca* forest remnant, which has an area of approximately 13 km². This area is predominantly covered by forest in the early to late successional stages that has previously experienced fires and deforestation. The surrounding vegetation is composed of pasture and sugar cane plantations (BERVEGLIERI et al., 2016). This region is known as *Pontal do Paranapanema*, and it is a transition zone between inland Atlantic forest and the Savanna, where species of both domains share the space in variable proportions throughout the transition zone (VELOSO; RANGEL FILHO; LIMA, 1991).

The field data collection was performed between 2014 and 2016. The location of the plots was based on the interpretation of aerial photographs of the study area from 1962 and 2010. The objective was to include the areas that suffered disturbances over time and the areas that remained preserved, so it was possible to cover all successional stages in the area from the initial successional stage, which is characterized by pioneer species, to the advanced successional stage, which is characterized by climax species. Pioneer species are undemanding

in terms of the environmental conditions and resources and are fast-growing (LAURANCE et al., 2006). Climax species present long-life cycles and slow growth, and they are very dependent on the environmental conditions and resources and the associated fauna for seed dispersal (LEAO et al., 2014; SOBRAL-SOUZA et al., 2017).

Figure 1 – Location of the study area in the State of São Paulo and aerial image (RGB) of *Ponte Branca* forest remnant where T1 to T8 represent the transects and P1 to P7 the plots. Adapted from Berveglieri et al., (2018).



According to the Köppen climate classification, the climate of the *Ponte Branca* Forest remnant is defined as Aw – tropical savanna climate with dry winter. The average precipitation is 1382 mm, which mostly occurs during the warmest months (September to April), when temperatures reach up to 40°C. In winter, the temperatures range from 15°C to 20°C. Due to the low precipitation in the driest months (May to August), 20–50% of the species in the forest lose their leaves (DITT, 2002), and the elevation ranges from 336 m in the southern part (more humid, alluvial forest) to 440 m in northern part (submontane forest) (BERVEGLIERI et al., 2016).

2.2 FOREST INVENTORY

For arboreal stratum sampling, 15 plots were established with two different designs (Figure 1): seven square plots measuring approximately 40 m x 40 m (P1 to P7) and a transect composed of 8 contiguous plots with sizes of 80 m x 20 m (T1 to T8). The corners of these

plots were located using a measuring tape, and their coordinates were acquired by a double frequency GNSS (global navigation satellite system) receiver (HIPER). All plots were at a minimum distance of 100 m from the forest borders to avoid edge effects. The sampling area corresponds to 0.2% of the total forest area.

All trees with diameter at breast height (DBH – measured at 1.30 m above the ground) larger than 3.5 cm were measured with a caliper, labelled, and identified at the species level, whenever possible (or at least to the genus level). The identification of species was made based on the dendrological attributes of the trees by a specialist in the identification of species in our study area, which was carried out in all field surveys and consultations in the relevant literature. The classification system used was APG IV (CHASE et al., 2016).

Due to the high density of trees and the large abundance of lianas, which hindered a clear view of the treetops, the variable tree heights were not measured, instead the height was obtained from the canopy height model (CHM) of the area obtained from airborne laser scanner (ALS) data. The ALS data were collected in 2017 at a flight height of 900 m using Riegl LMS – Q680i equipment, which is a full-waveform LiDAR. The density of the ALS point cloud was 18.1 points m⁻², and the data processing consisted of classification of the ground and nonground (vegetation) points and was used to generate a digital terrain model (DTM) and digital surface model (DSM) of the area. The DTM and DSM were used to compute the tree heights and to produce a CHM with a ground sample distance (GSD) of 0.50 m.

2.3 VEGETATION ANALYSIS

Data analysis was performed in the R environment (R CORE TEAM, 2017) with the vegan package version 2.5 -5 (OKSANEN, 2019). The statistical significance of the samples was checked to characterize the *in situ* floristic composition. The species accumulation curve was built by using a method that randomizes the samples for a number of selected iterations and plots the species accumulation curve and then fits a model to the curve (OKSANEN, 2019; UGLAND; GRAY; ELLINGSEN, 2003). In this study, a threshold of 1000 iterations were used.

The number of trees (N), average DBH (DBHm), total basal area (BA_t) were calculated for each species sampled along with the following phytosociological parameters as described by Mueller-Dombois; Ellenberg, (1974) and presented in Freitas; Magalhães, (2012): absolute

density (AD), relative density (RD), relative frequency (RF), relative dominance (RDo), and importance value index (IVI).

The diversity of the tree species in the whole sample was expressed by the Shannon-Weaver index (H'), which considers the logarithmic transformation of species density, and it is most influenced by low density or “rare” species in the sample (MAGURRAN, 1988). The Pielou equability evenness index (J) was calculated to estimate the uniformity of the community, measuring how close the diversity H' obtained would be to maximum hypothetical H' diversity (BROWER; ZAR; VON ENDE, 1984; MAGURRAN, 2013; MOREIRA; CARVALHO, 2013).

The diametric distribution was analyzed for all species sampled from the *Ponte Branca* Forest remnant, and for those with higher IVI values. The number of diameter classes was calculated by Sturges formula (Equation (1)) and the range of each class by Equation (2).

$$k = 1 + 3.322 * \log_{10}(N) \quad (1)$$

where k is the number of classes and N is the number of observations.

$$a = \frac{\text{Upper limit} - \text{Inferior limit}}{k} \quad (2)$$

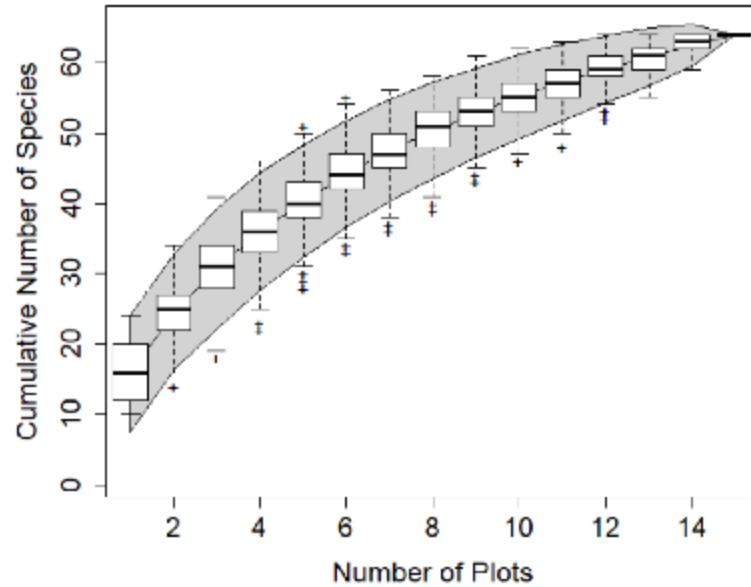
where a is the range of each class.

3 RESULTS AND DISCUSSION

In the *Ponte Branca* forest remnant, four successional stages were found: initial stage, which showed a predominance of primary species and a very low number of secondary and climax trees; initial–medium stage, which was characterized by primary species and growing secondary and climax species in the understory; medium-advanced stage, which had a very high abundance of secondary and climax species that had not yet reached maturity; and mature forested areas, which had trees that had reached the highest canopy layer, such as tall secondary and climax trees (BERVEGLIERI et al., 2016).

We verified a near stabilization with the increase in the density of sampled individuals (Figure 2). This meant that if more individuals were sampled, there would be no significant increment in the number of species, suggesting that the number of samples was adequate.

Figure 2 – Species accumulation curve with a boxplot of each plot. The dotted lines represent the confidence interval of 95%.



A total of 3181 trees were sampled, which corresponded to a density of 1277 ind ha⁻¹, and were distributed in 29 families and 64 species. Of these species, 58 were identified to the species level, and six were undetermined. Forty-seven percent of the total sampled species were classified as pioneer, 42% as secondary and 11% as climax. The families with the highest species richness were Fabaceae (12 species), Myrtaceae (7 species) and Apocynaceae (5 species). Two families (Bignoniaceae and Rutaceae) presented three species each. The families Anacardiaceae and Meliaceae presented two species each, and the other families were represented by only one species (Table 1).

Table 1 – Phytosociological parameters of the tree species sampled in the *Ponte Branca* forest remnant.

FAMILY/Species	N	DBHm (cm)	BAt (m ²)	AD (ind.ha ⁻¹)	RD (%)	RF (%)	RDo (%)	IVI (%)
ANACARDIACEAE								
<i>Astronium graveolens</i>	5	14.60	0.1529	2.0080	0.1572	1.2712	0.3583	0.60
<i>Tapirira guianensis</i>	5	9.52	0.0500	2.0080	0.1572	1.2712	0.1172	0.52
APOCYNACEAE								
<i>Aspidosperma cylindrocarpon</i>	40	10.53	0.4025	16.0638	1.2575	1.2712	0.9432	1.16
<i>Aspidosperma macrocarpon</i>	56	12.50	0.9106	22.4894	1.7605	5.0847	2.1337	2.99
<i>Aspidosperma polyneuron</i>	21	41.84	4.0188	8.4335	0.6602	2.9661	9.4162	4.35
<i>Aspidosperma ramiflorum</i>	3	12.99	0.0474	1.2048	0.0943	1.2712	0.1111	0.49
<i>Aspidosperma riedelii</i>	2	10.19	0.0171	0.8032	0.0629	0.4237	0.0400	0.18
ARALIACEAE								
<i>Dendropanax cuneatus</i>	672	8.68	5.2530	269.8724	21.1254	5.9322	12.3079	13.12
ARECACEAE								
<i>Syagrus romanzoffiana</i>	120	22.04	4.8289	48.1915	3.7724	2.9661	11.3144	6.02
ASTERACEAE								
<i>Gochnatia polymorpha</i>	12	24.26	0.6444	4.8191	0.3772	0.8475	1.5099	0.91

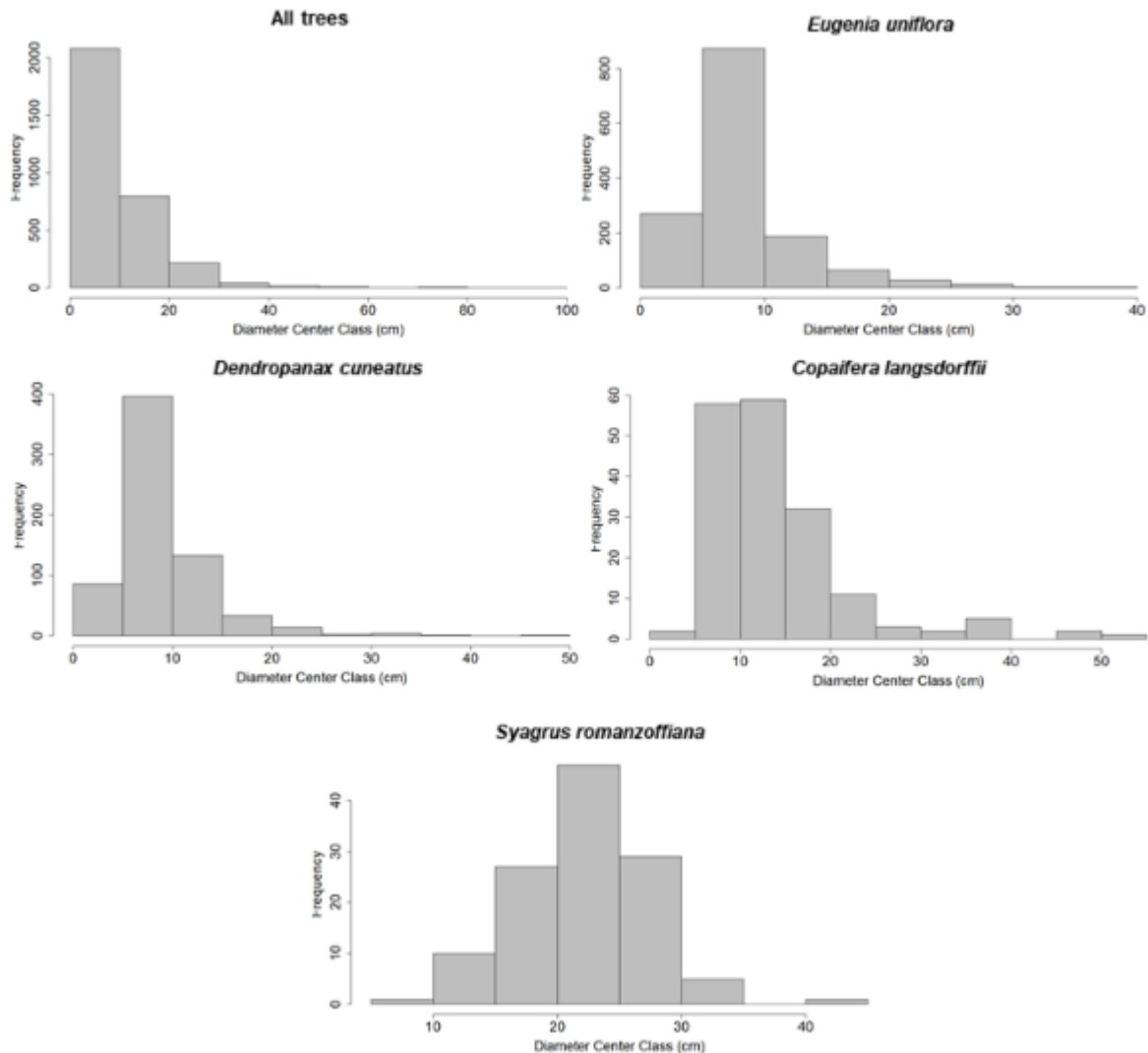
FAMILY/Species	N	DBHm (cm)	BAt (m ²)	AD (ind.ha ⁻¹)	RD (%)	RF (%)	RDo (%)	IVI (%)
BIGNONIACEAE								
<i>Tabebuia heptaphylla</i>	3	16.50	0.0788	1.2048	0.0943	0.4237	0.1846	0.23
<i>Tabebuia ochracea</i>	1	5.41	0.0023	0.4016	0.0314	0.4237	0.0054	0.15
<i>Zeyheria tuberculosa</i>	2	28.49	0.1536	0.8032	0.0629	0.8475	0.3599	0.42
BORAGINACEAE								
<i>Cordia trichotoma</i>	3	12.02	0.0347	1.2048	0.0943	1.2712	0.0813	0.48
BURSERACEAE								
<i>Patagonula americana</i>	1	6.20	0.0030	0.4016	0.0314	0.4237	0.0071	0.15
BURSERACEAE								
<i>Patagonula americana</i>	1	6.20	0.0030	0.4016	0.0314	0.4237	0.0071	0.15
CLUSIACEAE								
<i>Garcinia gardneriana</i>	1	18.00	0.0254	0.4016	0.0314	0.4237	0.0596	0.17
COMBRETACEAE								
<i>Terminalia glabrescens</i>	2	7.01	0.0087	0.8032	0.0629	0.4237	0.0204	0.17
EUPHORBIACEAE								
<i>Actinostemon concolor</i>	39	8.42	0.2581	15.6622	1.2260	1.6949	0.6048	1.18
<i>Croton floribundus</i>	10	11.69	0.1199	4.0160	0.3144	2.1186	0.2808	0.90
<i>Mabea fistulifera</i>	1	11.00	0.0095	0.4016	0.0314	0.4237	0.0223	0.16
FABACEAE	-							
CAESALPINIOIDEAE								
<i>Apuleia leiocarpa</i>	22	18.22	0.7375	8.8351	0.6916	2.1186	1.7280	1.51
<i>Bauhinia forficata</i>	1	7.50	0.0044	0.4016	0.0314	0.4237	0.0104	0.16
<i>Copaifera langsdorffii</i>	175	13.92	3.5797	70.2793	5.5014	5.0847	8.3874	6.32
<i>Guibourtia hymenifolia</i>	8	21.25	0.3633	3.2128	0.2515	1.2712	0.8513	0.79
<i>Hymenaea courbaril</i>	31	22.28	1.6400	12.4495	0.9745	3.8136	3.8425	2.88
<i>Pterogyne nitens</i>	2	55.52	0.5172	0.8032	0.0629	0.4237	1.2118	0.57
FABACEAE	-							
FABOIDEAE								
<i>Machaerium aculeatum</i>	4	8.80	0.0257	1.6064	0.1257	0.8475	0.0602	0.34
<i>Machaerium scleroxylon</i>	1	43.61	0.1494	0.4016	0.0314	0.4237	0.3500	0.27
<i>Pterodon pubescens</i>	18	33.99	2.2320	7.2287	0.5659	4.2373	5.2297	3.34
FABACEAE	-							
MIMOSOIDAE								
<i>Anadenanthera peregrina</i>	41	19.42	1.9248	16.4654	1.2889	4.6610	4.5100	3.49
<i>Inga vera</i>	12	13.50	0.2372	4.8191	0.3772	1.6949	0.5557	0.88
<i>Piptadenia gonoacantha</i>	2	9.55	0.0143	0.8032	0.0629	0.8475	0.0336	0.31
FLACOURTIACEAE								
<i>Casearia gossypiosperma</i>	3	17.19	0.0888	1.2048	0.0943	1.2712	0.2080	0.52
LAURACEAE								
<i>Endlicheria paniculata</i>	6	13.99	0.1492	2.4096	0.1886	1.6949	0.3496	0.74
LECYTHIDACEAE								
<i>Cariniana estrellensis</i>	7	16.17	0.1869	2.8112	0.2201	1.2712	0.4379	0.64
MAGNOLIACEAE								
<i>Talauma ovata</i>	4	17.35	0.0982	1.6064	0.1257	0.4237	0.2301	0.256
MALVACEAE								
<i>Luehea grandiflora</i>	24	10.26	0.2572	9.6383	0.7545	2.1186	0.6026	1.16
MELIACEAE								
<i>Cedrela fissilis</i>	5	17.83	0.1389	2.0080	0.1572	1.2712	0.3255	0.58
<i>Guarea macrophylla</i>	1	23.20	0.0423	0.4016	0.0314	0.4237	0.0990	0.18
MONIMIACEAE								
<i>Mollinedia elegans</i>	9	7.04	0.0362	3.6144	0.2829	0.4237	0.0848	0.26
MORACEAE								
<i>Ficus insipida</i>	1	5.73	0.0026	0.4016	0.0314	0.4237	0.0060	0.15
MYRTACEAE								
<i>Campomanesia xanthocarpa</i>	6	8.01	0.0336	2.4096	0.1886	0.8475	0.0788	0.37
<i>Eugenia psidiiflora</i>	3	8.91	0.0198	1.2048	0.0943	0.4237	0.0464	0.19
<i>Eugenia pyriformis</i>	10	8.37	0.0705	4.0160	0.3144	1.2712	0.1651	0.58
<i>Eugenia speciosa</i>	2	5.89	0.0055	0.8032	0.0629	0.8475	0.0128	0.31
<i>Eugenia uniflora</i>	1447	7.89	9.3458	581.1091	45.4888	6.3559	21.8978	24.58
<i>Eugenia uvalha</i>	7	9.37	0.0530	2.8112	0.2201	2.1186	0.1242	0.82
<i>Myrciaria ciliolata</i>	24	8.06	0.1697	9.6383	0.7545	1.2712	0.3977	0.81
PHYTOLACCACEAE								
<i>Gallesia integrifolia</i>	1	24.51	0.0472	0.4016	0.0314	0.4237	0.1105	0.19

FAMILY/Species	N	DBHm (cm)	BAt (m ²)	AD (ind.ha ⁻¹)	RD (%)	RF (%)	RDo (%)	IVI (%)
ROSACEAE								
<i>Prunus sellowii</i>	3	6.90	0.0126	1.2048	0.0943	1.2712	0.0294	0.46
RUTACEAE								
<i>Balfourodendron riedelianum</i>	3	25.78	0.1932	1.2048	0.0943	0.8475	0.4527	0.46
<i>Helietta apiculata</i>	37	11.74	0.5059	14.8590	1.1632	4.2373	1.1853	2.19
<i>Zanthoxylum rhoifolium</i>	8	16.26	0.2006	3.2128	0.2515	1.2712	0.4701	0.66
SAPINDACEAE								
<i>Dilodendron bipinnatum</i>	92	10.19	0.9432	36.9468	2.8922	5.0847	2.2100	3.39
SAPOTACEAE								
<i>Pouteria ramiflora</i>	8	12.40	0.1200	3.2128	0.2515	2.1186	0.2812	0.88
ULMACEAE								
<i>Celtis iguanaea</i>	137	8.74	1.2937	55.0186	4.3068	0.4237	3.0311	2.58
VERBENACEAE								
<i>Vitex montevidensis</i>	6	11.78	0.0668	2.4096	0.1886	0.4237	0.1565	0.26
VOCHYSIACEAE								
<i>Qualea multiflora</i>	1	12.10	0.0115	0.4016	0.0314	0.4237	0.0269	0.16
UNIDENTIFIED								
Indeterminate 1	3	5.33	0.0068	1.2048	0.0943	0.8475	0.0158	0.32
Indeterminate 2	2	19.26	0.0587	0.8032	0.0629	0.4237	0.1376	0.21
Indeterminate 3	2	12.73	0.0259	0.8032	0.0629	0.8475	0.0606	0.32
Indeterminate 4	1	22.92	0.0413	0.4016	0.0314	0.4237	0.0967	0.18
Indeterminate 5	1	9.55	0.0072	0.4016	0.0314	0.4237	0.0168	0.16
Indeterminate 6	1	4.46	0.0016	0.4016	0.0314	0.4237	0.0037	0.15

Where: N is the number of trees; DBHm is the mean diameter at breast height; BAt is the total basal area; AD is absolute density; RD is relative density; RF is relative frequency; RDo is relative dominance and IVI is importance value index.

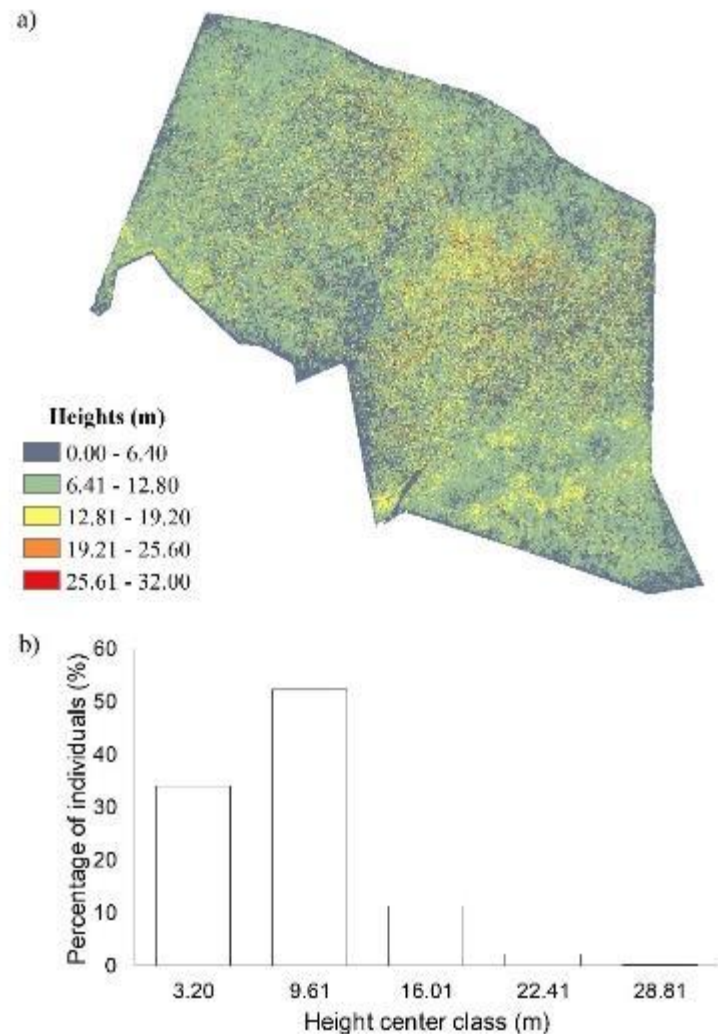
The four species with the highest IVIs were *Eugenia uniflora* (24.58%), *Dendropanax cuneatus* (13.12%), *Copaifera langsdorffii* (6.32%) and *Syagrus romanzoffiana* (6.02%). However, considering the relative levels of dominance, the *S. romanzoffiana* species had higher values than *C. langsdorffii*, since the *S. romanzoffiana* species presented more individuals with higher DBH values (DBHm of 10.31 cm), which occupied a larger basal area (BA 17.40 m² ha⁻¹) in the forest. The diameter distribution for all individuals had an inverted J-shaped curve, which is characteristic of unequal natural forests (DA ROCHA et al., 2018), with 79.8% of individuals in the smaller diameter classes (3.5 cm to 13.5 cm) and a progressively fewer individuals with increasing diameter class. The same pattern was observed for species with higher IVIs, except for *S. romanzoffiana*, which had an almost normal distribution, with most individuals in the middle diameter classes (Figure 3).

Figure 3 – Diameter distribution for all trees and for species with higher IVIs.



The value of the Shannon-Weaver diversity index (H') was 2.08. Heterogeneous forests usually have values between 1.5 and 3.5 for this index (TURNER, 1989). The value of Pielou equability evenness was 0.50, i.e., 50% of the hypothetical maximum H' was reached. Figure 4 shows the distribution of heights from the CHM of the study area with a class interval of 6.4 m. The tallest tree had a height of 31.95 m, and more than 50% of the trees were concentrated in the height classes between 6.41 m and 12.78 m.

Figure 4 – (a) Canopy height model of *Ponte Branca* forest remnant. (b) Frequency distribution of pixels referring to tree heights.



The areas in the initial successional stage (parts 1 through 5 of the transect), see Figure 1 were characterized by low-diameter trees that were smaller than 13.5 cm, a large presence of lianas and a lower canopy height, with an average of 13 m. There was a predominance of pioneer species of the Myrtaceae family, mainly *E. uniflora*, and a high density of *Dendropanax cuneatus*. However, some sparse secondary and climax tree species were found, such as *Aspidosperma polyneuron*, *Copaifera langsdorffii* and *Machaerium scleroxylon*, with DBH values between 40 cm and 50 cm.

A transition between the initial and medium successional stages was found in plots 3 and 4 (Figure 1). The average DBH of the trees measured at plots 3 and 4 was higher (up to 17 cm) than that observed in the plots in the initial regeneration stage. The upper canopy had a height between 13 and 20 m. Palm trees (*Syagrus romanzoffiana*) with DBH values ranging

from 19 cm to 42 cm were found. In the understory, individuals of the secondary and climax species, such as *Aspidosperma macrocarpon* and *Copaifera langsdorffii*, were appearing. The presence of *Aspidosperma polyneuron* in the upper stratum was also verified, indicating that there had been more canopy-height individuals of this climax species in the past. There were sparsely distributed trees with large DBHs (between 30 cm and 60 cm), such as *Anadenanthera peregrina*, *Hymenaea courbaril* and *Copaifera langsdorffii*.

Parts 6, 7 and 8 of the transect and plots 1 and 5 (see Figure 1) showed a medium-advanced successional stage. The canopy height varied between 18 m and 21 m and had a low presence of lianas. Species of the Myrtaceae family, *Dendropanax cuneatus* and the palm tree *Syagrus romanzoffiana* predominated in the understory and medium strata, with heights of approximately 8 m. A larger abundance of secondary and climax species of higher commercial value was found in the lower and middle strata, including *Aspidosperma cylindrocarpon*, *A. macrocarpon*, *A. polyneuron*, *Cariniana estrellensis* and *Copaifera langsdorffii*. The tree individuals in parts 6, 7 and 8 of the transect and plots 1 and 5 had diameters of approximately 25 cm; some representatives of the climax species had diameters larger than 35 cm. Plot 1 showed a particular feature: a riparian zone was found near a small stream with very distinct species, such as *Endlicheria paniculata*, *Mollinedia elegans*, *Talauma ovata* and *Vitex montevidensis* (MMA; IBAMA; ICMBIO, 2007), and a very dense understory of *Celtis iguanaea* (137 individuals).

Plots 2, 6 and 7 represented mature forest areas. Large trees (23 m to 29 m in height) and a continuous canopy of approximately 12 m were present. Secondary and climax species with high economic value appeared in the emergent and canopy layers, such as *Pterodon pubescens*, which was the species with the highest DBH value found in the survey (97.40 cm); *Aspidosperma polyneuron*, which had individuals with DBH > 60 cm; *Pterogyne nitens*; *Copaifera langsdorffii*; *Balfourodendron riedelianum*; *Cedrella fissilis* and *Zeyheria tuberculosa*. In the understory, many individuals of *Actinostemon concolor* (22 individuals), *Dendropanax cuneatus* (147 individuals) and *Eugenia uniflora* (306 individuals) were found.

The tree DBHs were concentrated in the first diameter class, as shown in Figure 3, forming an inverted J-shaped curve and indicating a balanced distribution trend due to the self-regenerative capacity of plant species, in which younger individuals are able to replace the older and larger diameter trees that produce larger amounts of seeds and continue forest regeneration. This concentration also suggested that the forests were now in the early stages of regeneration

due to the high density of thinner trees (HARPER, 1990; MOREIRA; CARVALHO, 2013; UHL; MURPHY, 1981).

Some typical species of Brazilian Savanna, such as *Anadenanthera peregrina*, *Celtis iguanaea* and *Qualea multiflora*, were found, which emphasized the transition between inland Atlantic forest and Savanna in western São Paulo state. All plots had high concentrations of Myrtaceae in the understory, especially *Eugenia uniflora*, which were a species with a higher IVI value (Table 1). This species had a wide distribution and a large regrowth capacity, mainly colonizing open woods in the intermediate stratum (LORENZI, 1992). As *E. uniflora* is a pioneer species, a higher concentration of individuals with smaller diameters and absence of individuals in the upper classes are expected (Figure 3). The largest *E. uniflora* tree found in the *Ponte Branca* area had a DBH of 38.20 cm. Another species that occurred frequently in the understory and that had similar characteristics was *Dendropanax cuneatus*, a widely dispersed pioneer species that inhabits mainly secondary formations and open wooded areas (LORENZI, 1992). The largest individual found had a DBH of 50 cm.

Copaifera langsdorffii is a climax species with large ecological flexibility that can be found in many habitats. Young individuals frequently appear in areas in the regeneration stage and those with secondary vegetation (CARVALHO, 2003). These species were noticed in almost all plots, and it was possible to verify in Figure 3 that there were a relevant number of trees in the middle diameter classes, indicating the succession of this species, with individuals migrating from the smaller to the larger DBH classes.

Syagrus romanzoffiana palm trees occurred in all plots that were in the initial and average regeneration stages. This is a common pioneer species in seasonal semideciduous forests and has also been found in other forest remnants in western São Paulo State, in some cases creating almost pure canopies (MMA; IBAMA; ICMBIO, 2007). The *S. romanzoffiana* fruit is much appreciated by the local fauna, which disperse it widely (LORENZI, 1992). The diameter distribution of this species differed from those of others by following the pattern of an almost Gaussian curve, with the presence of 91% of the sampled individuals in the intermediate diameter classes, which indicated a certain degree of homogeneity and development of this species within the forests. Moreover, palm trees present different growth patterns when compared to the other species: the production of the stem axis occurs from a single apical meristem, with regular production of leaves, very constant crown structure and the absence of typical secondary growth; the diameter can be decreased through death or destruction of the

external stem (stip) tissues, which allows these plants to reach higher heights relative to the diameter (TOMLINSON, 1990).

The Shannon-Weaver index computed for the studied forest remnant was lower when compared to those of other areas of seasonal semideciduous forest in the São Paulo state, as mentioned by Durigan; Santos; Gandara, (2002), who reported diversity values above 3.50, indicating different dynamics in these forest types in the different locations. For the *Ponte Branca* forest remnant, Ditt, (2002) found a diversity value of 2.29. The difference may be linked to the dynamics that occurred in the forest between the dates of the two studies (approximately 13 years of difference), in which the author showed a low abundance of pioneer trees by quadrant sampling. However, currently, there is an abundance of pioneer trees in the *Ponte Branca* forest remnant that have relevant ecological dominance, which was confirmed by the Pielou evenness, with a value of $J' = 0.50$; some species had the most individuals (Table 1) in comparison with the low densities of other species (FONSECA; RIBEIRO; CARVALHO, 2013).

The *Ponte Branca* forest remnant suffered from several anthropogenic disturbances over time that significantly changed forest coverage, including the advance of agriculture (sugarcane) and pastures, the application of agricultural chemicals in the cultivated areas and forest fires that caused edge effects. Therefore, it was difficult for the forest to regenerate in many areas. The selective logging of species of high economic value that presented high DBH values, such as *Aspidosperma polyneuron*, *Cariniana estrellensis*, *Copaifera langsdorffii*, *Hymenaea courbaril*, *Pterodon pubescens* and *Tabebuia heptaphylla*, has also intensified the deforestation process (BERVEGLIERI et al., 2016; DURIGAN; SANTOS; GANDARA, 2002). Furthermore, deforestation had led to an increase in the groundwater level in the northern region of *Ponte Branca*, where the altitude is higher, such that the greater flow of water into the ground has led to the rise of groundwater. As a result, the regions at lower altitudes have begun to present an increase in the humidity near the surface, causing changes in the forests, as seen in Plot 1, where a riparian zone was found with the species characteristics of that environment (BERVEGLIERI et al., 2016; MMA; IBAMA; ICMBIO, 2007).

Forest fragmentation has occurred in western São Paulo State, mainly during the second half of the 20th century. Most of the trees that currently form the canopy layer were established before fragmentation. In the long term, however, the ability of these forests to survive the impacts of fragmentation is unknown. The conservation of species with small populations in these forest remnants, especially those under selective logging, can seriously threatened, and

for other species, the maintenance of the structure and genetic diversity of the population in the fragments may already be seriously affected (DURIGAN; SANTOS; GANDARA, 2002).

4 CONCLUSIONS

This study showed the structure, diversity and floristic composition of a fragment of inland Atlantic Forest, in the western region of the state of São Paulo. Comparing the same forest type in other parts of the state, less diversity of trees was found in our study area. Anthropogenic intervention had a decisive influence on succession stages since there are stages of initial and intermediate succession in *Ponte Branca* Forest remnant. However, there are some areas that have remained intact, which have not suffered from deforestation and have been protected since the area became a federal protected reserve, with the presence of secondary tree species and climax with varying heights and DBHs. Despite pressure from the neighborhood (e.g. agriculture), there is a recovery, despite the fragment's territorial limitations, is still noticeable. For future studies, the use of digital image processing and photogrammetry techniques can be of great value for tropical forests, mainly due to the difficulty of obtaining field data and historical data from these areas. Therefore, to facilitate the measurement and understanding of the forest structure and dynamics, remote sensing techniques are needed for the estimation of structural attributes and their relationships, mainly for these degraded areas with lack of field data.

5 REFERENCES

- BERVEGLIERI, A.; TOMMASELLI, A. M. G.; IMAI, N. N.; RIBEIRO, E. A. W.; GUIMARAES, R. B.; HONKAVAARA, E. Identification of successional stages and cover changes of tropical forest based on digital surface model analysis. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, [s. l.], v. 9, n. 12, p. 5385–5397, 2016.
- BROWER, J. E.; ZAR, J. H.; VON ENDE, C. A. **Field and laboratory methods for general ecology**. Dubuque: Wm. C. [s.l.] : Brown Publishers, 1984.
- CARVALHO, P. E. R. **Espécies arbóreas brasileiras**. Brasília, DF: Embrapa Informação Tecnológica Brasília, 2003. v. 1
- CHASE, M. W.; CHRISTENHUSZ, M. J. M.; FAY, M. F.; BYNG, J. W.; JUDD, W. S.; SOLTIS, D. E.; MABBERLEY, D. J.; SENNIKOV, A. N.; SOLTIS, P. S.; STEVENS, P. F. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. **Botanical Journal of the Linnean Society**, [s. l.], v. 181, n. 1, p. 1–20, 2016.
- DA ROCHA, S. J. S. S.; TORRES, C. M. M. E.; JACOVINE, L. A. G.; LEITE, H. G.; GELCER, E. M.; NEVES, K. M.; SCHETTINI, B. L. S.; VILLANOVA, P. H.; DA SILVA, L. F.; REIS, L. P. Artificial neural networks: Modeling tree survival and mortality in the Atlantic Forest biome in Brazil. **Science of the total environment**, [s. l.], v. 645, p. 655–661, 2018.
- DITT, E. H. **Fragmentos florestais no Pontal do Paranapanema**. São Paulo: Annablume, 2002. . Acesso em: 12 nov. 2019.
- DURIGAN, G.; SANTOS, J. D. Dos; GANDARA, F. B. Fitossociologia de dois fragmentos de floresta estacional semidecidual no Pontal do Paranapanema, SP. **Revista do Instituto Florestal**, [s. l.], v. 14, n. 1, p. 13–26, 2002.
- FONSECA, S. N.; RIBEIRO, J. H. C.; CARVALHO, F. A. Estrutura e diversidade da regeneração arbórea em uma floresta secundária urbana (Juiz de Fora, MG, Brasil). **Floresta e Ambiente**, [s. l.], v. 20, n. 3, p. 307–315, 2013.
- FREITAS, W. K. De; MAGALHÃES, L. M. S. Métodos e parâmetros para estudo da vegetação com ênfase no estrato arbóreo. **Floresta e Ambiente**, [s. l.], v. 19, p. 520–539, 2012.
- HARPER, J. L. **Population biology of plants**. 8th impr. [s.l.] : London, UK: Academic Press, 1990.
- LAURANCE, W. F.; NASCIMENTO, H. E.; LAURANCE, S. G.; ANDRADE, A.; RIBEIRO, J. E.; GIRALDO, J. P.; LOVEJOY, T. E.; CONDIT, R.; CHAVE, J.; HARMS, K. E. Rapid decay of tree-community composition in Amazonian forest fragments. **Proceedings of the National Academy of Sciences**, [s. l.], v. 103, n. 50, p. 19010–19014, 2006.
- LEAO, T. C.; FONSECA, C. R.; PERES, C. A.; TABARELLI, M. Predicting extinction risk of Brazilian Atlantic Forest angiosperms. **Conservation Biology**, [s. l.], v. 28, n. 5, p. 1349–1359, 2014.

- LORENZI, H. Árvores brasileiras, vol. 1. **Plantarum**, Nova Odessa, [s. l.], 1992.
- MAGURRAN, A. E. **Ecological diversity and its measurement**. [s.l.] : Princeton university press, 1988.
- MAGURRAN, A. E. **Measuring biological diversity**. [s.l.] : John Wiley & Sons, 2013.
- MMA; IBAMA; ICMBIO. **Plano de manejo da Estação Ecológica Mico - Leão - Preto.**, 2007. Disponível em: <http://www.icmbio.gov.br/portal/images/stories/imgs-unidades-coservacao/esec_mico_leao_preto.pdf>. Acesso em: 6 nov. 2019.
- MOREIRA, B.; CARVALHO, F. A. A comunidade arbórea de um fragmento urbano de Floresta Atlântica após 40 anos de sucessão secundária (Juiz de Fora, Minas Gerais). **Biotemas**, [s. l.], v. 26, n. 2, p. 59–70, 2013.
- MUELLER-DOMBOIS, D.; ELLENBERG, H. **Aims and methods of vegetation ecology**. [s.l.] : Wiley, 1974.
- OKSANEN, J. **vegan: Community Ecology Package (R package Version 2.5-5)** <https://CRAN.R-project.org/package=vegan>. [s.l.] : Accessed, 2019.
- R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2017.
- SOBRAL-SOUZA, T.; LAUTENSCHLAGER, L.; MORCATTY, T. Q.; BELLO, C.; HANSEN, D.; GALETTI, M. Rewilding defaunated Atlantic Forests with tortoises to restore lost seed dispersal functions. **Perspectives in ecology and conservation**, [s. l.], v. 15, n. 4, p. 300–307, 2017.
- TOMLINSON, P. B. **The structural biology of palms**. [s.l.] : Oxford University Press, 1990.
- TURNER, M. G. Landscape ecology: the effect of pattern on process. **Annual review of ecology and systematics**, [s. l.], v. 20, n. 1, p. 171–197, 1989.
- UGLAND, K. I.; GRAY, J. S.; ELLINGSEN, K. E. The species–accumulation curve and estimation of species richness. **Journal of Animal Ecology**, [s. l.], v. 72, n. 5, p. 888–897, 2003.
- UHL, C.; MURPHY, P. G. **Composition, structure, and regeneration of a tierra firme forest in the Amazon basin of Venezuela**. [s.l.: s.n.]. v. 22
- VELOSO, H. P.; RANGEL FILHO, A. L. R.; LIMA, J. C. A. **Classificação da vegetação brasileira, adaptada a um sistema universal**. Rio de Janeiro: IBGE, 1991. Disponível em: <<https://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/classificacaovegetal.pdf>>. Acesso em: 8 nov. 2019.

CHAPTER 3 – IDENTIFICATION OF SIGNIFICATIVE LIDAR METRICS AND COMPARISON OF MACHINE LEARNING APPROACHES FOR ESTIMATING STAND AND DIVERSITY VARIABLES IN HETEROGENEOUS BRAZILIAN ATLANTIC FOREST

This chapter was published in “*Remote Sensing – MDP*” journal.

doi.org/10.3390/rs13132444

1 INTRODUCTION

Field surveys in the Brazilian tropical forests (e.g., Atlantic Forest) are laborious work. The high understory density, presence of lianas and vines, and aerial roots existent in this environment result in difficulties for the measurement of tree variables and a walk through the forest. Estimating variables that describe the forest structure, the tree species diversity, and richness are challenging tasks, because the Brazilian Atlantic forest is a very rich biome in plant species, with approximately 14000 vascular plant species, of which approximately 8000 are classified as endemic (DE SOUZA WERNECK et al., 2011). Additionally, measuring the canopy area and tree heights is troublesome because of the variation in tree height and overlapping of tree crowns. Due to these restrictions, tree height is usually estimated with the naked eye (HOPKINS, 2007). These data are often used as inputs for regression models to estimate biomass, volume, growth, and yield, but uncertainties in the field variables measurement propagate to the estimates through regression models (NOGUEIRA et al., 2008; NOGUEIRA; NELSON; FEARNSIDE, 2006).

Data from airborne laser scanner (ALS) have been widely used in forestry applications, due to their ability to penetrate through small openings (e.g., gaps between branches and leaves) in the forest canopy and collect three-dimensional information of vegetation and terrain (MILTADOU et al., 2020; SHAN; TOTH, 2018; YAO et al., 2014). ALS is based on light detection and ranging (LiDAR) technology, and with the resulting three-dimensional point cloud, it is possible to better understand the arrangement of the forest canopy, allowing the accurate estimation of structural parameters of the forest (FALKOWSKI et al., 2009). The information acquired by ALS is very valuable for forest inventory and modelling, especially for dense, complex forests that are not safe and/or easy to access.

It is possible to extract several metrics from the ALS point cloud. This includes descriptive statistics, percentiles, and distribution measures of heights, intensity, and laser pulse returns, providing a summary of the forest canopy structure. LiDAR metrics are usually used

as predictors in regression models for the estimation of forest biophysical variables (DONOGHUE et al., 2007; MORSDORF et al., 2008; NÆSSET, 2004a). Multiple linear regression is commonly used for modelling forest variables from LiDAR metrics due to its simplicity and clarity when interpreting the resulting model (GARCÍA-GUTIÉRREZ et al., 2015).

However, multiple linear regression requires the basic assumptions of classical statistics, which can be difficult to achieve when dealing with the modelling of biological data (SILVA et al., 2020). As a result, the use of computational techniques, such as machine learning, has been increased, including modelling forest inventory variables with LiDAR metrics as predictors. Machine learning techniques can model complex relationships between dependent and independent variables (i.e., a large number of LiDAR metrics) without requiring linear assumptions about the data distribution (SILVA et al., 2020; ZHAO et al., 2011). Therefore, machine learning techniques are suitable for predicting complex non-linear relationships. Additionally, interaction effects are modelled automatically which makes these methods very powerful and promising compared to multiple linear regression to estimate forest parameters from LiDAR data (SILVA et al., 2020; VENIER et al., 2019).

The use of LiDAR metrics to estimate forest variables using machine learning techniques has been used in forest plantations in Brazil. The estimate of total, commercial, and pulp volume of a *Pinus taeda* plantation was performed by Silva et al. (2017) using LiDAR metrics as input data for the random forest (RF) algorithm. The results obtained indicated a low prediction bias (average of -0.2%) and the average of the root-mean-square error (RMSE) was 8.1%. The authors concluded the use of RF to determine different types of volumes in homogeneous forests presents highly accurate estimates. In *Eucalyptus* spp. plantations, Görgens; Montagni; Rodriguez, (2015) estimated volume per stand comparing multiple linear regression with artificial neural network (ANN), RF, and support vector machine (SVM) regression methods. The results obtained reached high accuracy, with R^2 close to 0.90 and with bias tending to zero. Among the tested machine learning methods, RF was slightly better than the other methods and its results were similar to the results obtained with multiple linear regression. The assessment of ANN, k nearest neighbours, and RF for modelling trunk shape and volume in black wattle plantation, Schikowski et al. (2018) showed that ANN and RF presented the best results, with RMSE of 8% and 8.4%, respectively, against the RMSE of 9.15% for the polynomial model. The authors concluded that the machine learning techniques

are appropriate for forest modelling, however, their use should be cautious because of the greater possibility of overtraining and overfitting.

Regarding native Brazilian forests, the combination of LiDAR metrics and machine learning techniques is mainly focused on the Amazonian Forest to estimate the aboveground biomass. In low-intensity logging areas, Rex et al. (2020) estimated the aboveground biomass (AGB) stock by comparing multiple linear regression with some machine learning approaches. Linear regression was the most appropriate method for the case study, with an RMSE of 19.7%, slightly better than the methods of RF, ANN, and SVM with a RMSE of 22.8% for the three methods. However, the results demonstrate the potential for predicting AGB when a non-parametric method is required mainly in tropical forests, due to its great diversity and heterogeneity.

Nevertheless, there are other biomes in Brazil with high richness and species diversity such as the Atlantic Forest. This is the second-largest rainforest in America, which occurs mainly along the coast, extending far inland in some areas of south and southeastern Brazil Fiaschi; Pirani, (2009), whose composition, structure, and diversity remain mostly unknown. Due to the occupation of the national territory that mainly occurs along the coast and other anthropogenic activities (e.g., logging, disordered urban growth, agricultural encroachment, and industrialization), the Atlantic Forest is the most degraded biome in Brazil. It is approximated that only 11.6% of its original cover still remains, and these are very fragmented (HADDAD et al., 2015; HARGREAVES, 2008; RIBEIRO et al., 2009). However, this biome is a biodiversity hotspot because it has already lost more than 75% of its original cover, with very high fragmentation, the remaining forest fragments of this biome have a high species endemism (MYERS et al., 2000; WILLIAMS et al., 2011).

The semideciduous seasonal forest, also known as inland Atlantic Forest because of the inland location, is one of the phytophysiognomies and associated ecosystems defining and forming the Atlantic forest, as described by IBGE, (2012). Despite the importance of this native forest, it is often neglected, resulting in a lack of information about its composition, structure, and diversity (MMA; IBAMA; ICMBIO, 2007).

The lack of studies using LiDAR metrics in the Brazilian Atlantic Forest and their potential for estimating variables describing the forest structure was the motivation for this work. The main objective is to compare four machine learning approaches (ANN, ordinary least-squares multiple regression (OLS), RF, and SVM) with different numbers of LiDAR metrics as input data, to estimate seven stand and diversity variables: mean diameter at breast

height (MDBH), quadratic mean diameter (QMD), basal area (BA), density (DEN), number of tree species (NTS), Shannon–Waver diversity index (H'), and Simpson diversity index (D). An area-based approach was considered more applicable than individual tree-based techniques due to the difficulty of extracting individual trees in the tropical forest (HYYPPÄ et al., 2012; YU et al., 2011).

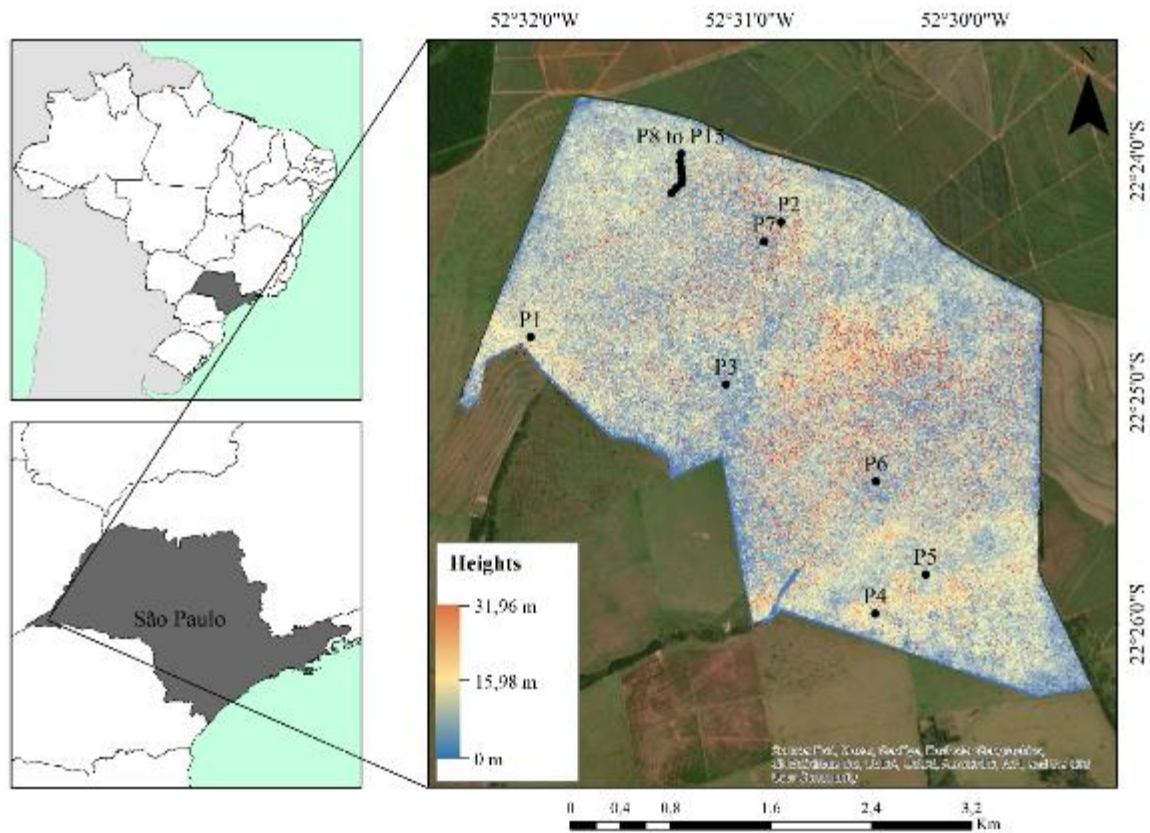
An extensive experimental assessment was made, combining the three input types of LiDAR metrics with the different regression methods tested with different architectures, to estimate the stand and diversity forest variables cited. The results obtained in this study may contribute to finding the best combination of a selection of metrics to deal with and a machine learning technique to estimate forest variables in an inland Atlantic Forest.

2 MATERIALS AND METHODS

2.1 FIELD SURVEY

The study was carried out in a remnant of inland Atlantic Forest, named *Ponte Branca* (White Bridge), located in the west of São Paulo State, Brazil. Its size is approximately 13 km² (Figure 1). As described by Berveglieri et al. (2016), this forest remnant has suffered from disturbances over time like selective logging and forest fires. However, there are still areas at a good conservation state. Different successional stages are found from pioneer formations to advanced regeneration stages with late secondary and climax species in the upper canopy. The understory is dominated by the Myrtaceae family, mainly by the *Eugenia uniflora* species, and with a high occurrence of *Dendropanax cuneatus*. In the upper canopy the species *Aspidosperma* spp., *Copaifera langsdorffii*, and *Hymenaea courbaril* are found.

Figure 1 – Location of the study area and the canopy height model representing the tree heights inside the *Ponte Branca* Forest remnant.



Field data were collected in seven square plots of 40 m x 40 m and eight rectangular plots with dimensions of 80 m x 20 m. This sums up to a total of 15 plots with an area of 1600 m² each (Figure 1). Plots 8 to Plot 15 have different dimensions due to greater observed heterogeneity in successional stages found in each plot. As a result, rectangular plots better represent the differences in these areas. The allocation of each plot was based on the previous interpretation of historical and recent aerial images and on the management plan provided by the environmental agency responsible for the study area, in such a way that the plots covered all successional stages present in the *Ponte Branca* Forest remnant. In Berveglieri et al. (2016, 2018) a detailed description of the successional stages in the study area is presented. The sampling area was 0.2% of the total forested area. The four corner positions of each plot were obtained by a dual-frequency Global Navigation Satellite System receiver with, at least, one-hour tracking, achieving an estimated precision of approximately 50 cm. The area did not have continuous inventory data due to the restricted accessibility caused by the high density of trees, and vines within the forest remnant. Despite the accessibility limitations, it was possible to survey a sample of plots for this study.

All trees with DBH (Diameter at Breast Height) above 3.5 cm were counted, individually measured, and their tree species were identified. From the measured DBH the variables MDBH (cm), QMD (cm), and BA (m² ha⁻¹) were calculated. From the tree counting, DEN (trees ha⁻¹) was obtained, and from the catalogued species, the total number of tree species per plot (NTS) was calculated. The Shannon–Weaver diversity index, H' (Equation (1)) (MAGURRAN, 1988), and the Simpson index D (Equation (2)) (BROWER; ZAR; VON ENDE, 1984), were also calculated to understand species composition and diversity. A summary of the seven variables obtained in the field is shown in Table 1, and the statistics of the variables for each of the 15 surveyed plots are presented in the Appendix A (Table A1).

$$H' = -\sum_{i=1}^S p_i \ln p_i \quad (1)$$

where H' is the Shannon–Weaver diversity index, S is the total number of species sampled and p_i is the ratio of the number of individual trees sampled from the i th species to the total number of individual trees.

$$D = \frac{\sum n_i(n_i-1)}{N(N-1)} \quad (2)$$

where D is the Simpson index, n is the number of individual trees of the i th species and N is the total number of individual trees.

Table 1 – Statistics about forest variables calculated from field data.

Field Variables	Minimum	Maximum	Mean	Standard Deviation	Coefficient of variation (%) ¹
MDBH	8.5	13.9	10.6	1.4	13.4
QMD	9.8	18.8	13.2	2.3	17.4
BA	5.6	30.7	16.2	7.5	46.2
DEN	380	2286	1193	569.4	47.7
NTS	10	24	15	3.8	25.9
H'	1.18	2.04	1.56	0.21	13.6
D	0.49	0.79	0.67	0.08	11.9

¹ Coefficient of variation (%) calculated by the ratio of the standard deviation to the average multiplied by 100. Field variables: mean diameter at breast height (MDBH), quadratic mean diameter (QMD), basal area (BA), density (DEN), number of tree species (NTS), Shannon Weaver diversity index (H'), and Simpson diversity index (D).

2.2 LiDAR DATA COLLECTION

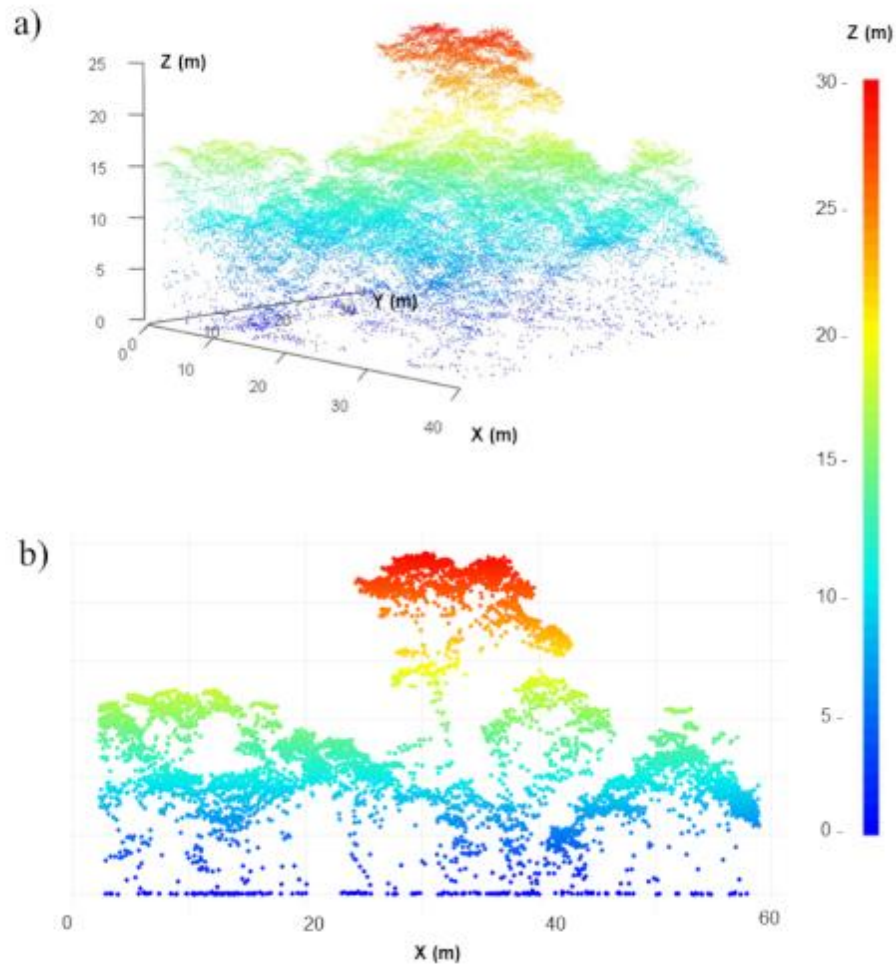
The LiDAR data covering the study area was acquired in October 2017. The airborne laser scanner used was a RIEGL LMS-Q680i, which is a full-waveform LiDAR sensor with a scan angle range of $\pm 30^\circ$. The waveforms were processed in post-processing mode and the point cloud was the peak returns of the waveforms that were delivered in discrete LiDAR file formats. Up to 5 returns per emitted pulse were recorded, which is allowed according to the specifications of the *.las* file (ASPRS, 2010). This scanner model uses the multiple time around (MTA) technique due to the high frequency of repetition (up to 400,000 Hz), being able to acquire echoes arriving after a delay of more than one pulse repetition interval, thus allowing measurements with a range beyond the maximum unambiguous measurement range (RIEGL, 2012).

The flight height of the LiDAR survey used in the study was 900 m. The LiDAR data was delivered in 16 flight lines acquired in the northeast-southwest direction and covering the entire *Ponte Branca* Forest remnant. The average sampling density was 19.8 pulses per m². Figure 2 depicts the ALS point cloud of a plot of the study area, showing a sample of the vertical structure of the forest.

2.3 LiDAR DATA PROCESSING

For LiDAR data processing, *LAStools* (ISENBURG, 2014) and R environment (R CORE TEAM, 2017) have been utilized, to obtain area-based metrics from the point clouds. First, with *LAStools* (ISENBURG, 2014), the classification of the point cloud into ground and nonground points was performed. Several parameters were empirically assessed to achieve suitable results, mainly ensuring that ground points always existed within the selected search window, since in tropical forests there is a dense understory, and ground points are sometimes not acquired by LiDAR systems.

Figure 2 – Airborne laser scanner (ALS) point cloud representing the vertical structure of a plot of the *Ponte Branca* Forest remnant. (a) Three-dimensional view of the plot. (b) Cross section of the same plot.



The following processing steps were performed in the R environment (R CORE TEAM, 2017) using the *LidR* package (ROUSSEL et al., 2018). The main aim was to extract LiDAR metrics, not only related to elevation and pulse returns, but also to intensity metrics, to understand their relationship in estimating forest variables. However, the distance between the sensor and the target is not constant during the survey, due to the variations of the platform (aircraft), the terrain topography (AHOKAS et al., 2006; HOPKINSON, 2007; KAASALAINEN et al., 2007; ROUSSEL; BOURDON; ACHIM, 2020), and the scan angle. These variations in the distances and atmospheric attenuation change the intensity recorded by the sensor (GATZIOLIS, 2011; HOPKINSON, 2007; LUO et al., 2017). Even with these variations, some authors used raw intensities (SHI et al., 2018; SUMNALL; HILL; HINSLEY, 2016; ZHANG; LIU, 2013). However, to achieve more accurate results, the use of intensity

metrics requires a priori correction also known as normalization. In this study, we used a range correction model (Equation (3)) developed by (GATZLIOLIS, 2011; HOPKINSON, 2007), which is based on the principle that the distance travelled from the sensor beam to the target is not the same (ROUSSEL; BOURDON; ACHIM, 2020).

$$I_{norm} = I_{obs} \left(\frac{R_{act}}{R_{ref}} \right)^f \quad (3)$$

where I_{norm} is the normalized intensity, I_{obs} is the observed intensity, R_{act} is the distance between the laser instrument and the returns, R_{ref} is an arbitrary reference distance, f represents the rate of energy attenuation sustained by the pulse as it travels through a medium back and forth from a target.

To calculate the R_{act} parameter for each return, it was necessary to determine the location of the sensor, which was performed using a method proposed by Roussel; Bourdon; Achim. (2020), which is an adaptation of the methodology developed by Gatziolis, (2011). The main idea is that a pair of very close pulses on different flight lines, assuming they reach the same object, must have the same intensity. If there is a difference it would be indicative of a difference of range (ROUSSEL; BOURDON; ACHIM, 2020). Thus, the sensor's position was determined by linear interpolation from the two closest points in the aircraft's trajectory positions, then the range for each return was determined as the Euclidean distance from the point to the sensor (GATZLIOLIS, 2011; ROUSSEL; BOURDON; ACHIM, 2020). To apply this method, the values of 'gpstime', 'ReturnNumber', 'NumberOfReturns', and 'PointSourceID' are necessary (ROUSSEL et al., 2018), which are not always made available by the data provider, making normalization of intensity unviable (GATZLIOLIS, 2011; HOPKINSON, 2007; KAASALAINEN et al., 2007; ROUSSEL; BOURDON; ACHIM, 2020). Our data had all these values, so it was possible to proceed with the intensity normalization.

We use the arbitrary reference distance as the flight height, that is, 900 m. In forest areas, the optimum value of the exponent f should be between 2.1–2.5 (GATZLIOLIS, 2011), but for practical purposes, the values between 2.2 to 2.4 are the most suitable, according to (ROUSSEL; BOURDON; ACHIM, 2020). We made some preliminary experiments and the value of 2.3 was the best choice for our area of study. After this procedure, the result was a point cloud with corrected (or normalized) intensity values, with which it was possible to extract intensity

metrics and to use them as predictor variables. The Riegl LMS-Q680i instrument records pulses with an intensity of 16 bits, but processing and analyzing huge datasets with this range is costly and thus, the intensity data were resampled to 8 bits.

A point cloud with corrected intensities, containing only ground points was modelled using a triangular irregular network (TIN) to generate a digital terrain model (DTM), with a sample distance of 0.50 m. The point cloud (once ground points were removed) was normalized using the DTM. The result is a normalized point cloud with only vegetation related points included, mapped on flat terrain. After that, outliers were removed according to the expected heights of the trees existing in the area; points below 0 m and above 40 m were removed. The normalized point cloud filtered for outliers was interpolated using the *p2r* algorithm, with a sub circle radius of 0.1 m, and the spatial interpolation to fill the empty pixels used a TIN. The *p2r* algorithm is based on the “*points-to-raster*” method (KHOSRAVIPOUR et al., 2014) for each pixel of the output raster and the height of the highest point is assigned, resulting in a canopy height model (CHM with a sampling distance of 0.50 m) (Figure 1). This method was chosen due to the processing speed and good smoothing of the canopy model.

The normalized point cloud was then clipped, based on vector files defining the perimeter of the plots surveyed in the field, from which LiDAR metrics at each plot level were extracted. Overall, 54 metrics were extracted using elevation, intensity, and pulse return values (Table 2) from each point cloud of each plot, which served as input data for the regression models.

2.4 INPUT DATA SELECTION

Different numbers and combinations of input data for the different machine learning models were tested. Thus, from the 54 LiDAR metrics previously extracted, two further analyses were performed to reduce the number of observations as input data.

The first analysis was a correlation calculation using the Pearson (*r*) method (BENESTY et al., 2009). The LiDAR metrics having a correlation coefficient greater than 0.70 and smaller than -0.70 were selected. If two or more metrics had a high correlation (direct or inverse), only one of these metrics was maintained and the others were excluded as explained by Hudak et al. (2012); Silva et al. (2014). We aimed to keep the number of uncorrelated metrics equal to the number of samples. For that reason, from the 54 extracted LiDAR metrics, 15 metrics were maintained after the aforementioned analysis.

The second analysis was the PCA (principal component analysis), which is a technique that makes a linear transformation of highly correlated variables generating a new set of uncorrelated orthogonal variables, called principal components (PCs) (ABDI; WILLIAMS, 2010; MANLY; ALBERTO, 2016). In this case, the 54 LiDAR metrics were transformed into a smaller set of uncorrelated metrics with the most ones (MANLY; ALBERTO, 2016). The PCAs were extracted by using the *FactoMineR* R package (LÊ; JOSSE; HUSSON, 2008). Due to the different scales of the LiDAR metrics, a preliminary step was inserted for normalizing the inputs to zero mean and unit variance. Five dimensions were retained in the results. The selection of the main components was based on the Kaiser criterion (KAISER, 1958), in which the components with eigenvalues greater than one should explain most of the variations of the LiDAR metrics (CLIFF, 1988; HONGYU; SANDANIELO; DE OLIVEIRA JUNIOR, 2016).

Having performed the previous analysis, 3 different sets of input data were available: 54 metrics extracted from LiDAR data, 15 non-correlated metrics, and 5 PCs which were used as new input variables in the regression and machine learning techniques.

Table 2 – Light detection and ranging (LiDAR) metrics extracted from normalized point clouds.

Metrics	Description
ZMAX	Maximum height
ZMEAN	Mean height
ZSD	Standard deviation of height distribution
ZSKEW	Skewness of height distribution
ZKURT	Kurtosis of height distribution
ZENTROPY	Entropy of height distribution
PZABOVEZMEAN	Percentage of returns above ZMEAN
PZABOVE2	Percentage of returns above 2 m
ZQ _x	Xth percentile (5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95) of height distribution
ZPCUM _x	Cumulative percentage of return in the Xth layer (1 to 9) with f(z) the probability distribution of elevations
ITOT	Sum of intensities for each return
IMAX	Maximum intensity
IMEAN	Mean intensity
ISD	Standard deviation of intensity
ISKEW	Skewness of intensity distribution
IKURT	Kurtosis of intensity distribution
IPGROUND	Percentage of intensity returned by points classified as ground
IPCUMZQ _x	Percentage of intensity returned below the Xth (10, 30, 50, 70, 90) percentile of height
PX _{th}	Percentage of Xth returns (1 to 5)
PGROUND	Percentage of returns classified as ground

2.5 REGRESSION TECHNIQUES SETTINGS

In this study, the forest stand and diversity variables were derived by machine learning techniques tested in the R environment (R CORE TEAM, 2017) and the associated packages for each method. We tested the OLS and three methods of regression by supervised machine learning: ANN, RF, and SVM, whose inferences are based on inductive reasoning, in which the process of approximation of functions is performed by the knowledge acquired (CARVALHO et al., 2011).

The OLS is a traditional approach to data modelling. It has advantages in terms of simplicity and ease-making inferences with good predictive performance (FREESE, 1964). OLS fitting was performed using the *Carret* package (KUHN, 2008). Considering that it is not possible to apply OLS with more predictors than samples, the test case with the 54 LiDAR metrics as input was not tested with this method.

Using as inputs the 5 PCs (OLS–5) and the 15 uncorrelated LiDAR metrics (OLS–15), a selection of the predictor variables was performed for each of the seven forest variables to be estimated. This selection was implemented in R with the *Leaps* package using the *regsubsets* function (LUMLEY; LUMLEY, 2013), as described by Shin et al. (2016). This function works in a similar way to the stepwise method, and we set the maximum number of predictor variables to not exceed a ratio of 1:3, compared to the number of samples, that is, only a maximum of 5 predictor variables could be selected in each set of input data for each forest variable to be estimated. Subsequently, from the set of 5 predictor variables, we selected those statistically significant, according to the p -value at an $\alpha = 0.05$, for a given variable. Then, after selecting the predictor variables, the OLS model was fitted.

ANNs are computational models inspired by the human nervous system, that acquire knowledge through a learning process, with synaptic weights that indicate the strength of the connections between neurons (HAYKIN, 2010). Multilayer perceptrons are a type of network having a universal approach for any continuous function and are typically defined by the input layer, at least a hidden layer made up of neurons, and an output layer (HAYKIN, 2010; HORNIK; STINCHCOMBE; WHITE, 1989). Five different ANNs architectures using the *neuralnet* package (GÜNTHER; FRITSCH, 2010) were tested (Table 3). The input metrics data were first normalized to belong to the [0,1] range. The backpropagation algorithm was used, with a learning rate of 0.01. The activation function was logistical, and the type of network was multilayer perceptron, in which one, two, and three hidden layers were tested. For one hidden

layer, one architecture was tested, and for two and three hidden layers, two architectures were tested.

Table 3 – Summary of the architectures adopted for modelling using artificial neural networks (ANNs).

Number of Hidden Layers	Inputs	Architecture*	Name
1	5 PCs	5–3–1	ANN–5–1
	15 Metrics	15–4–1	ANN–15–1
	54 Metrics	54–8–1	ANN–54–1
2A	5 PCs	5–16–8–1	ANN–5–2A
	15 Metrics	15–16–8–1	ANN–15–2A
	54 Metrics	54–16–8–1	ANN–54–2A
2B	5 PCs	5–6–3–1	ANN–5–2B
	54 Metrics	54–55–28–1	ANN–54–2B
3A	5 PCs	5–16–8–4–1	ANN–5–3A
	15 Metrics	15–16–8–4–1	ANN–15–3A
	54 Metrics	54–16–8–4–1	ANN–54–3A
3B	5 PCs	5–6–3–1–1	ANN–5–3B
	54 Metrics	54–55–28–14–1	ANN–54–3B

* The first number refers to the number of input data and the last number to the output data. The intermediate numbers are the number of neurons in each hidden layer.

In the case of one hidden layer, the method proposed by Shibata; Ikeda. (2009) was adopted. In this method, the input layer with the number of inputs was equal to the number of metrics (15 or 54) or PCs (5) adopted. The number of neurons was defined by the square root of the number of variables in the input layer times the number of outputs; in this case, one for each of the forest variables studied. The output layer contained the estimations of the forest variables.

For the ANNs with two and three hidden layers, configurations named “A” and “B” were assessed. The configurations “A” sets the number of LiDAR metrics or the PCs in the input layer. In the first hidden layer, the number of neurons was defined as the number of surveyed plots (15) plus the number of outputs (case one). In the second hidden layer, the number of neurons was half of the number of neurons of the previous layer. If the architecture had three hidden layers, the number of neurons in the third layer was half of the number of neurons in the second layer. The configuration “B” sets the number of neurons in the first hidden layer as the number of inputs (LiDAR metrics or PCs) plus the number of outputs (as we were doing regression, the number of outputs was one). The second hidden layer had half the number of the first. Having a third hidden layer, the number of neurons should be half of the second. It is worth noting that for the case with 15 metrics as input, only one configuration was tested

with two and three hidden layers since the number of metrics and surveyed plots was the same and there was no difference between configurations A and B.

The RF algorithm is based on decision trees. The rationale of this technique is to grow a set of decision trees (referred to it as “forest”) such that the correlation between these trees remains as low as possible. Randomness is placed in the forest using a different subset of training samples for each tree, so instances of the training set are randomly extracted, aiming to train a specific number of trees in the “forest”. In addition, for each tree node, a random subset of the input variables is used to learn the partition function, making the decision trees as independent as possible, improving the robustness and generalization of the data set (BERK, 2008; BREIMAN, 2001). For the estimation of variables using the RF method, the *randomForest* package (LIAW; WIENER, 2002) was used. The number of decision trees constructed was 1000, and in each node of the tree, the number of predictor variables randomly sampled was a third of the number of inputs. The nomenclature adopted for the inputs of this regression method using the PCs, 15 and 54 LiDAR metrics were RF-5, RF-15, and RF-54, respectively.

SVM is an algorithm that uses the concept of “margins”, which is the shortest distance between the decision surface and any of the samples (CORTES; VAPNIK, 1995). The main idea of this technique is to fit optimal decision surfaces (called hyperplanes) to a set of training samples, for deriving the linear dependency between unidimensional target variables and n-dimensional input vector pairs (BEN-HUR; WESTON, 2010; CORTES; VAPNIK, 1995; HUANG; DAVIS; TOWNSHEND, 2002). SVM with epsilon regression type (ϵ -SVM) was used, with two necessary packages of R environment: *kernelab* (KARATZOGLOU et al., 2004) and *e1071* (MEYER et al., 2012). The fitting with the ϵ -SVM method was performed for three sets of inputs: 54 LiDAR metrics, 15 uncorrelated metrics, and the 5 PCs. Three types of *kernels* were also tested for each of the three inputs: linear, polynomial, and radial. The value adopted for the *cost* was 1 (one), the *gamma* parameter was defined as the inverse of the number of samples (1/15), and the *epsilon* parameter was 0.1. A summary of each parameter and input data used for fitting with the ϵ -SVM approach is in Table 4.

Table 4 – Summary of the parameters adopted for modelling using epsilon regression type-support vector machine (ϵ -SVM).

Kernel Type	Inputs	Name
Linear	5 PCs	SVM-5-L
	15 Metrics	SVM-15-L
	54 Metrics	SVM-54-L
Polynomial	5 PCs	SVM-5-P
	15 Metrics	SVM-15-P
	54 Metrics	SVM-54-P
Radial	5 PCs	SVM-5-R
	15 Metrics	SVM-15-R
	54 Metrics	SVM-54-R

2.6 EVALUATION AND PERFORMANCE OF TESTED MODELS

The bootstrap resampling process was used for the estimation, with the performance of 1000 random bootstraps, with a set of data from this resampling being drawn in each analysis, consisting of all seven response variables, as well as the different input data for each technique of machine learning tested in this study (LATIFI; FASSNACHT; KOCH, 2012), as described in Section 2.5. The performance of all modelling techniques (OLS, ANN, RF, SVM) with all different settings was assessed using leave-one-out cross-validation (LOOCV). A total of 14 reference samples were used to train the model and the remaining one was used for calculating the prediction error; this was repeated for each reference sample for cross-validation. The following statistics were calculated: root-mean-square error (RMSE) (Equation (4)) and bias (Equation (5)), both in percentage values.

$$RMSE\% = \frac{\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}}{\frac{\sum y_i}{n}} \cdot 100 \quad (4)$$

$$BIAS\% = \frac{\sum(\hat{y}_i - y_i)}{\sum y_i} \cdot 100 \quad (5)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value and n is the number of observations.

Instead, the selection of the best regression technique to predict each of the seven forest variables was based on RMSE (%), bias, and the Akaike information criterion (AIC) (AKAIKE, 1974), which is a measure of the quality of an adjusted model, using the maximum likelihood

method (Equation (6)). According to Bozdogan, (1987), complex models with many variables tend to be penalized through the AIC, advantaging simpler models. A good model is one that has minimum AIC among all the other models (MOHAMMED; NAUGLER; FAR, 2015). Due to the small number of samples, we used an AIC correction (Equation (7)), developed by Hurvich; Tsai, (1989), since there is a tendency for AIC to select models with many parameters in the case of small samples. AICc adds an extra term penalty to the number of parameters (HURVICH; TSAI, 1989; MCQUARRIE; TSAI, 1998).

$$AIC = -2 \ln(L) + 2k \quad (6)$$

where L is the maximum likelihood of estimated parameters and k is the number of parameters in the model.

$$AICc = AIC + 2 \frac{k(k+1)}{n-k-1} \quad (7)$$

where k is the number of parameters in the model and n is the number of samples.

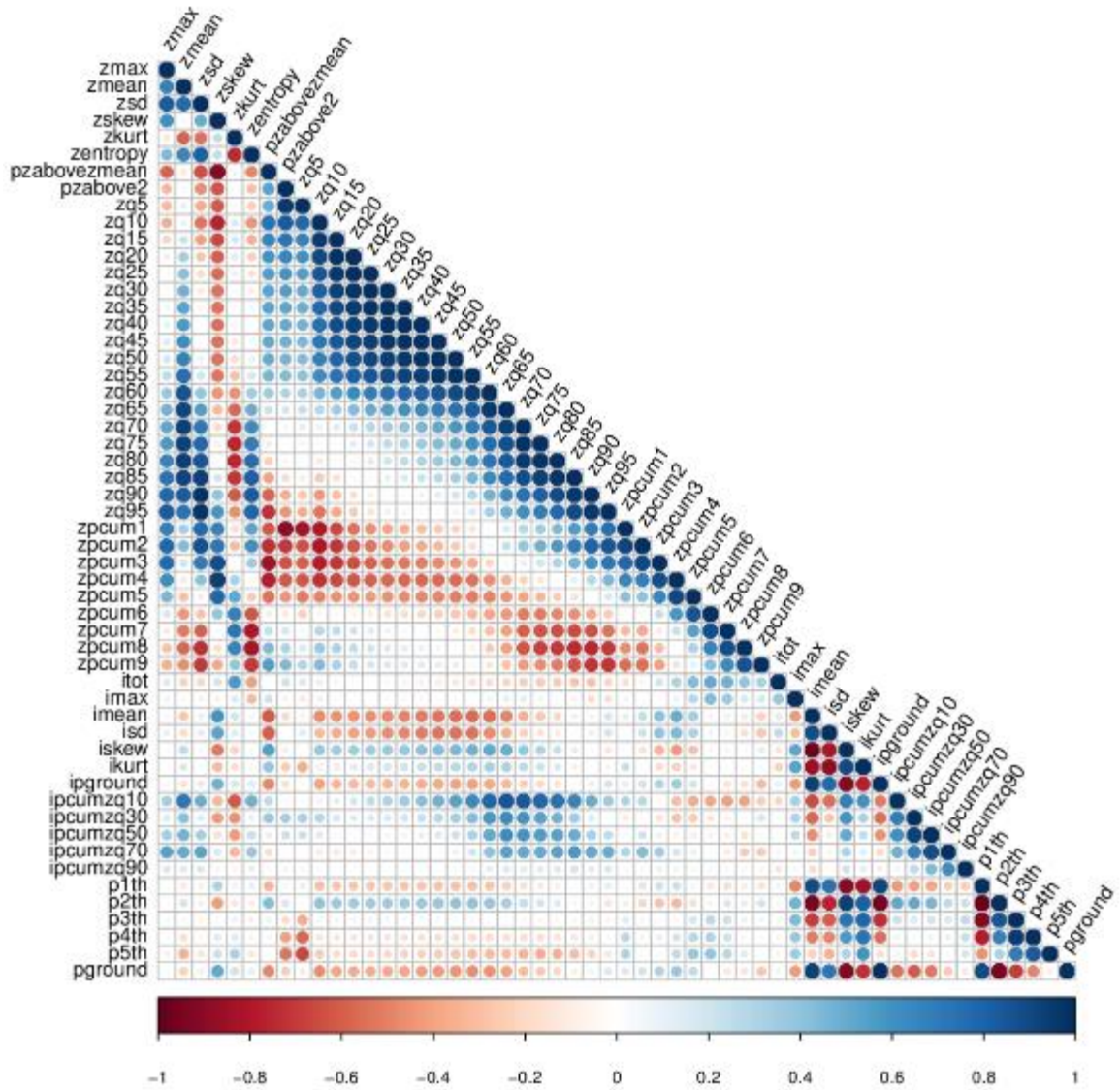
For each given variable, the best model presented a lower RMSE (%) value, a bias close to zero, and had the lowest AICc value among the models tested and with the various input configurations. The contribution of each input data (LiDAR metric or PC) to the best model selected, when estimating each forest variable, was shown for a better understanding of the physical meaning of these inputs in the response variables.

3 RESULTS

3.1 CORRELATION ANALYSIS AND PCAs

Some LiDAR metrics were highly correlated with each other, as shown in Figure 3, where the stronger the blue shade is, the greater the positive correlation is, and the stronger the red shade is, the greater the inverse correlation is. After correlation analysis, the 15 resulting metrics were: ZMAX, ZSKEW, ZQ5, ZQ30, ZQ75, ZPCUM5, ZPCUM7, ZPCUM9, PZABOVE2, P2TH, P5TH, PGROUND, IMAX, IPZCUMZQ50, IPZCUMZQ90. The meaning of these metrics is described in Table 2.

Figure 3 – Correlogram between the 54 extracted LiDAR metrics.



The first five PCs explained 90% of the variability of LiDAR metrics, according to the Kaiser criterion. Figure 4a shows the contribution of each PC in the total variance. Some authors, such as Rex et al. (2020) and Silva et al. (2020), have used the PCA to select variables. They selected one single metric from each PC, based on the highest eigenvector value. However, selecting a single metric may reduce the accuracy, since relevant information from the other metrics is missed.

Figure 4 – (a) First five principal components (PCs) and the percentage of variance explained by each one. (b) Projection of Table 5. PCs and the LiDAR metrics.

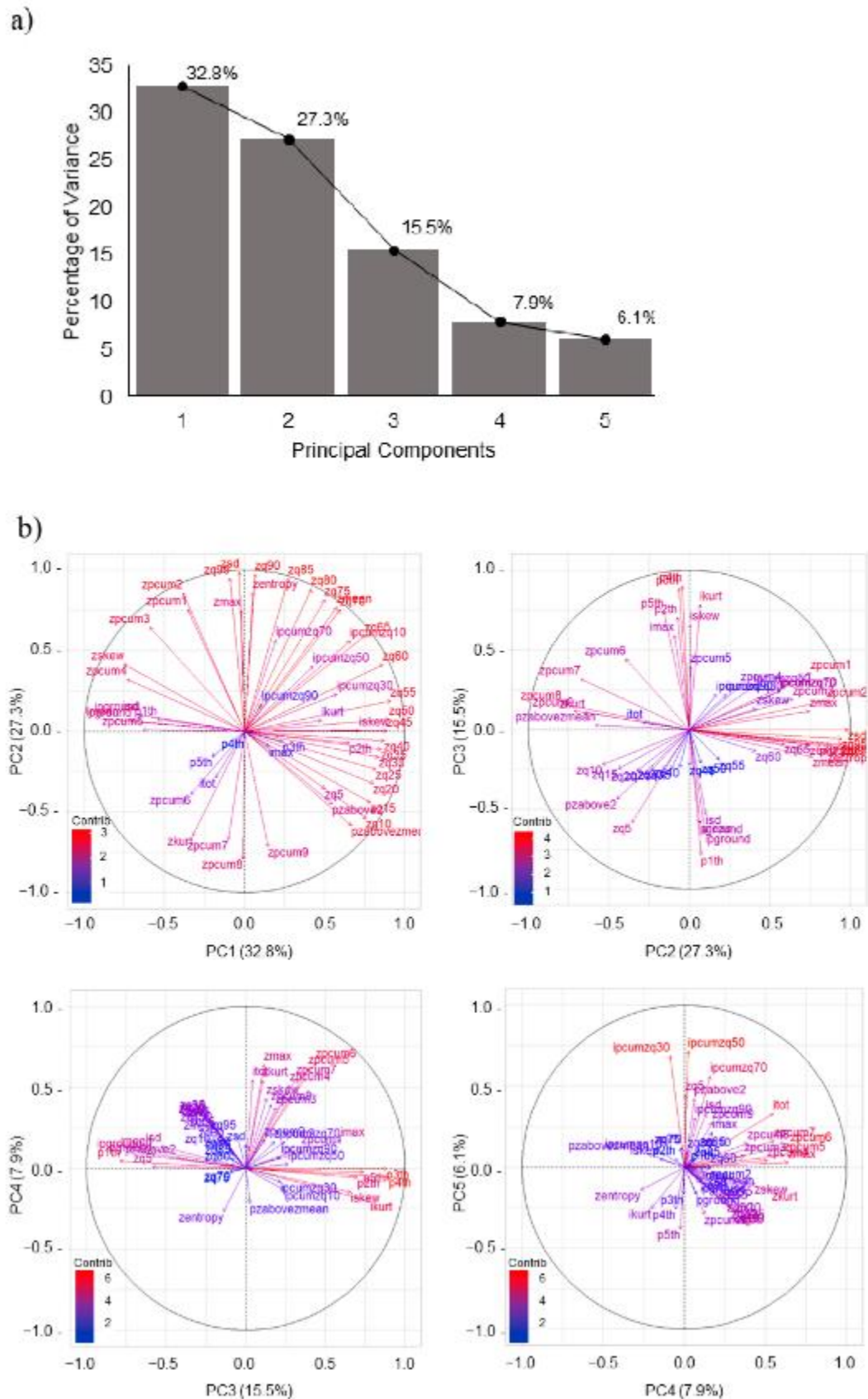


Table 5 – Selection of the best model for each estimated forest variable.

Variables	RMSE (%)	Bias (%)	AICc	Best Model Fitted
MDBH	5.6	0.60	15.03	ANN-5-2B
QMD	5.2	−0.03	14.44	ANN-5-2B
BA	22.5	−0.24	9.33	ANN-5-2B
DEN	16.3	−12.31	−6.75	ANN-5-2B
NTS	27.6	−12.49	4.90	ANN-5-3B
H'	10	−1.75	20.11	ANN-5-3B
D	8.4	3.64	24.55	ANN-5-3B

In this study, we decided to use the new variables created from the PCA as input, since they hold information about the importance of each metric in each of the 5 PCs. For further understanding of the physical meaning of each PC, the projection of the LiDAR metrics is shown for the 5 PCs that explain most of the data variance (Figure 4b).

3.2 MODEL PERFORMANCE AND EVALUATION

Considering all the machine learning techniques and test cases (in respect to input data and architecture), a total of 27 combinations were examined for estimating the seven forest variables used. Figure 5a shows the RMSE (%) of all the models (OLS, ANN, RF, and SVM) with the different settings for each one of the variables. The results of the adjusted linear models (OLS) with the significant input data, based on the p-value are shown in Table A2 in the Appendix A.

The MDBH variable, represented by the dark green, was estimated with the lowest RMSE, with an average RMSE of 8.05%, followed by QMD (orange bars) and D (dark blue bars), with RMSE average values of 9.3% and 10.9%, respectively. The RMSEs were the highest for NTS (light blue bars), whose average was 47.6%. As seen in Table 1, the variables with the lowest RMSE were the ones with the lowest CV%, below 20%, which is the range of values that do not have high variability between the surveyed plots.

Analysing the methods for variables estimation, the lowest RMSE values were achieved with ANNs, specifically with the PCs as inputs, whose RMSE bars (Figure 5a) were lower when compared to other methods, including the lowest RMSE value for the NTS variable. Further, the RF-5 and SVM-5 methods resulted in low RMSE (%), except for the NTS variable, but with a high bias when compared with the ANNs (Figure 5b). On the other hand, the highest RMSE value occurred with SVM-L, with 54 LiDAR metrics as inputs, with an average value for the method of 104%. With this method, the highest values of RMSE (%) were also found for all

seven modelled variables. The highest values of RMSE (%) were also found for all seven modelled variables, including values above 100%, found for the variables BA (139.1%), DEN (272.7%), and NTS (138.1%) which were not included in the chart because of the scale.

Figure 5 – (a) RMSE in percentage for the regression methods tested for the seven estimated forest variables; (b) Bias in the percentage of each modelled variable for each regression method tested.

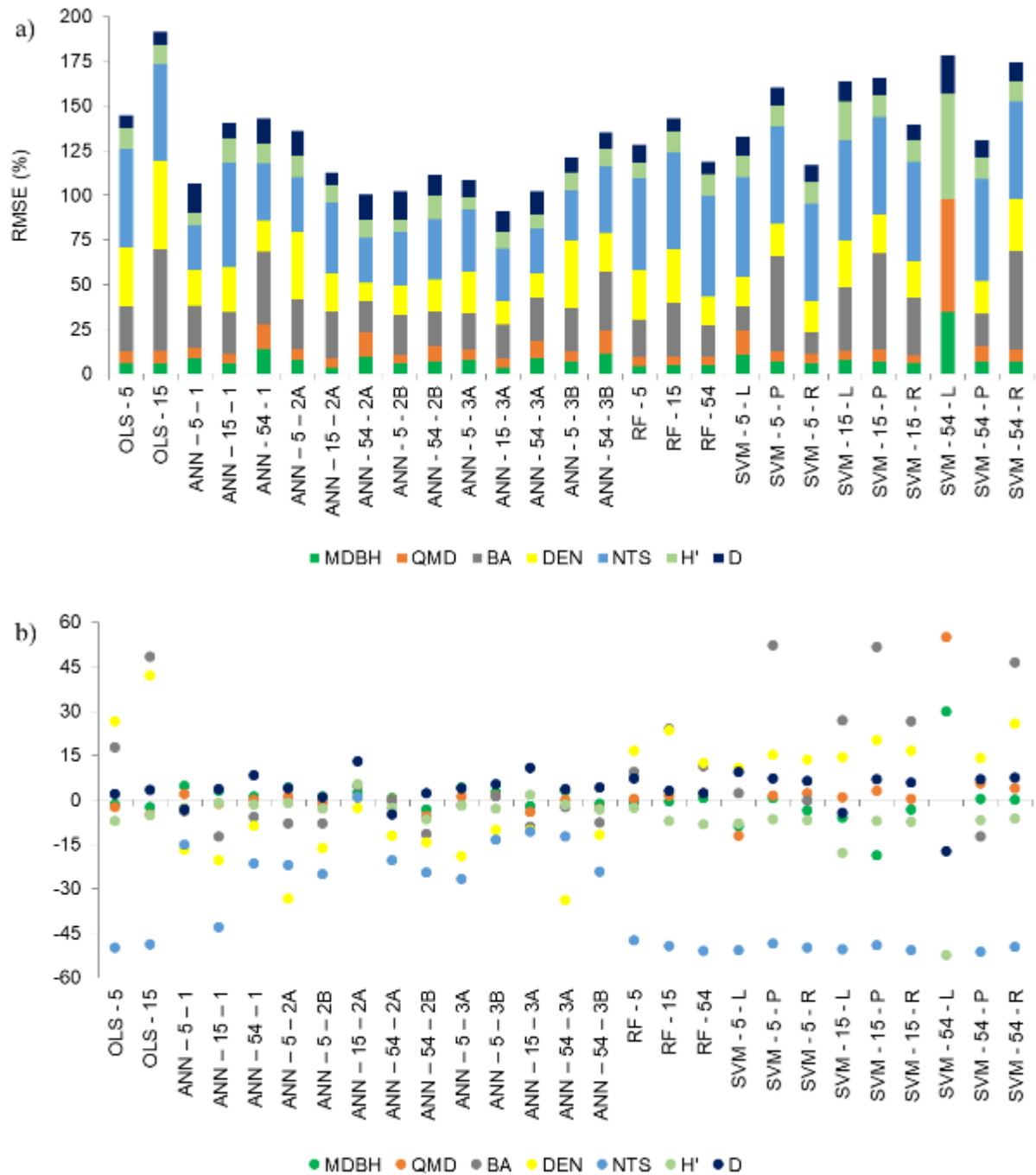


Figure 5b shows the bias for the 27 methods tested on the seven forest variables. These seven variables were estimated with high bias (%), and the highest among all tested approaches, was using the SVM-54-L method. The MDBH variable showed a $\pm 9\%$ bias except for the SVM-method 15-P, which was the second most biased method, with a value of -18.9%. The QMD variable presented the second largest bias for the SVM-5-L method, with a value of -12.17%, while the other values were between $\pm 6\%$. The Shannon-Weaver (H') and Simpson (D) indexes showed a bias of $\pm 8.5\%$, except for the SVM-15-L with a value of -18.1% for the variable H' and the ANN-54-2A with a bias of 12.9% for variable D. These results indicate that these variables (MDBH, QMD, H and D') were well estimated with a favourable errors distribution.

The estimation of the MDBH variable showed an underestimated trend in 13 regression methods and the QMD was underestimated in 11 of the 27 tested methods. The BA variable was underestimated in 14 of the 27 tested methods, with the most discrepant value of -199.6% occurring with the SVM-54-L method. An overestimation of about 50% was also observed for the OLS-15; SVM-5-P; SVM-15-P, and SVM-54-R methods. DEN was also underestimated in 13 methods. SVM-54-L method presented a bias of -237.3% and the OLS-15 method presented the highest overestimation with a bias of 42%. The NTS was underestimated in all tested models, except with the ANN-54-2A method. A bias lower than -45% in 15 of the 27 tested methods were found, with the most discrepant value of -123.5% with the SVM-54-L method. The bias values of the variables BA, DEN, and NTS, estimated with the SVM-54-L were not presented in Figure 5b due to the differences in scale. The underestimation of the values was also noted for variable H' in 25 trials and the variable D was overestimated in 23 regression trials.

To sum up, we used the AICc as a numerical value to drive the choice of the best regression method when predicting forest variables (Table 5). Thus, the best compromise model is the one with the lowest error (RMSE), low bias, and the minimum possible variables to be estimated, while at the same time, it best explains the behaviour of the response variable (BOZDOGAN, 1987).

According to the adopted criteria, as shown in Table 5, ANN using the five PCs as input was the regression method that best estimated the forest variables, with differences in the number and the configuration adopted for the hidden layers. The variables MDBH, QMD, BA, and DEN were better estimated using two hidden layers, with the "B" configuration (see Table 3). For the variables NTS, H', and D, the ANN with three hidden layers and "B" configuration,

was the method that best estimated these variables. The variables derived from DBH (QMD and BA) and derived from BA (DEN), have the same ANN configuration as the best modelling method (ANN-5-2B). This was also observed in the case of H' and D , whose calculation involves the number of species, the best method being ANN-5-3B, which was the same for the variable NTS. However, considering the limited number of sample data, the use of complex architectures with two and three layers in ANN, and the fact that the selected metrics are hidden, there is a risk of overfitting. The test case in which the most important inputs were selected before training the ANN presented a reduced risk of using irrelevant input metrics.

3.3 IMPORTANCE OF INPUT METRICS

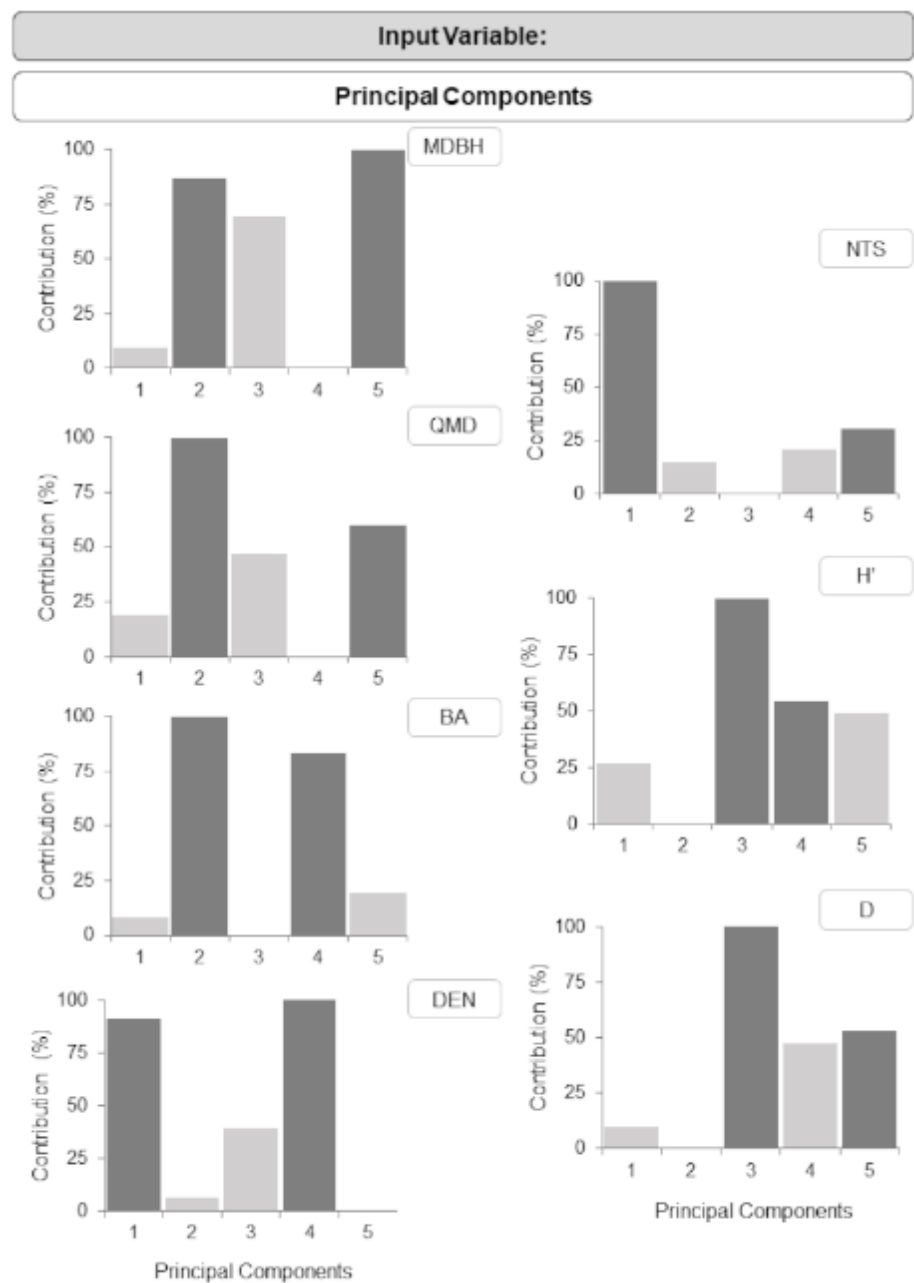
The relative importance of each PC for estimating forest variables for the best-selected regression method is shown in Figure 6. The dark grey bars indicate the two predictor variables with the greatest contribution to the estimation of a given forest variable. For MDBH, the fifth PC was the most important predictor variable, followed by the second PC.

On the other hand, for the QMD variable, the second PC was the most important predictor variable, as well as for BA, and then the fifth PC. For BA, the second predictor variable that most contributed to the estimate was the fourth PC. It is possible to verify some relationships, as in the case of QMD and BA, whose predictor variable with the greatest contribution is the same, since the QMD is obtained by the inverse formula of BA. Other relationships can be observed, e.g., MDBH and QMD, in which the predictor variables that most contribute to the estimate are the same but inverted (since QMD is also derived from MDBH). BA and DEN share the fourth PC in common, as the density considers the area occupied by the tree trunks.

The number of tree species (NTS) has the first PC as the most important predictor variable, followed by the fifth PC. Still, in this case, the contribution of the second predictor variable is less than 50%, different from that observed in the other estimated variables. The Shannon–Weaver (H') and Simpson (D) indexes share the third PC with the greatest relative importance, followed by the fourth PC and the fifth PC, respectively. As with the other estimated variables, we can observe some relationships here as well. In their formulation, both indexes H' and D use the number of species and individuals and, thus, it would be expected this similarity with the most important predictor. The variables D and NTS (Figure 6) have as the

second predictor variable, the fifth PC, indicating a relation of these two variables as mentioned above.

Figure 6 – Relative importance of PCs for each forest variable modelled by the best-selected regression method.



This interpretation was based on the empirical assessment of the results, since LiDAR metrics and their respective transformations, i.e., PCA, may not always have a physical meaning in the estimation of a given variable. Additionally, as shown in Figure 4b, the combination of several LiDAR metrics (i.e., the various metrics of pulse elevation, intensity, and return), when

transformed, provide relevant information on each PC and not a single metric with a higher eigenvalue, and this combination is what gave the results obtained.

4 DISCUSSION

Many combinations of input data have been tested with various regression techniques for estimating variables. This section critically discusses those with the most significant impact on the results, both positive and negative, based on the criteria established for choosing the regression technique.

According to Görgens; Montaghi; Rodriguez, (2015) and Silva et al. (2014), before using LiDAR metrics to build models, a pre-selection and/or transformation of these metrics is necessary to obtain better relations with the variable to be estimated. However, in this process, the selected metrics may not have physical meaning and may differ entirely according to the forest to be studied. This was confirmed in our study, in which the best results were obtained using a previous reduction of dimensionality by the PCA. Some methods, such as RF, also serve to select variables, which could also be tested to verify the consistency of the results obtained in this study (MILTADO et al., 2020).

The use of ANNs for the estimation of forest variables has been growing, and several studies have been developed in Brazilian forests using this technique as an alternative to OLS. Many of these studies are mainly concerned with estimating variables such as height, volume, the shape of the trunk and tapering, and prognosis of yield and production in forests of *Eucalyptus* spp, *Pinus* spp, and *Tectona grandis*, as listed by Chiarello et al. (2019), in Black Wattle plantations (SCHIKOWSKI et al., 2018), in native forests for prediction of the diametric distribution (REIS et al., 2016, 2018) and biomass (YU et al., 2011), both in the Amazon Forest and in the Atlantic Forest biome, aimed at estimating surviving individuals and mortality within the forest (DA ROCHA et al., 2018).

Most of these studies worked with a high sample size and the configuration of the networks for the estimation of the variables had only one hidden layer. This configuration was possibly adopted because the forests, were equian (trees with the same age) and homogeneous, with low variance between individuals, thus requiring the adjustment of less complex networks. However, even in studies with native forests, there is a trend to use only one hidden layer with a variable number of neurons within that layer.

In our study, ANNs with only one hidden layer, with the 54 metrics or 15 uncorrelated LiDAR metrics, presented the highest RMSE% for the studied variables (on average 20%),

compared with the other networks (Figure 5a) and underestimated five of the seven variables (Figure 5b). Silva et al. (2020) estimated the volume in clonal plantations of *Eucalyptus* spp. using several machine learning techniques with LiDAR metrics. They also assessed the different impacts of sample size on estimates, concluding that the sample size influences RMSE (%) and bias (%). The larger the sample size, the lower the values of the percentages are up to a certain level, where the sample size no longer influences them. Among the methods tested, the ANNs were the most susceptible to outliers. Tropical forests are more complex environments than planted forests, due to the great heterogeneity and diversity, requiring architectures with two hidden layers instead of one. This is aligned with Gorgens et al. (2014), who also stated that the processing capacity of a neural network is related to its connectivity; i.e., in more complex jobs, as the demand for hidden layers increases, the number of neurons in the first hidden layer increases as well. However, with the rise of the number of hidden layers, the chance of convergence to a local minima increases, resulting in overfitting (PANCHAL et al., 2011), especially when using a neural network with great learning capacity using few samples, in which the neural networks memorize the training data but lose the ability to generalize (DUMITRU; MARIA, 2013).

For this study case, the best neural network was the one with two and three hidden layers. In addition, careful selection of input metrics is important when using the ANN technique, since metrics that are unrelated to variables to be estimated can have a negative influence on the predictive power of the model (BROSOFKSKE et al., 2014; NISKA et al., 2009). Thus, the use of LiDAR metrics transformed by the PCAs was effective in the use of ANN, being the most appropriate method in the estimates in this study.

When working with a multilayer perceptrons neural network, usually one hidden layer is sufficient to estimate variables (PANCHAL et al., 2011). In complex problems, where discontinuous data modelling is required, two hidden layers can well represent complex functions, with better fitting (HAYKIN, 2010; PANCHAL et al., 2011). However, it is not usual to use more than two hidden layers to estimate variables, due to the risk of overfitting (PANCHAL et al., 2011).

OLS and RF were the methods presenting an intermediate performance in the estimation of forest variables, according to the criteria used in this study, for the selection of the most appropriate regression technique. The RMSE (%) bars in Figure 5a are higher with OLS, using uncorrelated LiDAR metrics, than with OLS using PCAs. However, these techniques showed large bias, especially for the BA, DEN, and NTS variables. The transformations of the

independent variables in the OLS, such as logarithm, square root, square, or cube, can improve estimates using this regression method, especially when the assumptions of classical statistics are not met (BEN-HUR; WESTON, 2010). However, these transformations do not guarantee unbiased estimates, and when returned to the original scale, a bias is introduced, requiring an appropriate adjustment to avoid introducing a large bias in the estimates (BROSOFKSKE et al., 2014; NÆSSET, 2002). The transformation of PCs may have introduced bias in the estimation by OLS (Figure 5b), but this has not been quantified. The estimate by the RF method, on the other hand, had similar behaviors, both according to RMSE (%) of the variables and to the bias. Some issues can influence the estimates with the RF (YU et al., 2011): the number of built decision trees, the number of variables randomly sampled as candidates at each split, and the number of training samples. According to the same authors, the amount of training data is an important issue when using RF, and was confirmed by Silva et al. (2020), who concluded that from 30% of the sample size, the method tends to improve and stabilize the RMSE (%) and bias. As the number of training data in this study was low, they may have negatively influenced the estimates with the RF.

Comparing all *kernels* for the SVM regression method, the linear model produced the least accurate estimates, mainly with the input of 54 LiDAR metrics. This may indicate that the training patterns are not linearly separable, presenting the largest RMSE (%) values and underestimating five of the seven estimated variables (Figure 5). In a forest located in the French Alps, the SVM technique was assessed with both a linear and a radial *kernel*, to estimate some forest parameters (MONNET; CHANUSSOT; BERGER, 2011). The mathematical combination of some metrics, as well as the use of PCA, to reduce the dimensionality of the data, were effective when using the linear *kernel*. In Figure 5a, the use of PCAs significantly improved the estimate with SVM–L and SVM-R, but in our case, even removing the most correlated variables, a less accurate result was obtained with the radial *kernel*. In addition, the same authors commented that the presence of outliers and the risk of overfitting in the SVM models can reduce the estimates by this method. In the estimate of the volume in commercial plantations of *Eucalyptus* spp., Görgens; Montaghi; Rodriguez, (2015) and Silva et al. (2020) used SVM with the radial base function *kernel*. Both authors mentioned the great estimation power of the SVM, but compared to the other methods, it presented slightly higher RMSE (%) values. This behaviour was also observed in this study, in which there was a small difference between the polynomial and radial *kernel* for the average value of RMSE (%) in the estimates, and was more evidenced with the use of PCs as inputs (22.9% for SVM–5–P and 16.7% for

SVM–5–R; 23.7% and 19.9% for SVM–15–P and SVM–15–R, respectively; and for SVM–54–P, 18.7% and 24.9% for SVM–54–R), but the values were slightly higher than those found for the RF (Figure 5a), for example.

The previous analysis was focused on the comparison of machine learning techniques for estimating forest variables. The following discussions will emphasize the comparison of the estimated variables. To the best of our knowledge, there are no studies estimating stand and diversity variables for native Brazilian Atlantic forests, and there are very few related to tropical forests, most of them focusing on biomass estimation. Thus, our comparisons were made with studies that estimated the same stand and diversity variables but for native forests in other countries.

Table 5 shows the RMSE (%) values obtained for each variable and the respective regression method that provided the best estimate. The MDBH was the variable estimated with higher accuracy (RMSE of 5.6%). Our results were better than those presented at Monnet; Chanussot; Berger, (2011) whose RMSE value was 14.6% and at Reis et al. (2016) whose RMSE was 33%. In the aforementioned studies, greater variability of MDBH was observed among survey plots, which may have resulted in higher RMSE values. In addition, our best result was achieved using the PCs with the ANN, while in those studies, multiple linear regressions using raw LiDAR metrics were performed. A similar observation was done about the results obtained for the QMD variable. The stepwise method was used by Næsset, (2002) to select the metrics that would feed the multiple linear regression model, and they estimated the QMD with a deviation from 12.5% to 14%. Other authors like Woods; Lim; Treitz, (2008) also used multiple linear regression and achieved deviations of 15.4% and 30.5% for the QMD, while our best result had a deviation (i.e., RMSE) of 5.2%.

Vincent et al. (2012) estimated the forest variables QMD, BA, and DEN in a tropical forest in French Guiana, using simple and multiple linear regression with LiDAR metrics and stand variables as inputs in various forest sites, such as mature, explored, and secondary forests. Adjusting general and specific equations by site, the regression by forest site showed lower RMSE values for BA and DEN (7.9% and 9.1%, respectively). For the QMD, the regression for the whole area better estimated this variable, with an RMSE of 4.9%, while DEN presented slight variations compared to that found in this study, in the best case (16.5%), BA already presented greater errors. Some authors like (MONNET; CHANUSSOT; BERGER, 2011; NÆSSET, 2002, 2004b; REIS et al., 2016; WOODS; LIM; TREITZ, 2008) have studied natural boreal and temperate forests, reporting the difficulty in determining these variables using

LiDAR metrics. They achieved RMSE ranges for BA between 18 and 46.8% and for DEN between 18.4 and 128.6%.

The relationships of the basal area with density within the forest structure are very complex, varying according to spacing and the stage of forest development, among others. This results in patterns that are more difficult to interpret and consequently estimate. Thus, there is a set of assumptions and site-specific considerations that must be made before estimating these highly variable variables (GÖRGENS; MONTAGHI; RODRIGUEZ, 2015). This confirms the results of Vincent et al. (2012), who improved the estimates by separating the forest into smaller sites. The same analogy for the variable NTS can be done, which varies greatly in different types and stages of forest development, especially for a tropical forest. In a natural forest in southern England, Sumnall; Hill; Hinsley, (2016) estimated NTS using LiDAR metrics and individual tree crown metrics as inputs in multiple linear regression. As a result, the RMSE (%) in that study was 25%, while our result was 27.6%. This means that, regardless of the forest typology, the variables BA, DEN, and NTS are difficult to estimate with LiDAR metrics.

Diversity indexes give more valuable data than the number of species since they provide information on the diversity and floristic composition of the forest. Due to the high variability within the site and the leaf-off and leaf-on conditions of the trees, Sumnall; Hill; Hinsley, (2016) estimated the Shannon–Waver and Simpson indexes with an RMSE of 37 and 24%, respectively, while these indexes were estimated in this study, respectively, with a RMSE of 10 and 8.4%.

Considering all issues, such as heterogeneous tropical forest, low field sample size, and the criteria used to select the best results (lowest value of RMSE (%), bias (%) close to zero, and low value of AICc) a neural network with complex architecture (two and three hidden layers) may overfit sampled data. As stated by Panchal et al. (2011), the use of one hidden layer is usually enough to solve problems using ANNs, however, an erroneous configuration of neurons inside the hidden layer can also cause overfitting. Thus, for future studies, it would be recommended to test ANN with one hidden layer, but varying the number of neurons, as it was done by Da Rocha et al. (2018); Gorgens et al. (2014); Reis et al. (2016, 2018).

5 CONCLUSIONS

The results obtained in this work demonstrated that it is feasible to use LiDAR metrics to estimate forest variables in a tropical forest with a particular focus on the Atlantic Forest of Brazil, using LiDAR metrics with different machine learning approaches.

Methods to reduce the data dimensionality or selection of variables were of particular importance to achieve the results presented, mainly using the principal component analysis (PCA). In this case, the combination of metrics of elevation, intensity, and pulse returns allowed the relevant information in these metrics to be contained on principal components (PCs).

Considering the adopted criteria for choosing the best modelling technique, principal components (PCs) as new variables for artificial neural networks (ANNs) achieved the best results. ANNs with two hidden layers better estimated the mean diameter at breast height (MDBH), quadratic mean diameter (QMD), basal area (BA), and density (DEN) variables. Three hidden layers were the best ANNs for a number of tree species (NTS) variables, Shannon–Weaver (H') and Simpson (D) diversity indexes.

For ANN, the predictor variables with the greatest contribution to the estimation of forest variables, were the fifth PC, for the MDBH; the second PC for QMD and BA; the fourth PC for DEN; the first PC for NTS, and the third PC for the H' and D indices.

While ANN was the most suitable regression technique for estimating the studied variables, support vector machine (SVM) with linear kernel, using 54 LiDAR metrics as input data, presented the worst performance, with the highest RMSE values (%) and more biased estimates.

It is important to note that it is a pioneering study to estimate the stand and diversity variables of this type of tropical forest, and the results presented can be improved later with more samples of field data and different areas for later validation. Nevertheless, these findings can be applied in the management and preservation of these endangered forest remnants with LiDAR data.

6 REFERENCES

- ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley interdisciplinary reviews: computational statistics**, [s. l.], v. 2, n. 4, p. 433–459, 2010.
- AHOKAS, E.; KAASALAINEN, S.; HYYPPÄ, J.; SUOMALAINEN, J. Calibration of the Optech ALTM 3100 laser scanner intensity data using brightness targets. **International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 36, n. Part 1, p. 1Á6, 2006.
- AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, [s. l.], v. 19, n. 6, p. 716–723, 1974.
- ASPRS. **LAS SPECIFICATION VERSION 1.3 – R11**. 2010. Disponível em: <http://www.asprs.org/a/society/committees/standards/LAS_1_3_r11.pdf>. Acesso em: 20 maio. 2021.
- BENESTY, J.; CHEN, J.; HUANG, Y.; COHEN, I. Pearson correlation coefficient. In: **Noise reduction in speech processing**. [s.l.] : Springer, 2009. p. 1–4.
- BEN-HUR, A.; WESTON, J. A user’s guide to support vector machines. In: **Data mining techniques for the life sciences**. [s.l.] : Springer, 2010. p. 223–239.
- BERK, R. A. **Statistical learning from a regression perspective**. [s.l.] : Springer, 2008. v. 14
- BERVEGLIERI, A.; IMAI, N. N.; TOMMASELLI, A. M.; CASAGRANDE, B.; HONKAVAARA, E. Successional stages and their evolution in tropical forests using multi-temporal photogrammetric surface models and superpixels. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 146, p. 548–558, 2018.
- BERVEGLIERI, A.; TOMMASELLI, A. M. G.; IMAI, N. N.; RIBEIRO, E. A. W.; GUIMARAES, R. B.; HONKAVAARA, E. Identification of successional stages and cover changes of tropical forest based on digital surface model analysis. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, [s. l.], v. 9, n. 12, p. 5385–5397, 2016.
- BOZDOGAN, H. Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. **Psychometrika**, [s. l.], v. 52, n. 3, p. 345–370, 1987.
- BREIMAN, L. Random forests, machine learning 45. **J. Clin. Microbiol**, [s. l.], v. 2, n. 30, p. 199–228, 2001.
- BROSOFKSKE, K. D.; FROESE, R. E.; FALKOWSKI, M. J.; BANSKOTA, A. A review of methods for mapping and prediction of inventory attributes for operational forest management. **Forest Science**, [s. l.], v. 60, n. 4, p. 733–756, 2014.
- BROWER, J. E.; ZAR, J. H.; VON ENDE, C. A. **Field and laboratory methods for general ecology**. Dubuque: Wm. C. [s.l.] : Brown Publishers, 1984.

CARVALHO, A.; FACELI, K.; LORENA, A.; GAMA, J. Inteligência Artificial—uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, [s. l.], 2011.

CHIARELLO, F.; STEINER, M. T. A.; OLIVEIRA, E. B. D.; ARCE, J. E.; FERREIRA, J. C. ARTIFICIAL NEURAL NETWORKS APPLIED IN FOREST BIOMETRICS AND MODELING: STATE OF THE ART (JANUARY/2007 TO JULY/2018). **Cerne**, [s. l.], v. 25, n. 2, p. 140–155, 2019.

CLIFF, N. The eigenvalues-greater-than-one rule and the reliability of components. **Psychological bulletin**, [s. l.], v. 103, n. 2, p. 276, 1988.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, [s. l.], v. 20, n. 3, p. 273–297, 1995.

DA ROCHA, S. J. S. S.; TORRES, C. M. M. E.; JACOVINE, L. A. G.; LEITE, H. G.; GELCER, E. M.; NEVES, K. M.; SCHETTINI, B. L. S.; VILLANOVA, P. H.; DA SILVA, L. F.; REIS, L. P. Artificial neural networks: Modeling tree survival and mortality in the Atlantic Forest biome in Brazil. **Science of the total environment**, [s. l.], v. 645, p. 655–661, 2018.

DE SOUZA WERNECK, M.; SOBRAL, M. E. G.; ROCHA, C. T. V.; LANDAU, E. C.; STEHMANN, J. R. Distribution and endemism of angiosperms in the Atlantic Forest. **Natureza & Conservação**, [s. l.], v. 9, p. 188–193, 2011.

DONOGHUE, D. N.; WATT, P. J.; COX, N. J.; WILSON, J. Remote sensing of species mixtures in conifer plantations using LiDAR height and intensity data. **Remote Sensing of Environment**, [s. l.], v. 110, n. 4, p. 509–522, 2007.

DUMITRU, C.; MARIA, V. Advantages and Disadvantages of Using Neural Networks for Predictions. **Ovidius University Annals, Series Economic Sciences**, [s. l.], v. 13, n. 1, 2013.

FALKOWSKI, M. J.; EVANS, J. S.; MARTINUZZI, S.; GESSLER, P. E.; HUDAK, A. T. Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. **Remote Sensing of Environment**, [s. l.], v. 113, n. 5, p. 946–956, 2009.

FIASCHI, P.; PIRANI, J. R. Review of plant biogeographic studies in Brazil. **Journal of systematics and evolution**, [s. l.], v. 47, n. 5, p. 477–496, 2009.

FREESE, F. **Linear regression methods for forest research**. [s.l.] : US Department of Agriculture, Forest Service, Forest Products Laboratory, 1964. v. 17

GARCÍA-GUTIÉRREZ, J.; MARTÍNEZ-ÁLVAREZ, F.; TRONCOSO, A.; RIQUELME, J. C. A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. **Neurocomputing**, [s. l.], v. 167, p. 24–31, 2015.

GATZIOLIS, D. Dynamic range-based intensity normalization for airborne, discrete return lidar data of forest canopies. **Photogrammetric Engineering & Remote Sensing**, [s. l.], v. 77, n. 3, p. 251–259, 2011.

GORGENS, E. B.; LEITE, H. G.; GLERIANI, J. M.; SOARES, C. P. B.; CEOLIN, A. Influência da arquitetura na estimativa de volume de árvores individuais por meio de redes neurais artificiais. **Revista Árvore**, [s. l.], v. 38, n. 2, p. 289–295, 2014.

GÖRGENS, E. B.; MONTAGHI, A.; RODRIGUEZ, L. C. E. 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**, [s. l.], v. 116, p. 221–227, 2015.

GÜNTHER, F.; FRITSCH, S. neuralnet: Training of neural networks. **The R journal**, [s. l.], v. 2, n. 1, p. 30–38, 2010.

HADDAD, N. M.; BRUDVIG, L. A.; CLOBERT, J.; DAVIES, K. F.; GONZALEZ, A.; HOLT, R. D.; LOVEJOY, T. E.; SEXTON, J. O.; AUSTIN, M. P.; COLLINS, C. D. Habitat fragmentation and its lasting impact on Earth's ecosystems. **Science advances**, [s. l.], v. 1, n. 2, p. e1500052, 2015.

HARGREAVES, P. Phytosociology in Brazil. **J Plant Sci Biotechnol**, [s. l.], v. 2, p. 12–20, 2008.

HAYKIN, S. **Neural Networks and Learning Machines, 3/E**. [s.l.] : Pearson Education India, 2010.

HONGYU, K.; SANDANIELO, V. L. M.; DE OLIVEIRA JUNIOR, G. J. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and Science**, [s. l.], v. 5, n. 1, p. 83–90, 2016.

HOPKINS, M. J. Modelling the known and unknown plant biodiversity of the Amazon Basin. **Journal of Biogeography**, [s. l.], v. 34, n. 8, p. 1400–1411, 2007.

HOPKINSON, C. The influence of flying altitude, beam divergence, and pulse repetition frequency on laser pulse return intensity and canopy frequency distribution. **Canadian Journal of Remote Sensing**, [s. l.], v. 33, n. 4, p. 312–324, 2007.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural networks**, [s. l.], v. 2, n. 5, p. 359–366, 1989.

HUANG, C.; DAVIS, L. S.; TOWNSHEND, J. R. G. An assessment of support vector machines for land cover classification. **International Journal of remote sensing**, [s. l.], v. 23, n. 4, p. 725–749, 2002.

HUDAK, A. T.; STRAND, E. K.; VIERLING, L. A.; BYRNE, J. C.; EITEL, J. U.; MARTINUZZI, S.; FALKOWSKI, M. J. Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. **Remote Sensing of Environment**, [s. l.], v. 123, p. 25–40, 2012.

HURVICH, C. M.; TSAI, C.-L. Regression and time series model selection in small samples. **Biometrika**, [s. l.], v. 76, n. 2, p. 297–307, 1989.

HYYPÄ, J.; YU, X.; HYYPPÄ, H.; VASTARANTA, M.; HOLOPAINEN, M.; KUKKO, A.; KAARTINEN, H.; JAAKKOLA, A.; VAAJA, M.; KOSKINEN, J. Advances in forest inventory using airborne laser scanning. **Remote sensing**, [s. l.], v. 4, n. 5, p. 1190–1207, 2012.

IBGE. **Manual Técnico da Vegetação Brasileira**. 2012. Disponível em: <<https://biblioteca.ibge.gov.br/visualizacao/livros/liv63011.pdf>>. Acesso em: 21 jan. 2021.

ISENBURG, M. **LAStools-efficient LiDAR processing software**. 2014. Disponível em: <<http://lastools.org/>>. Acesso em: 12 nov. 2020.

KAASALAINEN, S.; HYYPPÄ, J.; LITKEY, P.; HYYPPÄ, H.; AHOKAS, E.; KUKKO, A.; KAARTINEN, H. Radiometric calibration of ALS intensity. **Int. Arch. Photogramm. Remote Sens**, [s. l.], v. 36, p. 201–205, 2007.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, [s. l.], v. 23, n. 3, p. 187–200, 1958.

KARATZOGLOU, A.; SMOLA, A.; HORNIK, K.; ZEILEIS, A. kernlab-an S4 package for kernel methods in R. **Journal of statistical software**, [s. l.], v. 11, n. 9, p. 1–20, 2004.

KHOSRAVIPOUR, A.; SKIDMORE, A. K.; ISENBURG, M.; WANG, T.; HUSSIN, Y. A. Generating pit-free canopy height models from airborne lidar. **Photogrammetric Engineering & Remote Sensing**, [s. l.], v. 80, n. 9, p. 863–872, 2014.

KUHN, M. Building predictive models in R using the caret package. **Journal of statistical software**, [s. l.], v. 28, n. 5, p. 1–26, 2008.

LATIFI, H.; FASSNACHT, F.; KOCH, B. Forest structure modeling with combined airborne hyperspectral and LiDAR data. **Remote Sensing of Environment**, [s. l.], v. 121, p. 10–25, 2012.

LÊ, S.; JOSSE, J.; HUSSON, F. FactoMineR: an R package for multivariate analysis. **Journal of statistical software**, [s. l.], v. 25, n. 1, p. 1–18, 2008.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R news**, [s. l.], v. 2, n. 3, p. 18–22, 2002.

LUMLEY, T.; LUMLEY, M. T. Package ‘leaps’. **Regression Subset Selection. Thomas Lumley Based on Fortran Code by Alan Miller. Available online: <http://CRAN.R-project.org/package=leaps> (accessed on 18 March 2018)**, [s. l.], 2013.

LUO, S.; CHEN, J. M.; WANG, C.; GONSAMO, A.; XI, X.; LIN, Y.; QIAN, M.; PENG, D.; NIE, S.; QIN, H. Comparative performances of airborne LiDAR height and intensity data for leaf area index estimation. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, [s. l.], v. 11, n. 1, p. 300–310, 2017.

MAGURRAN, A. E. **Ecological diversity and its measurement**. [s.l.] : Princeton university press, 1988.

MANLY, B. F.; ALBERTO, J. A. N. **Multivariate statistical methods: a primer**. [s.l.] : CRC press, 2016.

MCQUARRIE, A. D.; TSAI, C.-L. **Regression and time series model selection**. [s.l.] : World Scientific, 1998.

MEYER, D.; DIMITRIADOU, E.; HORNIK, K.; WEINGEESSEL, A.; LEISCH, F. **e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1**. 2012. Disponível em: <<https://rdrr.io/rforge/e1071/>>. Acesso em: 25 fev. 2021.

MILTADOU, M.; AGAPIOU, A.; GONZALEZ ARACIL, S.; HADJIMITSIS, D. G. Detecting Dead Standing Eucalypt Trees from Voxellised Full-Waveform Lidar Using Multi-Scale 3D-Windows for Tackling Height and Size Variations. **Forests**, [s. l.], v. 11, n. 2, p. 161, 2020.

MMA; IBAMA; ICMBIO. **Plano de manejo da Estação Ecológica Mico - Leão - Preto.**, 2007. Disponível em: <http://www.icmbio.gov.br/portal/images/stories/imgs-unidades-coservacao/esec_mico_leao_preto.pdf>. Acesso em: 6 nov. 2019.

MOHAMMED, A. A.; NAUGLER, C.; FAR, B. H. Emerging business intelligence framework for a clinical laboratory through big data analytics. **Emerging trends in computational biology, bioinformatics, and systems biology: algorithms and software tools**. New York: Elsevier/Morgan Kaufmann, [s. l.], p. 577–602, 2015.

MONNET, J.-M.; CHANUSSOT, J.; BERGER, F. Support vector regression for the estimation of forest stand parameters using airborne laser scanning. **IEEE Geoscience and remote sensing letters**, [s. l.], v. 8, n. 3, p. 580–584, 2011.

MORSDORF, F.; FREY, O.; MEIER, E.; ITTEN, K. I.; ALLGÖWER, B. Assessment of the influence of flying altitude and scan angle on biophysical vegetation products derived from airborne laser scanning. **International journal of remote sensing**, [s. l.], v. 29, n. 5, p. 1387–1406, 2008.

MYERS, N.; MITTERMEIER, R. A.; MITTERMEIER, C. G.; DA FONSECA, G. A.; KENT, J. Biodiversity hotspots for conservation priorities. **Nature**, [s. l.], v. 403, n. 6772, p. 853, 2000.

NÆSSET, E. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. **Remote sensing of environment**, [s. l.], v. 80, n. 1, p. 88–99, 2002.

NÆSSET, E. Effects of different flying altitudes on biophysical stand properties estimated from canopy height and density measured with a small-footprint airborne scanning laser. **Remote Sensing of Environment**, [s. l.], v. 91, n. 2, p. 243–255, 2004. a.

NÆSSET, E. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. **Scandinavian Journal of Forest Research**, [s. l.], v. 19, n. 2, p. 164–179, 2004. b.

NISKA, H.; SKON, J.-P.; PACKALEN, P.; TOKOLA, T.; MALTAMO, M.; KOLEHMAINEN, M. Neural networks for the prediction of species-specific plot volumes using airborne laser scanning and aerial photographs. **IEEE Transactions on Geoscience and Remote Sensing**, [s. l.], v. 48, n. 3, p. 1076–1085, 2009.

NOGUEIRA, E. M.; FEARNSIDE, P. M.; NELSON, B. W.; BARBOSA, R. I.; KEIZER, E. W. H. Estimates of forest biomass in the Brazilian Amazon: new allometric equations and adjustments to biomass from wood-volume inventories. **Forest Ecology and Management**, [s. l.], v. 256, n. 11, p. 1853–1867, 2008.

NOGUEIRA, E. M.; NELSON, B. W.; FEARNSIDE, P. M. Volume and biomass of trees in central Amazonia: influence of irregularly shaped and hollow trunks. **Forest Ecology and Management**, [s. l.], v. 227, n. 1–2, p. 14–21, 2006.

PANCHAL, G.; GANATRA, A.; KOSTA, Y. P.; PANCHAL, D. Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. **International Journal of Computer Theory and Engineering**, [s. l.], v. 3, n. 2, p. 332–337, 2011.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2017.

REIS, L. P.; DE SOUZA, A. L.; DOS REIS REIS, P. C. M.; MAZZEI, L.; BINOTI, D. H. B.; LEITE, H. G. Prognose da distribuição diamétrica na Amazônia utilizando redes neurais artificiais e autômatos celulares. **Floresta**, [s. l.], v. 48, n. 1, p. 93–102, 2018.

REIS, L. P.; DE SOUZA, A. L.; MAZZEI, L.; DOS REIS, P. C. M.; LEITE, H. G.; SOARES, C. P. B.; TORRES, C. M. M. E.; DA SILVA, L. F.; RUSCHEL, A. R. Prognosis on the diameter of individual trees on the eastern region of the amazon using artificial neural networks. **Forest Ecology and Management**, [s. l.], v. 382, p. 161–167, 2016.

REX, F. E.; SILVA, C. A.; DALLA CORTE, A. P.; KLAUBERG, C.; MOHAN, M.; CARDIL, A.; SILVA, V. S. Da; ALMEIDA, D. R. A. De; GARCIA, M.; BROADBENT, E. N. Comparison of Statistical Modelling Approaches for Estimating Tropical Forest Aboveground Biomass Stock and Reporting their Changes in Low-intensity Logging Areas using Multi-temporal LiDAR Data. **Remote Sensing**, [s. l.], v. 12, n. 9, p. 1498, 2020.

RIBEIRO, M. C.; METZGER, J. P.; MARTENSEN, A. C.; PONZONI, F. J.; HIROTA, M. M. The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. **Biological conservation**, [s. l.], v. 142, n. 6, p. 1141–1153, 2009.

RIEGL. **DataSheet LMS-Q680i**. 2012. Disponível em: <http://www.riegl.com/uploads/tx_pxpriegldownloads/10_DataSheet_LMS-Q680i_28-09-2012_01.pdf>. Acesso em: 7 abr. 2021.

ROUSSEL, J.-R.; AUTY, D.; DE BOISSIEU, F.; MEADOR, A. S. **lidR: Airborne LiDAR data manipulation and visualization for forestry applications**. 2018. Disponível em: <<https://rdr.io/cran/lidR/>>. Acesso em: 21 jan. 2021.

ROUSSEL, J.-R.; BOURDON, J.-F.; ACHIM, A. Range-based intensity normalization of ALS data over forested areas using a sensor tracking method from multiple returns. [s. l.], 2020.

SCHIKOWSKI, A. B.; CORTE, A. P.; RUZA, M. S.; SANQUETTA, C. R.; MONTANO, R. A. Modeling of stem form and volume through machine learning. **Anais da Academia Brasileira de Ciências**, [s. l.], v. 90, n. 4, p. 3389–3401, 2018.

SHAN, J.; TOTH, C. K. **Topographic laser ranging and scanning: principles and processing**. [s.l.] : CRC press, 2018.

SHI, Y.; WANG, T.; SKIDMORE, A. K.; HEURICH, M. Important LiDAR metrics for discriminating forest tree species in Central Europe. **ISPRS journal of photogrammetry and remote sensing**, [s. l.], v. 137, p. 163–174, 2018.

SHIBATA, K.; IKEDA, Y. Effect of number of hidden neurons on learning in large-scale layered neural networks. In: 2009 ICCAS-SICE 2009, **Anais...** : IEEE, 2009.

- SHIN, J.; TEMESGEN, H.; STRUNK, J. L.; HILKER, T. Comparing modeling methods for predicting forest attributes using LiDAR metrics and ground measurements. **Canadian Journal of Remote Sensing**, [s. l.], v. 42, n. 6, p. 739–765, 2016.
- SILVA, C. A.; KLAUBERG, C.; E CARVALHO, S. de P. C.; HUDAK, A. T. Mapping aboveground carbon stocks using LiDAR data in Eucalyptus spp. plantations in the state of São Paulo, Brazil. **Scientia Forestalis**. **42 (104): 591-604.**, [s. l.], v. 42, n. 104, p. 591–604, 2014.
- SILVA, C. A.; KLAUBERG, C.; HUDAK, A. T.; VIERLING, L. A.; JAAFAR, W. S. W. M.; MOHAN, M.; GARCIA, M.; FERRAZ, A.; CARDIL, A.; SAATCHI, S. Predicting stem total and assortment volumes in an industrial Pinus taeda L. forest plantation using airborne laser scanning data and random forest. **Forests**, [s. l.], v. 8, n. 7, p. 254, 2017.
- SILVA, V. S. Da; SILVA, C. A.; MOHAN, M.; CARDIL, A.; REX, F. E.; LOUREIRO, G. H.; ALMEIDA, D. R. A. De; BROADBENT, E. N.; GORGENS, E. B.; DALLA CORTE, A. P. Combined Impact of Sample Size and Modeling Approaches for Predicting Stem Volume in Eucalyptus spp. Forest Plantations Using Field and LiDAR Data. **Remote Sensing**, [s. l.], v. 12, n. 9, p. 1438, 2020.
- SUMNALL, M. J.; HILL, R. A.; HINSLEY, S. A. Comparison of small-footprint discrete return and full waveform airborne LiDAR data for estimating multiple forest variables. **Remote Sensing of Environment**, [s. l.], v. 173, p. 214–223, 2016.
- VENIER, L. A.; SWYSTUN, T.; MAZEROLLE, M. J.; KREUTZWEISER, D. P.; WAINIO-KEIZER, K. L.; MCILWRICK, K. A.; WOODS, M. E.; WANG, X. Modelling vegetation understory cover using LiDAR metrics. **PloS one**, [s. l.], v. 14, n. 11, p. e0220096, 2019.
- VINCENT, G.; SABATIER, D.; BLANC, L.; CHAVE, J.; WEISSENBACHER, E.; PÉLISSIER, R.; FONTY, E.; MOLINO, J.-F.; COUTERON, P. Accuracy of small footprint airborne LiDAR in its predictions of tropical moist forest stand structure. **Remote Sensing of Environment**, [s. l.], v. 125, p. 23–33, 2012.
- WILLIAMS, K. J.; FORD, A.; ROSAUER, D. F.; DE SILVA, N.; MITTERMEIER, R.; BRUCE, C.; LARSEN, F. W.; MARGULES, C. Forests of East Australia: the 35th biodiversity hotspot. In: **Biodiversity hotspots**. [s.l.] : Springer, 2011. p. 295–310.
- WOODS, M.; LIM, K.; TREITZ, P. Predicting forest stand variables from LIDAR data in the Great Lakes St. Lawrence Forest of Ontario. **The Forestry Chronicle**, [s. l.], v. 84, n. 6, p. 827–839, 2008.
- YAO, W.; KRULL, J.; KRZYSZEK, P.; HEURICH, M. Sensitivity analysis of 3D individual tree detection from LiDAR point clouds of temperate forests. **Forests**, [s. l.], v. 5, n. 6, p. 1122–1142, 2014.
- YU, X.; HYYPPÄ, J.; VASTARANTA, M.; HOLOPAINEN, M.; VIITALA, R. Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. **ISPRS Journal of Photogrammetry and remote sensing**, [s. l.], v. 66, n. 1, p. 28–37, 2011.

ZHANG, Z.; LIU, X. Support vector machines for tree species identification using LiDAR-derived structure and intensity variables. **Geocarto International**, [s. l.], v. 28, n. 4, p. 364–378, 2013.

ZHAO, K.; POPESCU, S.; MENG, X.; PANG, Y.; AGCA, M. Characterizing forest canopy structure with lidar composite metrics and machine learning. **Remote Sensing of Environment**, [s. l.], v. 115, n. 8, p. 1978–1996, 2011.

APPENDIX A – CHAPTER 3

Table A1 – Statistics of forest variables calculated from each of 15 surveyed plots.

Plots	Field Variables								
	DBH (cm)			QMD (cm)	BA (m ² ha ⁻¹)	DEN (trees ha ⁻¹)	NTS	H'	D
	Minimum	Mean	Maximum						
1	3.5	11.2	74.8	13.8	25.2	1688	14	1.63	0.74
2	3.5	10.4	97.4	14.5	24.3	1477	17	1.61	0.72
3	3.5	8.5	41.4	9.8	15.9	2062	18	1.46	0.63
4	4.0	9.5	51.6	11.0	22.5	2286	24	1.81	0.72
5	4.0	9.6	63.7	11.6	21.2	1897	12	1.51	0.70
6	4.0	13.9	91.0	18.8	30.7	1101	18	1.71	0.68
7	3.6	11.1	76.4	15.3	16.0	869	16	1.51	0.60
8	3.8	10.9	42.7	13.1	10.0	741	12	1.74	0.74
9	3.5	9.4	50.9	12.5	10.9	889	13	1.37	0.60
10	3.8	9.1	30.9	10.7	9.0	991	18	2.04	0.79
11	3.8	9.6	33.7	11.0	5.6	593	10	1.45	0.67
12	3.8	10.6	48.4	13.2	17.8	1298	14	1.18	0.49
13	3.8	10.5	54.4	14.0	5.8	380	11	1.62	0.71
14	3.8	12.4	49.7	15.0	15.1	853	12	1.46	0.66
15	3.8	11.8	49.3	14.3	12.4	779	11	1.32	0.57

Table A2 – The best multiple linear regression models fitted for the 7 forest variables using the 5 PCs and 15 uncorrelated LiDAR metrics as input data.

Variable	Input data - PCA			
MDBH	Estimate	Std. Error	t value	p-value
(Intercept)	10.3703	0.3265	31.759	1.05E-13***
PC2	0.1323	0.0558	2.371	0.0339*
QMD	Estimate	Std. Error	t value	p-value
(Intercept)	12.815	0.42004	30.509	1.75E-13***
PC2	0.29324	0.07178	4.086	0.00129**
BA	Estimate	Std. Error	t value	p-value
(Intercept)	13.8749	1.6515	8.402	2.27E-06***
PC1	0.7026	0.2673	2.629	0.022*
PC2	0.4919	0.2651	1.856	0.0482*
DEN	Estimate	Std. Error	t value	p-value
(Intercept)	1099	138.32	7.945	4.03E-06***
PC1	56.61	22.39	2.529	0.0265*
PC2	-22.75	22.2	-1.025	0.0326*
NTS	Estimate	Std. Error	t value	p-value
(Intercept)	13.7278	0.8217	16.707	1.13E-09***
PC1	0.4221	0.1356	3.112	0.00898**
PC3	-0.1778	0.1618	-1.099	0.0293*
H'	Estimate	Std. Error	t value	p-value
(Intercept)	1.52863	0.056597	27.009	8.34E-13***
PC1	0.01422	0.009341	1.522	0.042*
D	Estimate	Std. Error	t value	p-value
(Intercept)	0.65872	0.021761	30.27	1.94E-13***
PC1	0.00433	0.003592	1.205	0.025*

Values followed by '*' are significant at the level $\alpha = 0.05$; values followed by '**' are significant at level $\alpha = 0.01$ and values followed by '***' are significant at level $\alpha = 0.001$.

Variable	Input data – 15 uncorrelated LiDAR metrics			
MDBH	Estimate	Std. Error	t value	p-value
(Intercept)	50.8028	14.0312	3.621	0.00311**
zpcum9	-0.4116	0.1435	-2.869	0.01317*
QMD	Estimate	Std. Error	t value	p-value
(Intercept)	60.0498	22.0737	2.72	0.0186*
zpcum9	-0.5153	0.2174	-2.37	0.0354*
BA	Estimate	Std. Error	t value	p-value
(Intercept)	-45.687	16.0257	-2.851	0.0158*
zmax	1.1912	0.3905	3.051	0.011*
zskew	-6.8411	2.928	-2.336	0.0394*
imax	0.2826	0.1085	2.604	0.0245*
DEN	Estimate	Std. Error	t value	p-value
(Intercept)	-4274.3	2299.93	-1.858	0.0401*
zskew	-546.53	222.04	-2.461	0.03168
zpcum7	52.64	21.71	2.424	0.0338*
NTS	Estimate	Std. Error	t value	p-value
(Intercept)	-25.154	12.67328	-1.985	0.03267*
zq75	1.12544	0.28003	4.019	0.00202**
zpcum5	-0.171	0.05446	-3.14	0.00941**
zpcum7	0.4479	0.13257	3.379	0.00616**
H'	Estimate	Std. Error	t value	p-value
(Intercept)	-8.569	3.6518	-2.346	0.0355*
ipcumzq90	0.1115	0.0402	2.774	0.0158*
D	Estimate	Std. Error	t value	p-value
(Intercept)	-3.0444	1.37869	-2.208	0.0458*
ipcumzq90	0.04088	0.01518	2.693	0.0184*

Values followed by '*' are significant at the level $\alpha = 0.05$; values followed by '**' are significant at level $\alpha = 0.01$ and values followed by '***' are significant at level $\alpha = 0.001$.

CHAPTER 4 – EXTRACTING TROPICAL FOREST VARIABLES BY VOXEL APPROACH – HOW DO LIDAR DATA TYPE AND FLIGHT HEIGHT AFFECT ESTIMATES?

This chapter was submitted to the ISPRS Open Journal of Photogrammetry and Remote Sensing

1 INTRODUCTION

Light detection and ranging (LiDAR) data are a powerful tool for monitoring forest environments. The ability of the pulses to penetrate through small openings in the vegetation canopy allows to obtain information on forest structure and terrain (LIM et al., 2003; SHAN; TOTH, 2018), which in turn can be used to estimate important biophysical parameters of forests (HYYPPÄ et al., 2008; VAUHKONEN et al., 2012). LiDAR systems can be classified according to the types of return, i.e., as either discrete return (DR) or full waveform (FWF) systems.

Discrete return systems emit a pulse, and the receiving sensor records only those peaks detected above a threshold that is determined by a discriminating fractional constant (usually 50%) of the peak amplitude energy (SHAN; TOTH, 2018; THIEL; WEHR, 2004). This approach substantially reduces the volume of data recorded compared to the FWF data, in which all the returned signals are stored. During the development of laser scanners, the first systems could record only one pulse return, and these were configured to detect and record the highest or the lowest peak. Later, the technology evolved to record two discrete returns (the first and last) for each emitted pulse (LIM et al., 2003; SHAN; TOTH, 2018). The first pulse is best suited for measuring partially penetrable objects, such as tree tops; the last pulse is for measuring non-penetrable surfaces, such as terrain (FAVORSKAYA; JAIN, 2017; SHAN; TOTH, 2018; THIEL; WEHR, 2004). The evolution of LiDAR systems has allowed the recording of up to ten returns for each pulse emitted (SHAN; TOTH, 2018).

Full-waveform LiDAR systems record and digitize the entire amount of energy returned to the sensor after being backscattered by objects present on the surface and are sampled at equal time intervals (LIM et al., 2003; MALLET; BRETAR, 2009). With this approach, more information is recorded than by using DR systems. As the result, the waveform contains the properties of all elements intercepting the path of the emitted beam, which can improve the representation of the forest structure, including its vertical structure, canopy volume,

understory, and terrain (PIROTTI, 2011). The analysis of the complete waveform allows a better interpretation of the physical structure and geometric backscatter properties of the intercepted objects, presenting new opportunities for forestry studies (MALLET; BRETAR, 2009; REITBERGER; KRZYTEK; STILLA, 2006; SHAN; TOTH, 2018).

DR data are widely used because of a large number of tools available for the visualization and extraction of information from the point cloud (more details in Crespo-Peremarch and Ruiz, 2020), mainly metrics related to height distribution, pulse returns, and intensity (HYYPÄÄ et al., 2012; MORSDORF et al., 2008; NÆSSET, 2004a). These metrics serve as input data for regression models for estimating various forest variables, including tree height, diameter, above-ground biomass (AGB), tree density, volume, and basal area (more details about this topic is available in Maltamo et al., 2014). On the other hand, FWF data are less frequently used for estimating forest variables, owing to difficulties in visualization and interpretation and the large volume of data generated compared to the DR data. In addition, there are only a few computational tools for analysing and extracting information, e.g., metrics from FWF data (MALLET; BRETAR, 2009; MILTIADOU et al., 2021; ZHOU et al., 2018b).

Sumnall et al. (2016) noted in their research that the use of FWF LiDAR metrics had little importance in estimating the composition, structure, and diversity variables of forests in southern England. On the other hand, studies as Allouis et al. (2012); Anderson et al. (2016); Brown et al. (2018); Cao et al. (2016b, 2014); Hermosilla et al. (2013) and Shen et al. (2018) found great potential in using metrics obtained from FWF data to estimate forest variables, as they provide detailed information on the vertical structure of forests, with geometric and radiometric features, thus providing a better understanding of multi-layered forest structures. These authors also reported that the combination of FWF metrics with DR metrics are complementary, and the use of both in predictive models could allow an increase in the accuracy of estimates. However, there is a gap in the literature regarding the use of FWF data in tropical forests for the estimation of forest attributes beyond AGB.

A very important step in the processing of waveforms is the reduction of noisy data signals (de-noising) because even when the emitted pulse does not reach any object, the system registers low-amplitude signals, which can be considered as noise. Thus, low-level filtering must be applied, and samples with an amplitude lower than the noise level must be discarded (MILTIADOU et al., 2019). The filtering threshold is normally calculated based on the mean and standard deviation of the signal segments representing noise (DRAKE et al., 2002).

However, these considerations are based on large-footprint FWF data and can differ for small-footprint data.

The filtering threshold must be chosen carefully, as low values can still transmit data with noise, and high threshold values can approximate waveforms to discrete return data and even eliminate important signals. In addition, Liu. (2011) *apud* (LAI; ZHENG, 2015) mentioned that inappropriate choice of the de-noising method can smoothen the signal causing distortions in waveform, for instance, leading to peak amplitude reduction and increase in width pulse, which will hinder the extraction of information from waveforms. In our study, we decided to use low-level filtering, which implies that it removes very low amplitude values and does not distort waveforms.

There is great potential in using LiDAR metrics to model biophysical parameters of forests, but some variables, such as basal area, density, number of tree species, and above-ground biomass, are difficult to estimate. The relationships between LiDAR data and such forest attributes are complex, especially in case of dense tropical forests (BROWN et al., 2018), as they relate to the degree of forest development, successional stage, and spacing, making it difficult to establish standards for subsequent estimates. Thus, there is great variability in the errors for these estimated variables regardless of forest type. Errors can vary from 13.0% to 139.0% for basal area measurements; from 13.0% to 128.6% for density values (GOBAKKEN; NÆSSET, 2004; HOLMGREN, 2004; MALTAMO; SUVANTO; PACKALÉN, 2007; MARTINS-NETO et al., 2021; MONNET; CHANUSSOT; BERGER, 2011; NÆSSET, 2002, 2004b; SHIN et al., 2016); from 25.0% to 138.0% for the number of tree species (MARTINS-NETO et al., 2021; SUMNALL; HILL; HINSLEY, 2016) and between 11.8% and 95% for above-ground biomass estimations (CAO et al., 2014; GARCÍA-GUTIÉRREZ et al., 2015; GLEASON; IM, 2012; MCROBERTS et al., 2015; REX et al., 2020).

New approaches for extracting information from LiDAR data with different types of pulse returns have garnered interest aiming at increasing the accuracy of estimates in variables that are difficult to model, especially for tropical forests that require further studies. There are a few methods for processing waveforms, such as deconvolution, Gaussian decomposition for peak detection (MALLET et al., 2009; WAGNER et al., 2006) and voxel-based approaches (PERSSON et al., 2005).

With technological advances, many LiDAR sensors now record full waveform, while older models have an additive digitizer, thus recording discrete returns and waveform data

separately. Nevertheless, because the amount of full waveform LiDAR data is huge in comparison to discrete data, recent systems do not provide FWF but only processed peak points. More information about FWF data types is available in (ULLRICH; PFENNIGBAUER, 2011). Peak point extraction is convenient because existing workflows support point cloud interpretations, but there is still information discarded between peaks. Thus, this paper examines whether it concerns the extra cost of processing FWF LiDAR in comparison to using only extracted peak points (Peak Returns – PR) for the estimation of four forest variables (basal area, density, number of tree species, and above-ground biomass) in a remnant of the Brazilian Atlantic Forest.

The advantage of using the voxel-based approach for processing FWF data is the decrease in the influence of off-nadir scan angles and the possibility of taking full advantage of the information provided by waveforms (WANG et al., 2020). The voxel approach used in this study is available through the open-source software DASOS (MILTADOU et al., 2019), which analyses both PR and FWF data based on the interpretation of peaks or the entire waveform intercepting each voxel. This approach makes it feasible to compare the effects of using two different data types (PR and FWF) in the estimation of forest variables.

In addition to the return type, we investigated the influence of flight height on the results. Previous studies with discrete return data have verified such influences, e.g., Yu et al. (2004) concluded that the accuracy in estimating tree heights and the number of detected trees decreased as the flight height increased. According to Morsdorf et al. (2008); Næsset. (2004) and Persson et al. (2002), despite the lower pulse density on higher flights, the estimated parameter values did not differ significantly with data from lower and higher flight heights. To the best of our knowledge, no studies till date have analysed how flight height can affect FWF data.

Thus, this study aimed to compare the estimated values of four forest variables from attributes extracted from PR and FWF LiDAR data at 900 m and 2000 m flight heights, to assess the improvements when processing all waveform samples, as there is still extra information between peak returns, which can be useful in case of complex forests, such as tropical forests. In addition, the objective was to verify: (i) the effect of voxel size at different flight heights, (ii) the effect of different noise-level filtering thresholds on FWF data, and (iii) the effect of the same voxel approach for different data types obtained at different flight heights.

2 BACKGROUND ON FULL-WAVEFORM LIDAR DATA

The first LiDAR sensor, capable of digitizing waveforms with a large footprint was the laser vegetation image sensor (LVIS) developed by NASA in 1999 (CHAUVE et al., 2007). In 2004, the first LiDAR system with a small footprint, the RIEGL LMS-Q560 laser scanner, was developed. It can record the waveform of backscattered signal (HUG; ULLRICH; GRIMM, 2004). Thus, in addition to distance measurements, other physical properties of the objects in the illuminated path can be derived by analysing the digitized signal (MALLET; BRETAR, 2009). This analysis allowed the identification of multiple pulses and calculated the respective coordinates of points based on the peaks contained in the waveforms. The returned pulse provides more detailed information on the vertical structure of the targets, better describing the nature of the surfaces that can be used to locate discontinuities. In vegetated areas, it is possible to distinguish bare soil from the understory, as well as to use reflectivity and roughness parameters to model forest variables and classify complex forest areas (ALEXANDER et al., 2014; FERRAZ et al., 2016; OLIVEIRA; CENTENO; HAINOSZ, 2018; TOTH et al., 2011).

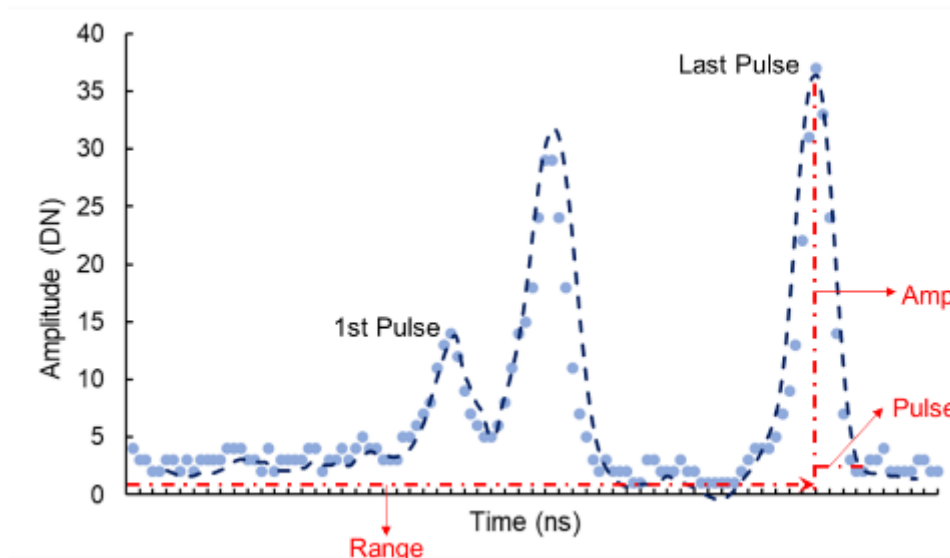
In most small-footprint LiDAR sensor systems (e.g., Riegl, TopEye, Optech), the returned pulses of the digitized waveform are very similar to ideal Gaussian functions (BRUGGISSER et al., 2017; WAGNER et al., 2006). A sample distance of 1 ns corresponds to 15 cm of distance (considering that the distance corresponds to half of the signal trajectory). However, the effective measuring distance depends on the nature and manner in which objects reflect the pulse (TOTH et al., 2011).

As FWF data record the full return of the emitted pulse, it is possible to obtain more detailed information about the vertical structure of the vegetation (TOTH et al., 2011). However, there are some peculiarities in FWF data that make its use and processing difficult, such as the large volume of data compared to discrete return data, increased processing time, and lack of tools capable of visualizing and processing FWF data (MILTADOU et al., 2021; TOTH et al., 2011). Crespo-Peremarch and Ruiz, (2020) and Miltiadou et al. (2021) described the available software and waveform processing techniques.

Some approaches are used for waveform processing including peak detection from the digitized signal, correlation between emitted and received signals, deconvolution, and decomposition (MALLET et al., 2009; OLIVEIRA; CENTENO; HAINOSZ, 2018). Decomposition is the most used method for processing FWF data, and it consists of decomposing the waveform into a sum of peaks, which are identified by fitting parametric

functions (MALLET et al., 2009; WAGNER et al., 2006). The aim is to extract as many peaks as possible from the raw signal (Figure 1), as well as other information, such as intensity, amplitude, and pulse width (DUCIC et al., 2006; TOTH et al., 2011). The number of components and initial values are required and the parameters can be estimated using various methods, such as expectation-maximization that takes into account the maximum likelihood, and Levenberg-Marquardt that deals with the problem using the least squares approach (CHAUVE et al., 2007; MALLET et al., 2009; MALLET; BRETAR, 2009). In forestry studies, the waveform decomposition method is widely used to classify tree species and determine forestry variables, such as diameter, volume, above-ground biomass, and diversity-related variables (CAO et al., 2016b, 2016a; CHAUVE et al., 2007, 2008; REITBERGER et al., 2009; SHENDRYK et al., 2016; SUMNALL; HILL; HINSLEY, 2016; YAO; KRZYSZEK; HEURICH, 2012; ZHOU et al., 2018a; ZHOU; POPESCU, 2017).

Figure 1 – Representation of a modelled waveform. The light blue dots are the LiDAR pulses, and the dark blue line is the Gaussian fitted model. Adapted from Ducic et al. (2006).

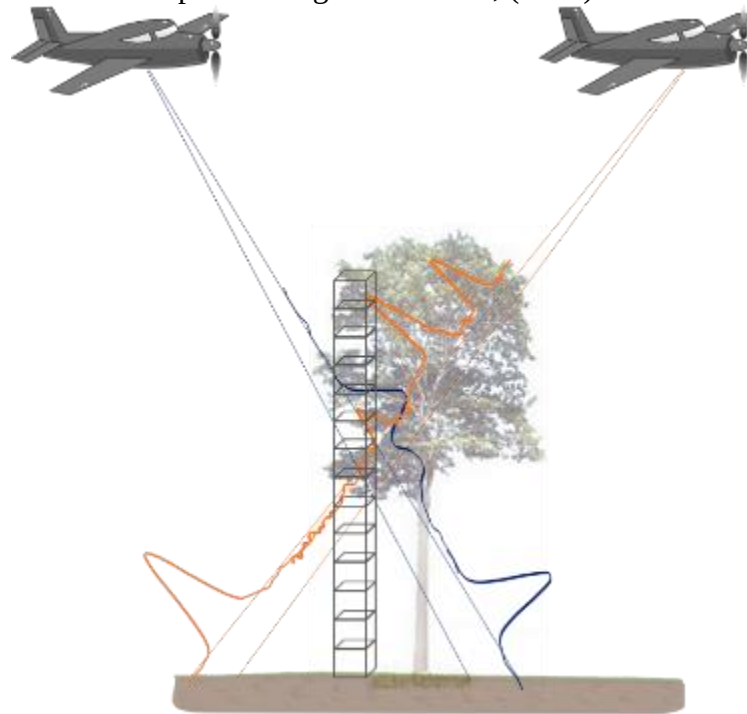


Another alternative to analyze the waveforms is to use a voxel-based approach, first developed by Persson et al. (2005). In this technique, the information contained in the waveform samples is entered into a three-dimensional grid, which accumulates the intensity values of multiple shots intercepting the voxel (volume element) (Figure 2). Depending on the method used, the value that the voxel assumes is the maximum intensity or the average intensity of the waveform samples crossing the voxel (HERMOSILLA et al., 2014; MILTIADOU et al., 2015; PERSSON et al., 2005). Hermosilla et al. (2014, 2013) developed a voxel-based approach to

analyze waveforms that correct the off-nadir effect due to different scan angles and non-vertical waveforms for different flight paths. Thus, pseudo-vertical waveforms are constructed using the maximum intensity value of the waveform sample intercepting each voxel, which synthesizes the information from the raw waveforms. From these pseudo-vertical waveforms, metrics are derived using Gaussian decomposition to extract peaks, the intensity of Gaussian components, and shape. In addition, metrics related to surface roughness, vertical distribution, height, and FWF energy statistics can be derived (CRESPO-PEREMARCH; RUIZ, 2020; HERMOSILLA et al., 2013).

Miltiadou et al. (2019, 2015) developed an open-source software called DASOS (used in this study and explained in section 3.5.1), which analysis the waveforms within the voxels while preserving information about the echo width. This is an advantage compared to the decomposition technique. In contrast to the previously mentioned approach, in this case, the average value of the intensities, coming from the multiple waveforms intercepting the voxel, are assigned to the voxel. Other information is extracted as well, such as voxel distribution, pitch, intensity, and whether or not the voxel has been filled with a waveform sample. Furthermore, the voxel approach allows for a more robust and complete representation, giving greater resistance to noisy signals (MILTIADOU et al., 2019), and this can be very useful for complex forests, such as tropical forests.

Figure 2 – LiDAR pulses emitted at different scanning angles intercepting the voxelized space. Wang and Glennie, (2015).

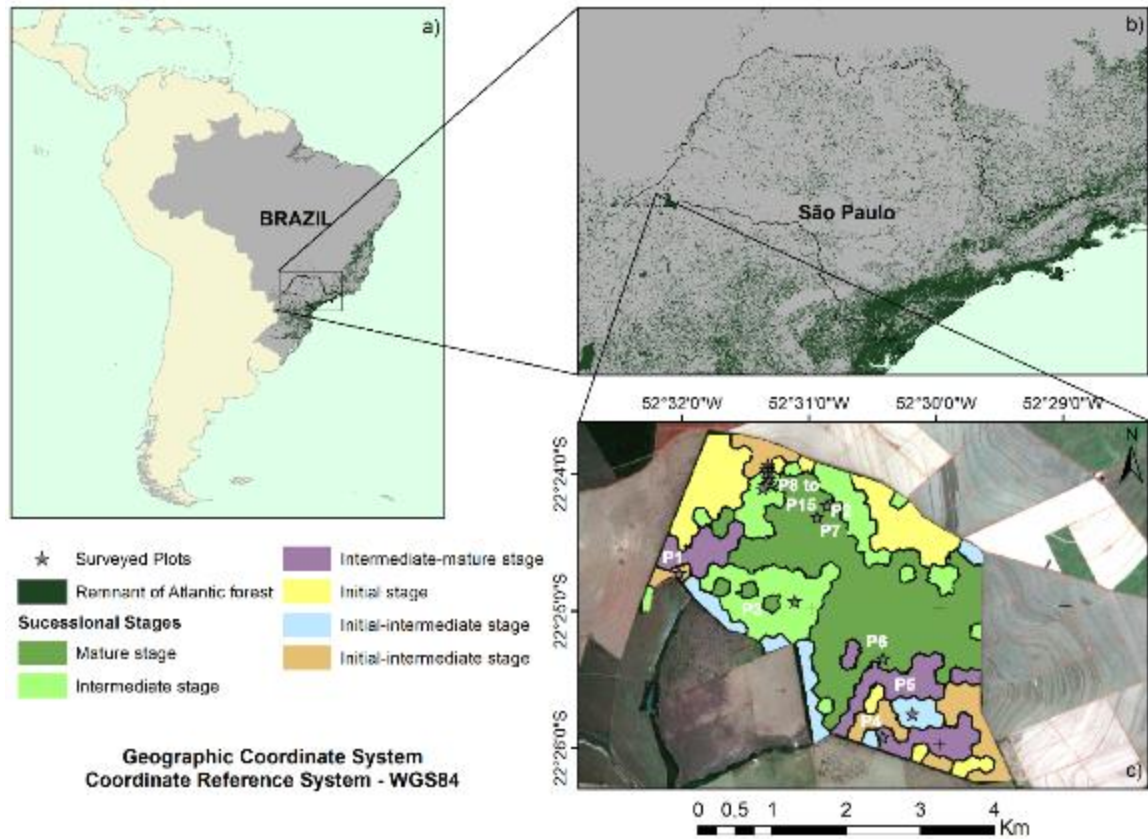


3 MATERIALS AND METHODS

3.1 STUDY AREA

The study area is a conservation unit, i.e., forest remnants protected by environmental laws, located in the interior of the state of São Paulo, Brazil (Figure 3). This is one of the few areas of the seasonal Atlantic Forest (*Floresta Estacional*, in Portuguese), also known as inland Atlantic Forest, considering that most of the conservation units are in coastal regions (Figure 3a). This fact makes it an area of fundamental ecological importance for biodiversity maintenance (MMA; IBAMA; ICMBIO, 2007). Of the four forest remnants composing the conservation unit, the *Ponte Branca* remnant was chosen for this study (Figure 3).

Figure 3 – Location of the study area. (a) Map of Brazil with emphasis on the state of São Paulo. (b) West of the state of São Paulo. (c) *Ponte Branca* forest remnant with the location of the surveyed plots in the different successional stages found. Adapted from Berveglieri et al. (2018); Martins-Neto et al. (2021).



The remnant is covered by natural tropical forest, featuring the initial to mature successional stages (Figure 3c) that had experienced fires and/or deforestation. The surrounding vegetation is composed of pasture and sugarcane plantations (BERVEGLIERI et al., 2016). The forest structure is diverse, varying according to the altitudinal gradient (alluvial in the southern part of *Ponte Branca* and submontane in the northern part). In the early stages of forest regeneration, palm trees like *Syagrus romanzoffiana* and Myrtaceae species of the genus *Eugenia* spp., are found in the understory. In the intermediate stage, there are some secondary species in the upper canopy like *Hymenaea courbaril*, *Cedrella fissilis*, and *Handroanthus* spp.; and in the mature stage; species like *Aspidosperma* spp., *Copaifera langsdorffii*, and *Cariniana estrellensis* dominate the upper canopy (BERVEGLIERI et al., 2016, 2018; MMA; IBAMA; ICMBIO, 2007).

3.2 FIELD PLOT AND VARIABLES ACQUISITION

Fifteen field plots were surveyed, and their locations were established to cover all the existing successional stages in the *Ponte Branca* remnant (Figure 3c), based on previous interpretations of aerial images. Two different plot designs were used: seven square plots with dimensions of 40 m × 40 m and eight rectangular plots of 80 m × 20 m dimension. A rectangular shape was chosen as it better encompasses the heterogeneity observed in Plots 8 to 15, which cover three successional stages (Figure 3c). The coordinates of the four corners in each plot were obtained by dual-frequency global navigation satellite system (GNSS) receivers, tracking for at least one hour to mitigate signal occlusion due to the dense tree canopy, thereby achieving an estimated positional accuracy of 50 cm.

The number of surveyed plots was considered sufficient based on the analysis of the species accumulation curve, in which a near stabilization was observed, suggesting that augmenting the number of samples would not significantly increase the number of inventoried species. For each plot, the following variables were obtained: (1) diameter at breast height (DBH) of all trees above 3.5 cm; (2) tree species, identified by a taxonomic specialist in the Atlantic forest tree species, based on dendrological aspects and using the APG IV classification system (CHASE et al., 2016); and (3) the number of trees. From these variables, the basal area (unit: m² ha⁻¹), number of tree species, and density (unit: trees ha⁻¹) were obtained. A statistical summary of the measured and obtained forest variables among the plots is presented in Table 1. A total of 3181 trees were observed in all plots.

Table 1 – Descriptive statistics of measured and derived forest variables.

Variables	Mean	Range	Standard deviation
DBH (cm)	10.3	3.5 – 97.4	± 8.0
Basal area (m ² ha ⁻¹),	16.2	5.6 – 30.7	± 7.5
Number of tree species	15	10 – 24	± 3.8
Density (trees ha ⁻¹)	1193	380 – 2286	± 569

3.3 ABOVEGROUND BIOMASS ESTIMATION

The plot-level above-ground biomass (AGB) was estimated through the Tropical Biomass & Carbon – TB&C App (<https://www.tropicalbiomass.com/>). The TB&C app implemented four biomass equations. These equations output biomass at stand level (in Mg ha⁻¹) and require plot basal area (BA in m² ha⁻¹) and density (N in trees ha⁻¹) as input variables.

The TB&C App's equations have specificity to the forest types in Brazil, including "Seasonal Forest" (forest type inherent to remnant *Ponte Branca*). The equation specific to "Seasonal Forest" is (Equation 1):

$$\widehat{AGB} = e^{[2.35855 + 1.18178 \ln(BA) - 0.19365 \ln(N)]} \quad (1)$$

The equations implemented in the TB&C App were calibrated with 72000 bootstrapped samples of plots stemming from 1428 trees with AGB destructively measured. The TB&C App also provides a confidence interval (CI) for the estimated plot AGB (i.e., $\widehat{AGB}$) based on a model related uncertainty approach. The analysis of reliability indicates that the probability of the true AGB to fit into the CI provided by the App was $\geq 95\%$ ($n = 72000$). The mean value of AGB found using this approach was 78.6 Mg.ha^{-1} with a confidence interval at a 95% probability level between 64.4 and 92.9 Mg.ha^{-1} . The estimated relative error was 20.4%.

3.4 LiDAR DATA SURVEY

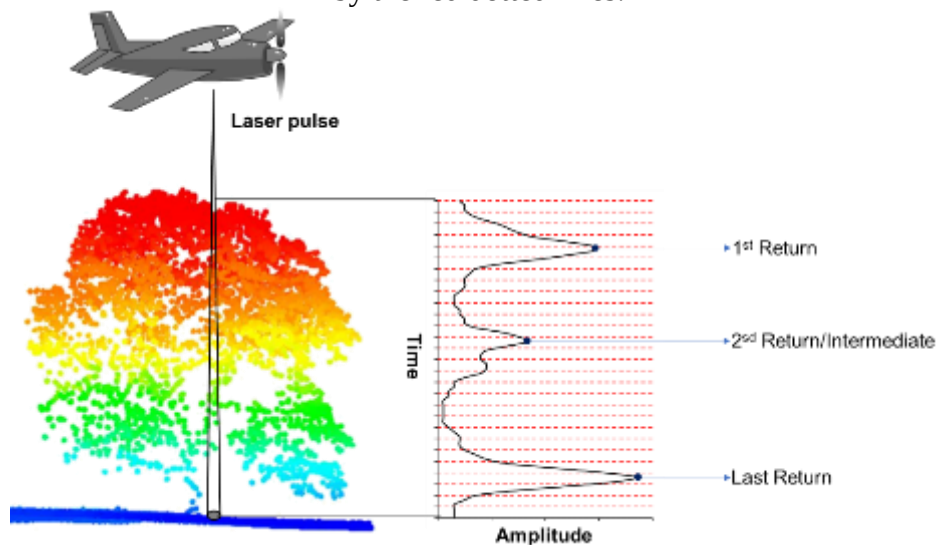
LiDAR data were acquired at two different flight heights: 900 and 2000 m. It was necessary to acquire 16 flight lines at a height of 900 m and three flight lines at a height of 2000 m to cover the entire *Ponte Branca* Forest remnant. The equipment used for data collection was a Riegl LMS-Q680i (Table 2), which is a long-range airborne laser scanner with digital full-waveform analysis and multiple-time-around (MTA) processing. MTA acquires echoes arriving after a delay of more than one pulse repetition interval, making measurements beyond the maximum unambiguous measurement range feasible (RIEGL, 2012). This enables a high repetition rate, which is fundamental for using this equipment in very complex areas, such as tropical forests (BERVEGLIERI et al., 2021; RIEGL, 2012).

Table 2 – Technical specifications of Riegl LMS-Q680i equipment (RIEGL, 2012).

Specification	Performance
Range (min/max)	30 m/3000 m @ reflectivity $\geq 60\%$
Precision	20 mm
Maximum pulse repetition rate	400,000 Hz
Wavelength	1550 nm
Intensity	16 bits
Scan angle range	$\pm 30^\circ$
Scan mechanism/pattern	Rotating polygon mirror/parallel scan lines
Beam divergence	≤ 0.5 mrad
Scan speed	10 – 200 lines/second
Dimensions (L x W x H)/weight	480 x 212 x 229 mm/17.5 kg

The data were delivered in two formats: (i) LAS 1.2 containing only the peak returns (PR) of the waveform data and (ii) LAS 1.3 with full-waveform data (FWF). In the LAS 1.2 file format, the returns are labelled up to number 5, but since the peaks are extracted from waveform data, there is no limitation on the number of peak points saved into the file (ASPRS, 2008). An additional file was included in LAS 1.3 with an extension *.wdp* (waveform data packets), which contains the stored pulse sections of the waveform in the vicinity of the declared returns (ASPRS, 2010) (Figure 4). The density was 16.4 points m^{-2} and 18.1 points m^{-2} in the PR and FWF data, respectively, for the 900 m flight height and for the 2000m flight height the density was 1.1 points m^{-2} and 1.2 points m^{-2} for PR and FWF data, respectively.

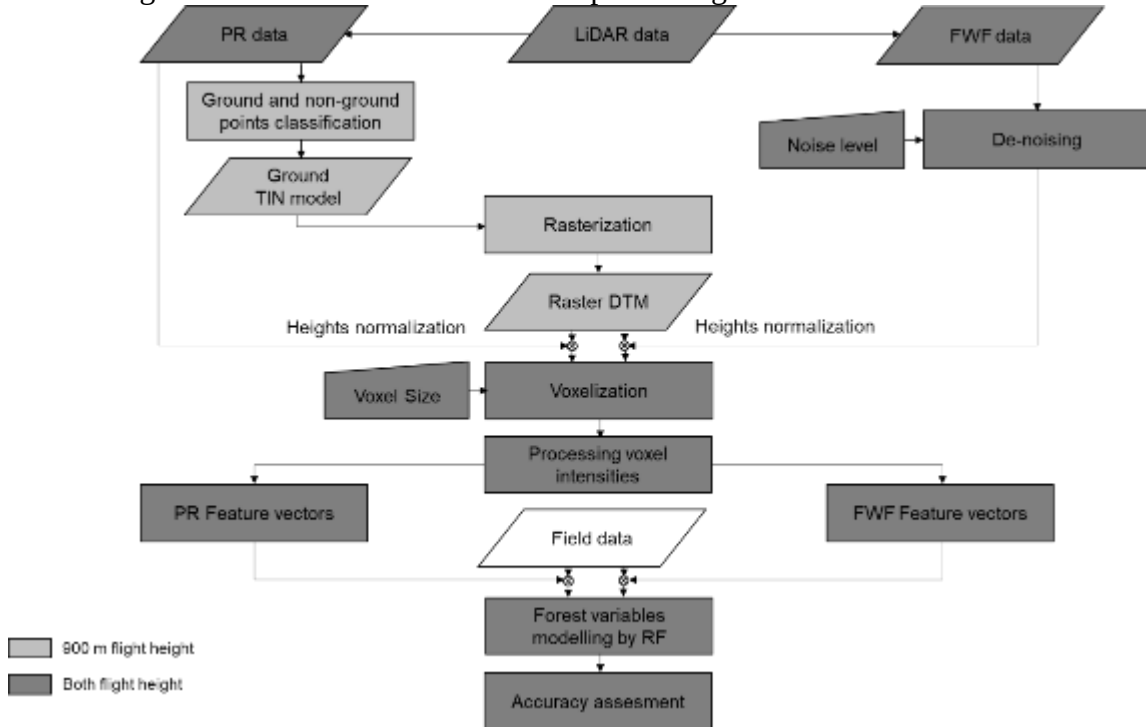
Figure 4 – Representation of LiDAR datatypes. The blue dots are the discrete peak returns peaks, and the solid black line is the FWF data that is digitized at intervals of 1 ns as denoted by the red dotted lines.



3.5 LiDAR DATA PROCESSING

LiDAR data were processed using two different tracks. The first track used the PR point cloud to generate a digital terrain model (DTM), while the second track used the data with the two data types (PR and FWF) to voxelize and extract the feature vectors containing the intensity values in each voxel (Figure 5).

Figure 5 – Flowchart of LiDAR data processing to extract the forest variables.



Using the PR point cloud, a DTM was produced to normalize the heights. For this task, the data acquired at flight height of 900 m were utilized owing to the higher pulse density per square meter (average of 16.4 points.m⁻²), which allows a higher number of pulses to penetrate through the canopy and collect information about the ground.

First, we classified the point cloud into the ground and non-ground points. This step was performed using software *LAStools* (ISENBURG, 2014), in which the parameters for classification were empirically tested until the vegetation points (non-ground points) and ground points were well separated. In the R environment (R CORE TEAM, 2017) using the *lidR* package (ROUSSEL et al., 2018) a surface was modelled by a triangular irregular network (TIN) using the ground points, and then rasterized to produce a raster DTM with a resolution of 0.50 m.

3.5.1 Voxelization

Voxelization was performed using the open-source software DASOS (MULTIADOU et al., 2019), which enables the analysis, accumulation, and visualization of FWF LiDAR data (<https://github.com/Art-n-MathS/DASOS>). Based on Persson et al. (2005), the waveform samples from multiple pulses with different intensities are accumulated in a three-dimensional discrete density array of voxels. DASOS processes both FWF and PR data, but in the latter, the peak returns are voxelized instead of the waveform samples.

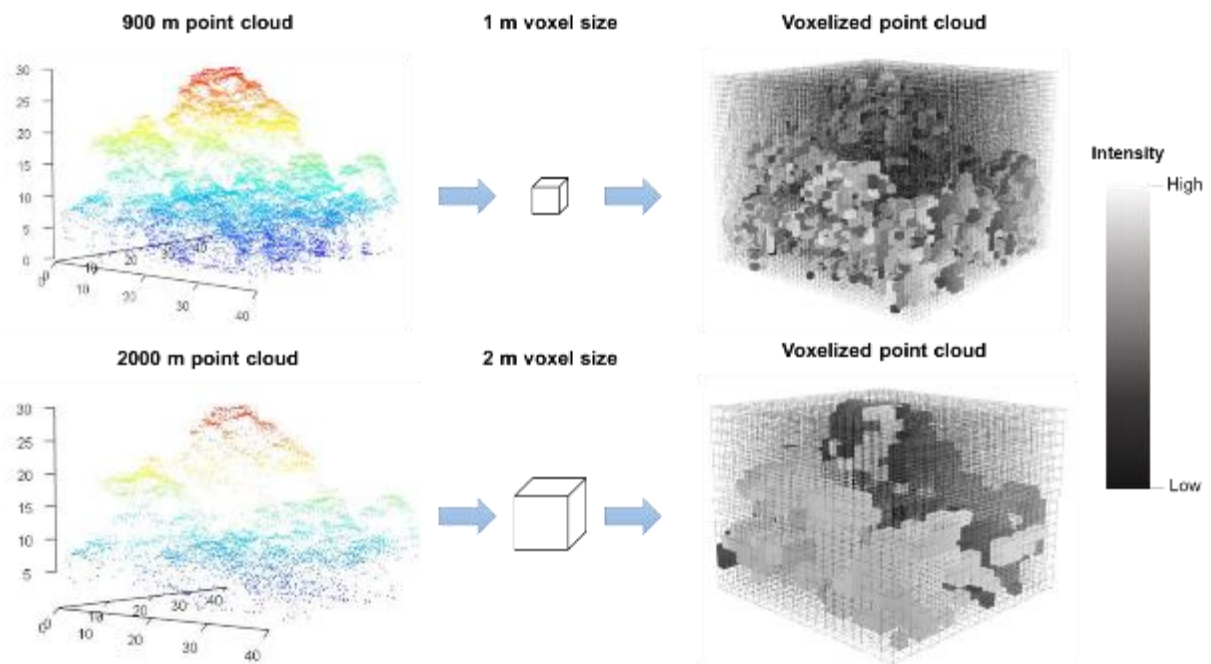
Waveform data contain more information than PR data, but they also contain much more noise. Sometimes, noise comes from the laser returned to the system after being scattered a few times. In this case, the signal is low, but at the same time, the estimated location of the target is incorrect because the time elapsed to reach the sensor is longer than a round-trip time. Low-level filtering was applied to remove these types of noises. The flight lines covering the surveyed plots were used as the input data. Different filtering thresholds based on pulse amplitudes were tested to verify how different noise thresholds can modify the estimates results. These include the following: no filtering (noise level 0 – nl_0), noise level 2 (nl_2), noise level 5 (nl_5), noise level 7 (nl_7), and noise level 10 (nl_10). Pulse signals below the selected amplitude thresholds were eliminated. Regarding the PR data, as only the peaks of each pulse are selected, the application of low-level filtering is not relevant.

Once the FWF data were filtered, voxelization was performed. To overcome some of the current limitations of DASOS relating to memory consumption and reducing processing time, each test was conducted using small selected areas containing the field plots. During voxelization, the position of each waveform sample was subtracted from the DTM (produced in R, as explained above). Therefore, the obtained set of voxels lies on flat terrain, and the tree height values are rounded in intervals corresponding to the voxel size. For the PR data, the height of each peak return is subtracted from the corresponding height of the interpolated DTM. Therefore, both sets of voxelized data lie on flat terrain.

The voxel size was defined based on the footprint and density (points m^{-2}) of the acquired LiDAR data at flight heights of 900 m and 2000 m and the representation of the trees in the voxels. Choosing the appropriate voxel size for data analysis depends on the density of the point cloud and the level of abstraction that is desired to be reached in each voxel. The selection of voxel size influences the accuracy of forest parameter estimates (CRESPO-PEREMARCH et al., 2018b; WANG et al., 2020). A small voxel size results in many empty

voxels because of the gaps between the LiDAR pulses; the smaller the size of the voxel, the more voxels lie within those gaps. When the size of the voxel increases, the number of empty voxels is reduced, thus filling the gaps between the pulses (CRESPO-PEREMARCH et al., 2018a). Consequently, the waveform samples from the smaller trees may be merged with the nearby samples. This may affect the modelling of variables that depend on the individual trees in the plot, for example, the variables density and number of tree species. In this study, it was decided that the appropriate voxel sizes used for the 900 m flight height and the 2000 m flight height were $1 \times 1 \times 1$ m and $2 \times 2 \times 2$ m, respectively (Figure 6). Voxels were visualized in an R environment (R CORE TEAM, 2017) using the voxR package (LECIGNE; DELAGRANGE; MESSIER, 2018).

Figure 6 – Point clouds at different heights of flights and division of the area into voxels of equal size and voxels with different intensities.



As LiDAR pulses are emitted at different scan angles and as multiple pulses can pass through the same voxel, it is necessary to normalize the values stored in the voxels to make them comparable to each other. A waveform sample is recorded as an intensity value and a point location, but it represents the pulse width of a time interval of 1 ns. Thus, normalization is performed so that the intensity of the equal pulse width corresponds to each voxel. It should be noted that the performed normalization is not equivalent to calibration, since the intensity decreases while traveling from the plane until it reaches a target, and occlusions are not taken into consideration. Initially, DASOS iterates through all the waveform samples and inserts each

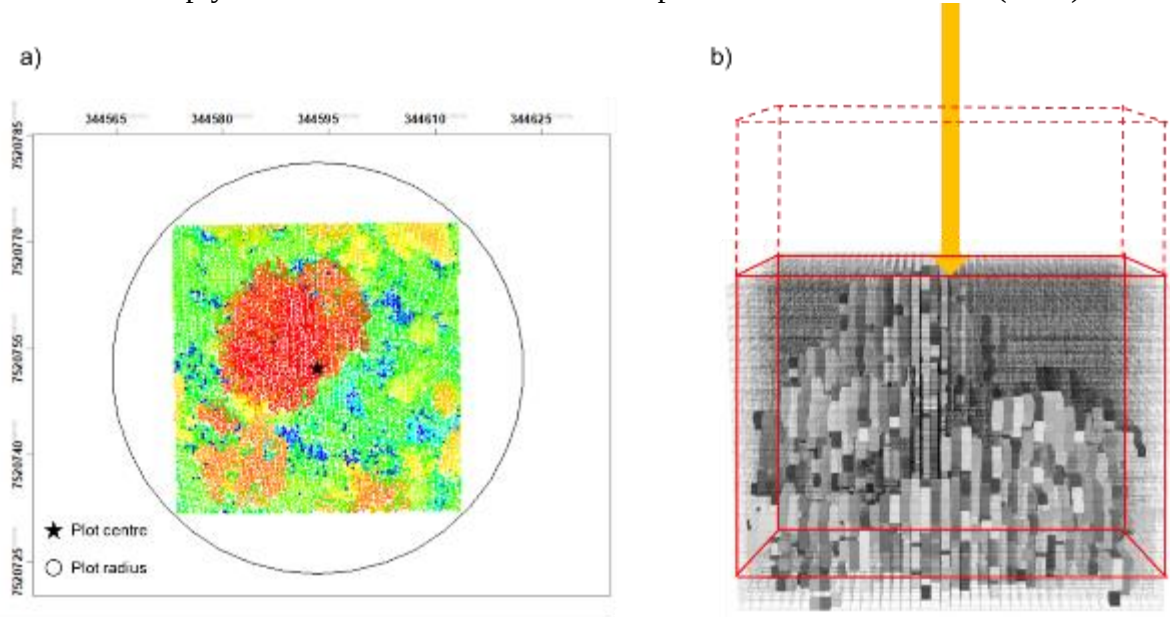
waveform sample into the voxel in which it lies inside. During normalization, a single intensity value was assigned to each voxel. This single intensity value is equal to the average intensity of the intensities of the waveform samples inserted inside a voxel. As a result, the uneven footprints of the scanning patterns were resolved (MULTIADOU et al., 2019, 2020). This methodology differs from that of Persson et al. (2005), which selected the waveform with the highest intensity inside the voxel. In contrast, while interpreting peak returns, the pulse width could vary depending on the interpretation because each Gaussian curve has different pulse widths. Therefore, when discrete data are voxelized, normalization is performed, but it does not guarantee equal pulse width per voxel.

3.5.2 Extraction of feature vectors

Another functionality existing in the open-source software DASOS is the option to extract attributes from 3D-windows (cylindrical or cuboid) placed within the voxel space and exporting these attributes into vectors. This feature was initially developed to characterize individual trees within the three-dimensional space for dead-standing tree detection (MULTIADOU et al., 2018). In our case, instead of extracting attribute vectors for characterizing individual trees, we extracted attribute vectors to characterize the entire surveyed plots. To restrict the 3D window only over the plot, the radius of coverage from the centre of the plot was defined (Figure 7a) to characterize an entire plot.

A cylindrical or cuboid 3D-window, with a size set by the user, is defined with the column of interest at the centre of the plot. This window moves downward from the top of the voxelized space until it reaches the first non-empty voxel, that is, the voxel that contains an accumulated intensity value derived either from waveform samples or peak returns. Information is extracted from the area within the 3D-window and exported as an attribute vector in .csv format (MULTIADOU et al., 2018, 2020). This vector is used as an input for modelling forest variables and characterizes the plots in 3D space. As the surveyed plots in this study have quadrangular shapes, cuboid geometry for the 3D-window was used for extracting attribute vectors (Figure 7b).

Figure 7 – (a) Point cloud representation of a plot with an established centre and a radius that contains the plot area. (b) Plot after the voxelization process and feature vector extraction from a cuboid three-dimensional window, with the movement of the top window to the first non-empty voxel of the middle column. Adapted from Miltiadou et al. (2020).



It is possible to extract vectors in two ways: (1) as raw data, which correspond to the intensity value of the voxels inside the three-dimensional window, and (2) in a processed form. The processed intensities provide a vector with 27 attributes as statistics about the voxels, their intensities, and voxel distribution in the 3D window (MILTIADOU et al., 2018, 2019). However, as the pulse intensities were not previously calibrated, their use is not recommended as it is not possible to make any comparisons with other data. Therefore, we removed the attributes related to the pulse intensity. We used 19 values from the set of 27 attributes extracted by the DASOS (Table 3). Attribute vectors were extracted for all surveyed plots corresponding to the four types of data obtained, i.e., PR and FWF at flight heights of 900 m and 2000 m each.

Table 3 – Attributes extracted in DASOS.

Feature Name	Description
Height_Middle_Column	The height of the middle column of the cuboid.
Height_Mean	Mean height of all the columns inside the 3D-window.
Height_Median	Median height of all the columns inside the 3D window.
Height_Std	Standard Deviation of the column heights inside the 3D-window.
Dis_Mean	Mean distance from the central voxel to non-empty voxels.
Dis_Median	Median distance from the central voxel to non-empty voxels.
Dis_Std	The Standard Deviation of the distances between the central voxel to non-empty voxels.
Top_Patch_Len_Middle_Col	The length of the top patch ¹ of the middle column of the cuboid.
Top_Patch_Len_Mean	The Mean length of all the top patches.
Top_Patch_Len_Median	The Median length of all the top patches.
Top_Patch_Len_Std	The Standard Deviation of all the top patches.
Mirror_Diff_X_Mean	The Mean Mirror Difference of the voxel intensities with the middle column of the x-axis as the axis of symmetry.
Mirror_Diff_X_Median	The Median Mirror Difference of the voxel intensities with the middle column of the x-axis as the axis of symmetry.
Mirror_Diff_X_Std	The Standard Deviation Mirror Difference of the voxel intensities with the middle column of the x-axis as the symmetric axis.
Mirror_Diff_Y_Mean	The Mean Mirror Difference of the voxel intensities with the middle column of the y-axis as the axis of symmetry.
Mirror_Diff_Y_Median	The Median Mirror Difference of the voxel intensities with the middle column of the y-axis as the axis of symmetry.
Mirror_Diff_Y_Std	The Standard Deviation Mirror Difference of the voxel intensities with the middle column of the y-axis as the axis of symmetry.
Mirror_Diff_Z_Mean	The Mean Mirror Difference of the voxel intensities with the middle column of the z-axis as the axis of symmetry.
Mirror_Diff_Z_Median	The Median Mirror Difference of the voxel intensities with the middle column of the z-axis as the axis of symmetry.

¹ The top patch is defined as the number of non-empty adjacent voxels starting from the top of the column (Miltiadou et al., 2021, 2019)

3.6 MODELING FOREST VARIABLES

The four forest variables (basal area, density, number of tree species, and AGB) were estimated using the attribute vectors extracted from the PR and FWF LiDAR data. The chosen modelling technique was random forest (RF), based on multiple decision trees constructed in an uncorrelated manner. Thereafter, the average of all predictions was calculated in the regression cases to estimate the variables (BREIMAN, 2001; FRIEDMAN; TIBSHIRANI; HASTIE, 2015). Decision trees are composed of different features, and the same feature can appear in more than one tree (BELGIU; DRĂGUȚ, 2016). There are several advantages of RF

for regression problems include its high ability to model complex structures, calculation of measures of variable importance, lack of assumptions about data distribution, and low sensitivity to outliers, even when increasing the number of decision trees and when using noisy data (BELGIU; DRĂGUȚ, 2016; BREIMAN, 2001; PAL, 2005; PRASAD; IVERSON; LIAW, 2006).

Prior to RF modelling, the feature vectors were normalized in the range [-1, 1] owing to the scale differences for each feature. The RF was executed with the R package *randomForest* (LIAW; WIENER, 2002), and two parameters were defined: (i) *ntree*, the number of trees to be created from the selection of samples by the bootstrap method, with a value of 1000; and (ii) *mtry*, which refers to the number of predictors to be randomly sampled in each of the tree nodes. In regression problems, the *mtry* value is one-third of the input data. As 19 feature vectors were extracted from each plot, the set *mtry* value was six.

3.7 MODEL GOODNESS-OF-FIT

The estimation of the four forest variables using the feature vectors with different LiDAR return types and flight height scenarios was done using leave-one-out cross-validation (LOOCV), in which the regression model was trained using all samples except one, and the prediction error was obtained from the average accuracy of each interaction with the remaining sample (BROVELLI et al., 2008; MARTINS-NETO et al., 2021). The accuracy of the adjusted models and the verification of the differences between the return type and the flight height in the estimation of forest variables were assessed by statistics of root-mean-square error (RMSE) (Equation 2), relative RMSE (rRMSE) (Equation 3) and relative bias (rBias) (Equation 4).

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$rRMSE(\%) = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}}{\frac{\sum y_i}{n}} \cdot 100 \quad (3)$$

$$rBias\% = \frac{\sum (\hat{y}_i - y_i)}{\sum y_i} \cdot 100 \quad (4)$$

where y_i is the value measured in the field (observed), $\hat{y}_i$ is the estimated value, and n is the number of samples.

4 RESULTS

4.1 SELECTING NOISE LEVEL THRESHOLD

Different noise level filtering thresholds were tested for the FWF data, and the results were further compared to the PR data. The thresholds with filtering levels chosen to remove noise were noise level nl_0, nl_2, nl_5, nl_7, and nl_10 for the two flight heights. After filtering the data with different thresholds, maps were produced by storing thickness information (Figure 8), which is the distance between the first and the last voxel with waveform samples or peak returns (not empty) in each column of the 3D volume (MILTADOU et al., 2019). The ground sample distance (GSD) of these maps were 1 m and 2 m for 2000 m flight height, for the scenarios mentioned above (Figure 9). Furthermore, a waveform sample with different thresholds of filtering levels is depicted in Figure 9 to understand how the waveform behaves at each adopted filtering threshold.

Figure 8 – Pixel value computed from the thickness information in a column of voxels.
Adapted from Miltiadou et al., 2019.

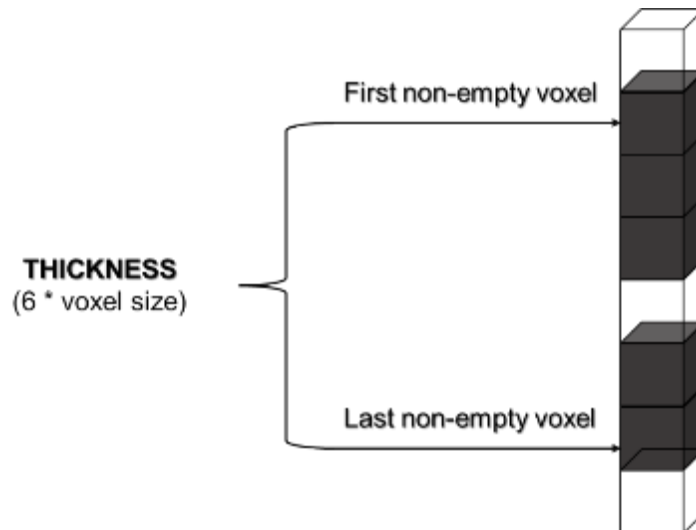
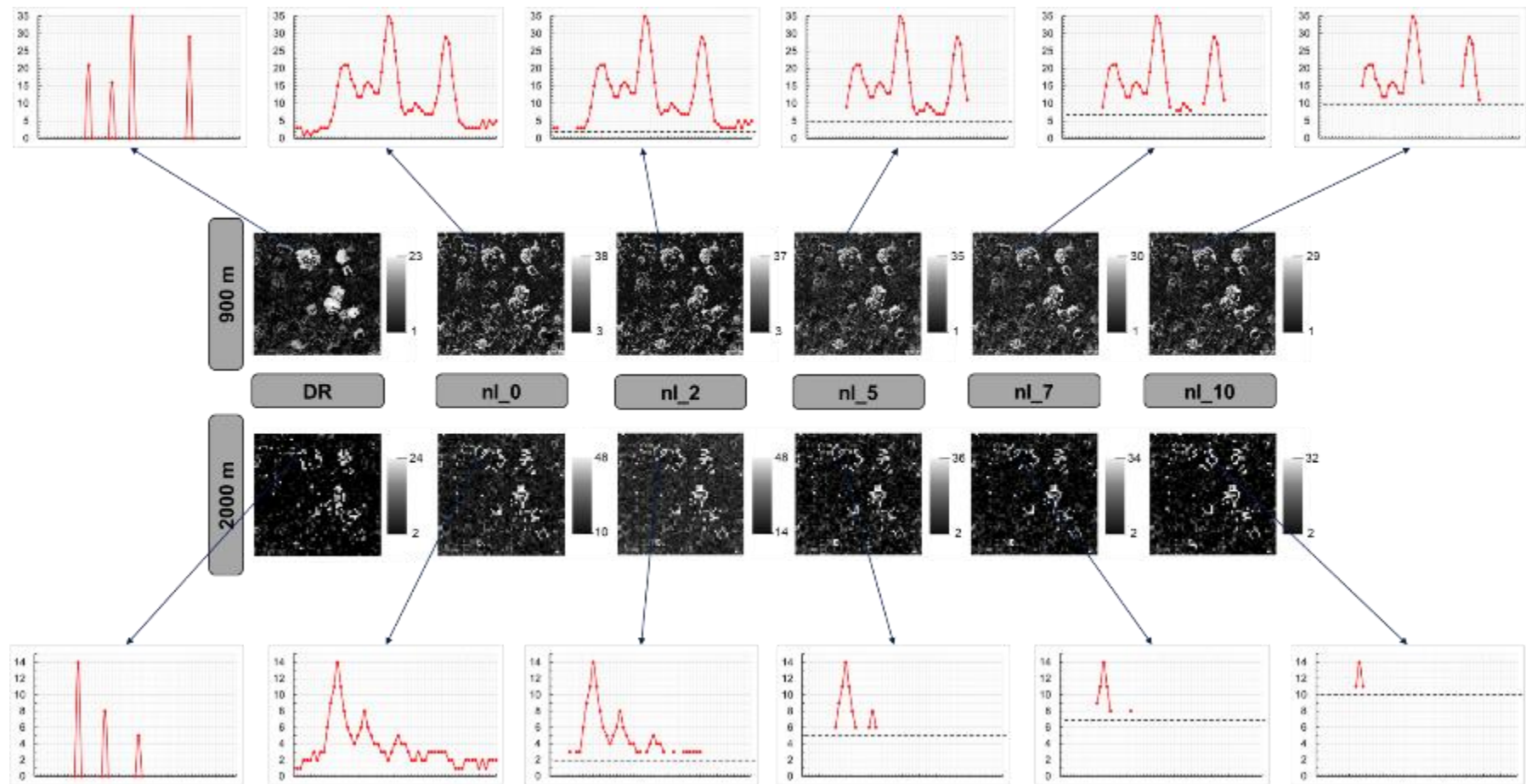


Figure 9 – Maps containing the distance between the first and the last non-empty voxel for PR and FWF data with different noise filtering thresholds at flight heights of 900 m and 2000 m. Details of waveform samples at different thresholds and flight heights are also shown.



By comparing the generated maps with different returns and thresholds (Figure 9), it was verified that the distances between the first and the last non-empty voxels were smaller for the voxelized PR data than the FWF data. This is because there are fewer voxels containing information, as only the peak returns are assigned to the voxels. This occurs because when filling voxels with discrete returns, there is a smaller probability of a voxel containing a peak return than a waveform sample. As the low-level noise threshold increases, more waveform samples are discarded, and the FWF becomes similar to the PR. Simultaneously, the possibility of empty voxels increases (Figure 9). Notably, the noise threshold is site- and sensor-dependent, which is justified by the comparison of height flights (900 m and 2000 m). As shown in Figure 9, more waveform samples were discarded from 2000 m flights than from 900 m flights. Therefore, it is important to identify the appropriate threshold for each site and sensor to avoid discarding too much information or not reducing enough noise.

Regarding the flight height, the lower the flight height, the greater the points density per square meter and the smaller the footprint size. When the pulse density decreased, the number of empty voxels increased. About the waveform data, even though the emitted pulse has a conical shape, the current voxelization method of DASOS treats each waveform sample as a point, and it does not interpolate the intensity of the conical shape that corresponds to each waveform within the voxelized space. As a result, once the pulse density decreased, the number of empty voxels increased in both the PR and FWF data. To mitigate this effect, a larger voxel size was used to voxelize the data acquired from the 2000 m flight height. The shortest distance between the first and last voxels, containing waveform samples, is lower for a flight height of 900 m, because more voxels store waveform samples. For the map generated with data coming from the 2000 m flight, greater distances between filled voxels were observed, because there are more empty spaces between the first and the last voxels containing waveform samples in the entire voxel column (see the distance between the highest and lowest thickness values for the same noise level threshold at different flight heights in Figure 9). However, despite the differences between the maximum and minimum values, it was observed that the ranges at different flight heights were very similar for the same noise level threshold.

4.2 FOREST VARIABLES ESTIMATION

Table 4 presents the results of the statistics from the estimation of forest variables through the RF method, using as input data the vector of attributes extracted with software DASOS for the different return types and flight height scenarios. The RMSE values for the estimated variables were very similar for different return types, thresholds, and flight heights (Table 4). The differences between the highest and lowest RMSE values for the estimations were, on average, 3.0% and 4.4%, respectively, for flight heights of 900 m and 2000 m. However, the results for data from 2000 m flight height showed a slightly higher RMSE than those for the data from 900 m flight height. Regarding the bias values, all estimates showed good distribution of errors, with values ranging between -2.0% and 2.0%, although for the data resulting from the highest flight (2000 m), there was a slight tendency to underestimate the variables, in comparison with data from the lowest flight (900 m).

For the variables basal area, density, and AGB, estimated with FWF data coming from the flight height of 900 m, it is possible to see a trend: the lower the filtering threshold, the higher the RMSE value, especially the rBias, until the threshold nl_7, indicating that these variables are influenced by noise in the FWF signal. This trend was not observed for the number of tree species, and the best result was obtained with threshold nl_2, as this variable has great variability in tropical forests where modelling the number of tree species is difficult. We can infer that, most probably in a lower flight, low-intensity signals are important to discriminate the number of species in the plots.

The results obtained with FWF data from the highest flight (2000 m) confirmed for all the estimated variables, the lowest RMSE values and the rBias values close to zero were obtained with FWF data with a threshold nl_2. Because of the lower point density at this flight height (1.2 points.m⁻²), less noisy signals were recorded by the sensor, and a lower filter threshold was required.

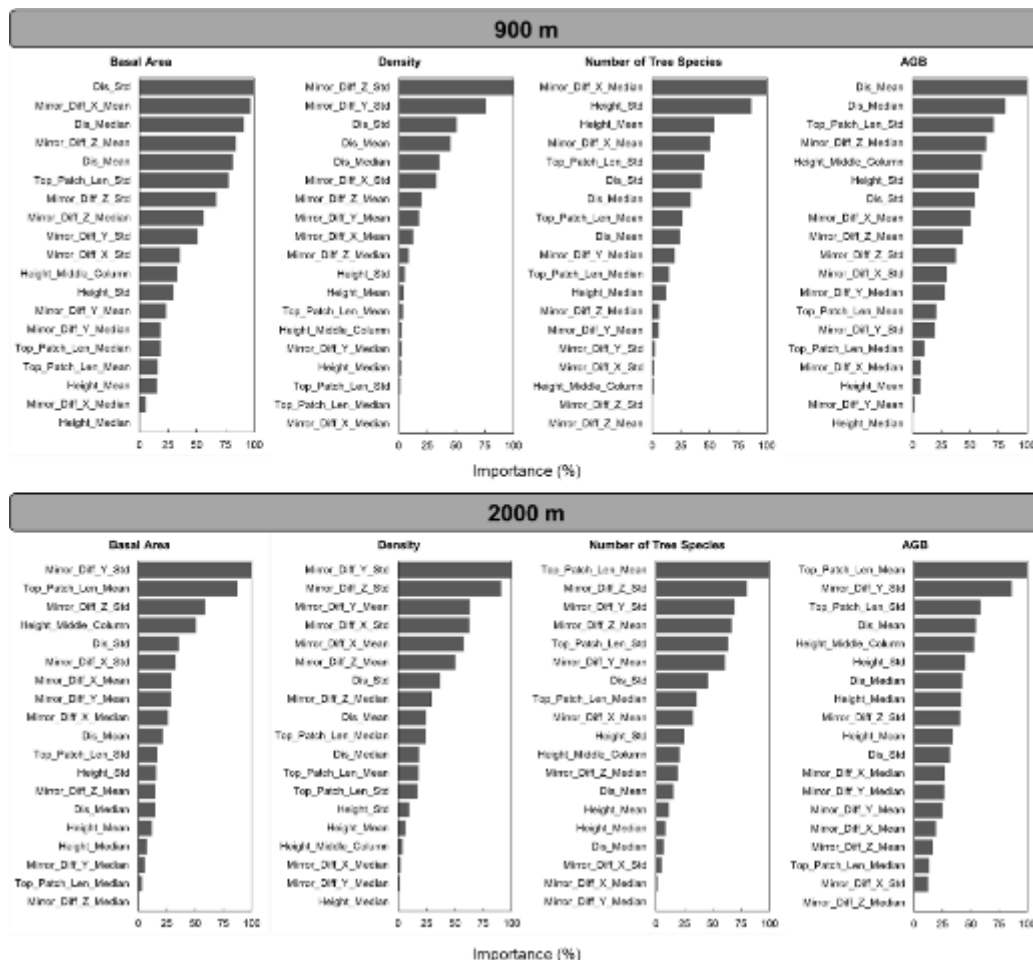
Table 4 – Accuracy assessment of forest variables estimated with PR and FWF data from different flight heights. The underlined values are those with the smallest bias for each variable estimated in each of the tested scenarios.

Variables	900 m																	
	PR			FWF_nl_0			FWF_nl_2			FWF_nl_5			FWF_nl_7			FWF_nl_10		
	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias
Basal area	2.9	17.8	-2.004	3.3	20.3	-1.523	3.3	20.1	-1.042	3.0	18.7	0.455	3.2	19.5	<u>0.354</u>	3.0	18.7	-1.118
Density	198.2	16.6	-0.207	216.6	18.2	1.186	233.6	19.6	0.699	224.9	18.8	0.346	213.6	17.9	<u>0.153</u>	219.1	18.4	-0.600
Number of tree species	1.7	12.9	-1.031	1.6	12.2	1.015	1.6	11.1	<u>0.163</u>	2.0	13.5	1.363	1.9	13.2	0.891	2.0	13.6	1.110
AGB	15.9	20.3	-3.048	16.9	21.5	-1.455	18.0	22.8	-1.320	16.3	20.8	0.874	16.6	21.1	<u>0.500</u>	15.9	20.2	-0.517
Variables	2000 m																	
	PR			FWF_nl_0			FWF_nl_2			FWF_nl_5			FWF_nl_7			FWF_nl_10		
	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias	RMSE	rRMSE	rBias
Basal area	3.2	20.1	-0.458	3.6	22.7	-0.801	3.4	21.3	<u>-0.072</u>	3.0	19.3	-0.411	3.0	19.3	-0.301	2.9	18.7	-1.108
Density	249.7	21.0	0.584	181.8	15.3	-0.511	159.7	13.4	<u>-0.160</u>	221.2	18.6	0.602	221.8	18.7	0.369	228.0	19.2	0.396
Number of tree species	1.9	12.7	1.008	1.9	13.0	1.738	1.9	12.3	<u>0.753</u>	2.1	14.3	2.043	2.0	13.0	0.885	1.8	12.2	0.846
AGB	17.0	22.2	-0.399	18.3	23.9	-1.964	19.0	24.8	<u>-0.394</u>	16.3	21.3	-1.121	15.8	20.7	-1.078	16.4	21.4	-1.957

A good estimate must lead not only to a low RMSE but also to a bias close to zero, with no overestimation or underestimation trends. Thus, the chosen scenarios at different flight heights were based not only on the lowest RMSE value, but also on the unbiased estimates. The best estimates are listed in Table 4.

Using the FWF data acquired at a 900 m flight height, the best estimations with bias close to zero for the basal area, density, and AGB variables were obtained at threshold nl_7 (Table 4). It is worth noting that these three variables are correlated, since AGB is strongly influenced by DBH and consequently by the basal area, while density is also influenced by the basal area. Therefore, the relationship between these variables was maintained. The variable number of tree species had the best result with FWF data at threshold nl_2. The best estimations of the four variables using FWF data acquired at a flight height of 2000 m were obtained with the lowest threshold, i.e., nl_2. The relative importance of the attributes selected by the RF for the best scenarios in the estimates of forest variables are shown in Figure 10.

Figure 10 – Percentage of importance of the attributes extracted by DASOS in the best scenarios for the estimation of the four forest variables.



Considering the two attributes with the greatest relative importance in the most favourable scenario, the “Mirror_Diff” attribute had great importance for all variables estimated in both flight heights, except for the AGB with data from the 900 m flight. The most important attributes extracted from the lowest flight (900 m) LiDAR data were as follows: the mean and median of “Mirror_Diff” on the X-axis; the standard deviation of “Mirror_Diff” on the Y and Z axes; the standard deviation of voxel column heights; and the mean and median of the distance from the central voxel to the non-empty voxel. For the LiDAR data generated with the highest flight (2000 m), the most important attributes were the standard deviation of “Mirror_Diff” in the Y and Z axes and the mean of “Top_Patch_Len” (Figure 10).

The “Mirror_Diff_X” and “Mirror_Diff_Y” statistics are site-specific, as they divide the voxelized space into two spaces by finding the average value of the X and Y axes and comparing these two spaces. If the two sides are very similar, the correlation will be high, and the mirror difference will be low. However, if one side has taller and dominant trees and the other has shorter trees with the presence of an understory, this attribute can be important because the mirror difference is high. Thus, it is possible to infer that due to the great heterogeneity of our studied forest, the mirror difference in the X and Y axes was high, making this attribute important for the estimation of the variable.

“Mirror_Diff_Z” statistics is meaningful, since it divides the space horizontally and gives values to compare different layers of the forest, like foliage and tree stems. This attribute was mainly important for the density variable (100% of relative importance for the 900 m flight and 91% for the 2000 m flight), because density is correlated with the number of tree stems in a given space, and the mirror difference in the horizontal space provides this information. In addition, this attribute was also important for estimating the number of tree species with data from the 2000 m flight, also demonstrating the ability of this attribute to differentiate this variable.

The “Height_Std” variable is dependent on the trees’ heights, which have high variability in a tropical forest. This variability was important to differentiate the tree species, with a relative importance of 87% in case of the 900 m flight.

The attributes “Dis_Mean” and “Dis_Median” (Table 3) were important for the estimation of basal area and AGB variables, with data from the 900 m flight height. As mentioned in Section 4.1, lower flight heights produce data with higher point density and more voxels with waveform samples. Consequently, the distances between the central voxel and non-empty voxels were smaller. These smaller distances and more voxels with waveform samples

can be correlated with the structure of the middle stratum of the forest (i.e., tree diameters), and these are very important for basal area and AGB estimations.

For the data collected with a 2000 m flight height, the attribute "Top_Patch_Len_Mean" was important for three of the four estimated variables (basal area, number of tree species, and AGB). This attribute can provide information about the composition of the top tree canopy. Tree crowns have a high correlation with diameter, which is very important when determining basal area and AGB, and each tree species has a different composition and distribution of leaves, so this attribute was also important for separating the number of tree species.

Considering that the data from the highest flight had a lower point density, there were more empty voxels below the canopy. The Top_Patch_Len_Mean attribute provides information from the top, where there is a higher probability of having filled voxels; thus, it provides better information for estimating the forest variables.

5 DISCUSSION

In this study, we tested a voxel-based approach to extract attributes from peak points extracted from waveform data, using all the waveform samples of the full waveform LiDAR data. The data were collected at two different flight heights (900 and 2000 m). The waveform samples were filtered and tested with different low-level noise-filtering thresholds. One of the fundamental steps in processing FWF data is noise removal, also known as de-noising. The sensor records low-energy signals that are not related to the reflection of the targets, and these signals have a negative influence on FWF data applications (LAI; ZHENG, 2015). The most commonly used method for de-noising (LEFSKY et al., 2005), was initially used for large footprint FWF sensors, whose threshold is defined from the mean of the waveform amplitude plus four times the standard deviation. Then, the remaining noise was removed using a Gaussian filter (HERMOSILLA et al., 2014).

We defined the filtering thresholds empirically by analysing a sample of waveforms (Figure 9) to understand how low-intensity signals behave in a tropical forest. It is worth highlighting that low-intensity signals may contain important structural information about the understory because in tropical forests there is a large presence of vines, shrubs, and branches as part of the forest structure. In this case, the threshold must be selected carefully to discard enough noise while preserving the crucial understory structure of the forest. To the best of our

knowledge, this is the first study to investigate how different waveform low-level filtering thresholds affect the modelling of forest variables.

For the data acquired at the 900 m flight height, the threshold nl_7 leads to the best estimations for three of the four forest variables (basal area, density, and AGB) and the value of the threshold approximates the mean of the waveform amplitudes. Figure 9 shows that this filtering threshold maintains few low-intensity signals; nevertheless, these signals are relevant in the modelling. For the data generated from the highest flight (2000 m), the threshold nl_2 leads to the best estimations for the four forest variables in the study. It approximates the mean of the waveform amplitudes at this flight height. These values differ from those found in the literature, since the threshold developed by Lefsky et al. (2005) (mean plus four times the standard deviation) is greater than the filtering threshold, which provided the best results in our study and which was the mean value of the amplitudes of the waveform samples.

At threshold nl_2, more waveform samples are discarded from the data acquired at a flight height of 2000 m compared to that acquired at 900 m. This may be due to the lower point density, larger footprint size, and larger distance between the target and receiver for the data acquired at the higher flight (2000 m). Therefore, lower flights in dense tropical forests may contain more low-level noise. Therefore, a higher low-level threshold for filtering is recommended, and vice versa.

Additionally, it was observed that the RMSE and rBias values (Table 4) obtained using data acquired at a flight height of 900 m, after removing low-level noise with a noise threshold of 10 (nl_10), are similar to PR data. Regarding the results with data acquired at a height of 2000 m, noise thresholds with a high filtering level can remove some high-amplitude peaks. This occurs because as the filtering level threshold increases, the selected amplitude value increases also, and more pulses are removed (Figure 9), until a point is reached where only a few waveform samples remain and most of them are close to the peak returns. Miltiadou et al. (2019) showed the effect of the noise level threshold on waveforms that were inserted into the voxel space. At a certain point, the chosen threshold is greater than the pulse amplitude recorded by the sensor; in this case, no waveforms were inserted into the voxel. This effect is more pronounced at the highest flight height, wherein there is a tendency to wrongly characterize the waveform when using a larger threshold. This is because they also eliminate peaks that have high return intensity, as seen when a threshold nl_10 was used for the 2000 m flight (Figure 9).

Other studies using FWF data for characterization and extraction of forest variables used several different flight heights: an average flight height of 710 m was used by Crespo-

Peremarch et al. (2018); Nie et al. (2018) used LiDAR data obtained from a flight height of 800 m; Hermosilla et al. (2013); Cao et al. (2014) and Shen et al. (2018) used a 900 m flight height; Sumnall et al. (2016) obtained LiDAR data at a flight height of 1600 m and Brown et al. (2018) with an average flight height of 2160 m. These flight heights were chosen because of the greater coverage of the study area, project costs, among others. Neither their effect on waveforms nor the effect of the filtering level threshold in the waveforms was evaluated, and the filtering threshold used was the same as suggested by Hermosilla et al. (2014) and Lefsky et al. (2005).

Thus, our comparison is based on the experimental choice of the threshold. Considering our area of study (tropical forest) and using the methodology implemented in the software DASOS, the average amplitude of the waveform samples obtained at each flight height (900 m and 2000 m) was found to be the most suitable filtering threshold value.

Another result to be discussed is the approach used to analyse the waveforms inside the voxel, which in our case was the average intensity value of the waveform samples intercepting the voxel (MILTADO et al., 2019). According to Hermosilla et al. (2014), approaches that use the average value can mask the information from the most significant peaks of the waveforms, because the average will combine the highest and most relevant values with the less relevant intensity values. However, choosing the highest intensity value is not always the most relevant, as it is possible to have intensities from different pulses, whose emitted intensity is different, and occlusions that occurred before the pulse reached the target. In our case, low-intensity values may contain information from the lower layers of the forest, which are important for estimating variables such as above-ground biomass and density. We choose a voxel size that is large enough to avoid gaps between the voxels, but small enough to avoid too much averaging the intensity of the waveforms that intercept the voxel, thus there will likely be only some waveform samples in each voxel to get the average. Furthermore, as the intensities were not calibrated, we used attributes related to structure rather than intensity. In the end it is as if each voxel assumes a Boolean value (empty or non-empty) so the mean intensity value of each voxel is not of great importance.

In most studies involving the analysis of FWF data in forest ecosystems, return peaks are extracted and metrics related to these return peaks are obtained and used as input data for biomass modelling, with or without combination with DR LiDAR metrics. Cao et al. (2014) estimated biomass in subtropical forests of China by comparing three sets of LiDAR metrics using a multiple linear regression method: metrics from DR, metrics extracted from peaks of FWF data, and a combination of both metrics as predictor variables. DR data metrics were

related to the distribution of LiDAR pulse heights (statistics of heights, percentiles, and canopy cover), and FWF metrics were extracted based on Hermosilla et al. (2014) from pseudo-vertical waveforms within the voxel space, including waveform distance, number of peaks, roughness, and median energy height. The estimated rRMSE of above-ground biomass was 17.9% using discrete data, 22.1% for full-waveform data, and 15.8% for combined models. The results achieved in their study are similar to our findings for AGB, for both PR and FWF data (the best scenario with noise level 7), with an RMSE of 20.3% and 21.1%, respectively. In both the studies, data were collected with a flight height of 900 m. The combination of discrete return information about the 3D canopy structure and detailed information from the peaks extracted from FWF data on geometry and radiometry of the canopy provide more accurate information for biomass estimation than data from each separate return (CAO et al., 2014). In our case, we used the same attribute sets for the PR and FWF data, and despite the similarities, results with FWF data showed less biased estimations of AGB.

Above-ground biomass (AGB) was also estimated by Brown et al. (2018) in peat swamp forests in Malaysia. LiDAR data were acquired at a higher flight height (2160 m, on average), and they also used DR LiDAR metrics, metrics of peak returns (PR) extracted from FWF data, and a combination of both as input data in the predictive models. The difference is that Gaussian decomposition was used to identify peaks and later extract the FWF metrics. The results of their study also indicated that the combination of DR and FWF metrics provided better estimates for the variable AGB, with rRMSE of 10.8% and rRMSEs of 12.3%, and 11.6%, using only DR and FWF metrics, respectively. These results were better than those achieved in this work at a similar flight height (2000 m), with a lower rRMSE of 20.7% with FWF data using the threshold of nl_7, and the rBias closest to zero was obtained with FWF data using the threshold nl_2 (-0.39%). This suggests that a higher flight height is more important for preserving and analyzing full waveforms in comparison to peak returns. This is most likely due to the footprint size, which acquires information from multiple targets within a larger area, so the waveform samples between the peak returns contain important structural information that is discarded with Gaussian decomposition. In addition, the greater the distance is between the target and the sensor, the higher the possibility is to record low-level noise from the backscattered signal before it returns to the sensor. Thus, LiDAR data acquisition at higher flight heights, as in the case of large footprint systems like LVIS (ANDERSON et al., 2006, 2008; BLAIR; RABINE; HOFTON, 1999; DRAKE et al., 2002) and satellites like the Global Ecosystem Dynamics

Investigation (GEDI) (BLAIR et al., 2014; DUBAYAH et al., 2014, 2020; MARSELIS et al., 2020) are better interpreted by full waveform and not by peak extraction.

Regarding other forest variables, Sumnall et al. (2016) estimated composition and structure variables in mixed deciduous and coniferous forests in southern England by comparing the metrics from DR and FWF data. The flight height used to acquire LiDAR data was 1600 m. A point cloud with more returns was produced from FWF data by Gaussian decomposition to identify the peaks. Then, metrics related to the heights and pulse intensity were extracted from the DR and FWF point clouds, and metrics related to peaks, amplitude, and echo-width attributes were extracted from the FWF point cloud. While the structure variables were better determined with the DR data, the composition variables showed the best results with FWF data. However, as in our study, the differences in rRMSE in both cases were low (less than 7%). In their study, the basal area had an rRMSEs of 26% and 28% with the DR and FWF metrics, respectively; for the density the rRMSEs were 24% and 22% for DR and FWF metrics, respectively and for the number of tree species, the rRMSEs were 25% for DR data and 28% for FWF data. Regarding our results, the rRMSE of PR data, compared to the best scenario for FWF data, were 17.8% and 19.5% for the basal area; 16.6% and 17.9% for density and number of tree species, and 12.9% and 11.1% for the 900 m flight height. The rRMSEs with PR and FWF data with the best scenario from the 2000 m flight height were 20.1% and 21.3% for the basal area, 21.0% and 13.4% for density, and 12.7% and 12.3% for the number of tree species. Our results were better for both flight heights and data types compared to that presented by Sumnall et al. (2016). This may be because those authors use LiDAR intensities were not calibrated before extracting the metrics, so the estimations that use intensity metrics are difficult to compare. Moreover, it was reported that FWF data echo-width variables were relatively unimportant; however, it was found that the system stored an average of 42.4% of waveforms over emitted pulses, as reported by (MILTADO et al., 2021) who used the same dataset, thus reducing the potential of using waveform data to characterize the studied forest.

Shen et al. (2018) also estimated basal area, density, and AGB variables in Chinese subtropical forests using metrics extracted from DR, FWF LiDAR data, and vegetation indices from hyperspectral images. All data were acquired at a flight height of 900 m. First, noisy data were eliminated from FWF based on Hermosilla et al. (2014) and Lefsky et al. (2005) and Gaussian decomposition was used to determine peak positions. Voxelization was then performed to generate vertical pseudo-waveforms, based on and then to extract the FWF

metrics. Again, models that combine discrete metrics (e.g., percentile height and canopy coverage) with FWF metrics (e.g., median energy height, waveform distance, and vertical distribution ratio) were the ones that best estimated the basal area, with an RMSE of 31.3 %, density of 21.9%, and AGB of 15.4%. The authors mentioned that the FWF metrics complemented the DR metrics with information from the intermediate canopy and the understory, in addition to the arrangement and density of the forest, which best represented its structure. Except for the variable AGB, as shown in Table 4, for the 900 m flight height, our estimates were relatively better. According to the authors, this may be related to the way the system records the signals reflected in different forest typologies, and especially the use of non-parametric modelling techniques, which can improve the capacity to predict structural parameters. While Shen et al. (2018) estimated the variables using the parametric method of multiple linear regression, in our study, we used RF, a non-parametric method.

In case of Brazilian forests, Martins-Neto et al. (2021) commented on the lack of studies using LiDAR in other forest typologies besides forest plantations and the Amazon Forest. They estimated seven stand and diversity variables in the same study area. Peak returns from FWF data were used, and several metrics based on heights, pulse returns, and calibrated intensities were used as input data for estimating the variables. The best set of input data used were the LiDAR metrics transformed into new metrics by principal component analysis, as they aggregate information from different types of metrics, thus reducing the autocorrelation between them. Among the tested regression techniques, artificial neural networks with different architectures best estimated the study variables; however, even in the best case, the basal area, density, and the number of tree species presented rRMSE values of 22.5%, 16.2%, and 27.6%, respectively, with an underestimation of -12%, except for the basal area, for which the underestimation was -0.24%.

From our results, we have observed that even when using PR data, the voxel approach, as implemented in the software DASOS, estimated the forest variables with lower rRMSE and rBias than those presented by a previous study in the same area by Martins-Neto et al. (2021). The approach implemented in the software DASOS extracted the attributes at the level of voxelized space as statistics of the height of the voxel columns, distances between voxels, number of voxels that contain or do not contain waveform samples, and symmetry. In other voxel-based approaches, the computed metrics are related to the energy, heights, and peaks extracted from the Gaussian decomposition. Although it was developed for the detection of living or dead trees, the results of this study show that the extracted attributes were also adequate

for the characterization of plots in tropical forests. The rRMSE values were similar between the results obtained with PR and FWF data. However, the estimates were less biased with the FWF data from both flight heights, showing that voxelizing all waveform samples reduces bias, since the samples that are discarded when selecting peaks contain important information that can reduce the bias of estimations. Anderson et al. (2016) studying urban forests, also found less biased estimates when using FWF LiDAR data, compared to DR data, to estimate woodland canopy height. They also mentioned the larger capacity of FWF data to describe the understory. However, non-penetrable targets, such as roads and pavements, are better described by the DR data.

6 CONCLUSION

In this study, we compared the estimations of four forest variables using two LiDAR data types from waveforms: peak returns and full waveform, obtained at two flight heights (900 m and 2000 m). For comparison, we used the same voxel-based approach, implemented in the software DASOS, to extract attributes related to voxels and use them as inputs for the estimations.

The software DASOS is object-oriented and open sources and allows the easy addition of new attributes. The attributes extracted from the software DASOS were developed to characterize individual trees. In this study, we have shown that it is possible to extract attributes at the plot level as well. In future studies, new attributes may be added to DASOS for the better interpretation of point clouds of LiDAR data, which may improve the accuracy of estimating forest variables.

For the data from the lowest flight height (900 m), a higher low-level filtering threshold was necessary to achieve good results. This happens because at lower flight heights, the noise of the sensor can be confused with the signal, and a great deal of information between peaks can be discarded. The variables basal area, density, and AGB had their best estimates at low-level filtering threshold nl_7. The number of tree species did not present the same relationship as the other variables for the lowest flight, where low-level filtering threshold was nl_2. This may be related to the influence of low-intensity signals in the estimation of this variable and the difficulty of modelling, owing to the great heterogeneity of the number of species in the surveyed plots.

As the flight height increases, Low-level filtering threshold tends to decrease, as the system stores less noisy signals and the information contained between the waveform peaks is more relevant, so the best threshold is equal to 2 (nl_2) for all estimated variables.

Choosing the correct voxel size is crucial to get good results, the values chosen were adequate for the variables we estimated. Voxelizing both peak returns and waveform samples resulted in estimates with similar errors. This indicates that the method present in the software DASOS is also indicated when only peak return LiDAR data is available. However, estimations using full waveform analysis resulted in less biased estimates, showing the great potential of processing complete waveforms for the characterization and estimation of variables in tropical forests.

7 REFERENCES

- ALEXANDER, C.; BØCHER, P. K.; ARGE, L.; SVENNING, J.-C. Regional-scale mapping of tree cover, height and main phenological tree types using airborne laser scanning data. **Remote sensing of environment**, [s. l.], v. 147, p. 156–172, 2014.
- ALLOUIS, T.; DURRIEU, S.; VÉGA, C.; COUTERON, P. Stem volume and above-ground biomass estimation of individual pine trees from LiDAR data: Contribution of full-waveform signals. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, [s. l.], v. 6, n. 2, p. 924–934, 2012.
- ANDERSON, J. E.; PLOURDE, L. C.; MARTIN, M. E.; BRASWELL, B. H.; SMITH, M.-L.; DUBAYAH, R. O.; HOFTON, M. A.; BLAIR, J. B. Integrating waveform lidar with hyperspectral imagery for inventory of a northern temperate forest. **Remote Sensing of Environment**, [s. l.], v. 112, n. 4, p. 1856–1870, 2008.
- ANDERSON, J.; MARTIN, M. E.; SMITH, M. L.; DUBAYAH, R. O.; HOFTON, M. A.; HYDE, P.; PETERSON, B. E.; BLAIR, J. B.; KNOX, R. G. The use of waveform lidar to measure northern temperate mixed conifer and deciduous forest structure in New Hampshire. **Remote Sensing of Environment**, [s. l.], v. 105, n. 3, p. 248–261, 2006.
- ANDERSON, K.; HANCOCK, S.; DISNEY, M.; GASTON, K. J. Is waveform worth it? A comparison of Li DAR approaches for vegetation and landscape characterization. **Remote Sensing in Ecology and Conservation**, [s. l.], v. 2, n. 1, p. 5–15, 2016.
- ASPRS. **LAS Specification Version 1.2**. 2008. Disponível em: <https://www.asprs.org/a/society/committees/standards/asprs_las_format_v12.pdf>. Acesso em: 1 jul. 2021.
- ASPRS. **LAS SPECIFICATION VERSION 1.3 – R11**. 2010. Disponível em: <http://www.asprs.org/a/society/committees/standards/LAS_1_3_r11.pdf>. Acesso em: 20 maio. 2021.
- BELGIU, M.; DRĂGUȚ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 114, p. 24–31, 2016.
- BERVEGLIERI, A.; IMAI, N. N.; TOMMASELLI, A. M.; CASAGRANDE, B.; HONKAVAARA, E. Successional stages and their evolution in tropical forests using multi-temporal photogrammetric surface models and superpixels. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 146, p. 548–558, 2018.
- BERVEGLIERI, A.; IMAI, N. N.; TOMMASELLI, A. M.; MARTINS-NETO, R. P.; MIYOSHI, G. T.; HONKAVAARA, E. Forest cover change analysis based on temporal gradients of the vertical structure and density. **Ecological Indicators**, [s. l.], p. 107597, 2021.
- BERVEGLIERI, A.; TOMMASELLI, A. M. G.; IMAI, N. N.; RIBEIRO, E. A. W.; GUIMARAES, R. B.; HONKAVAARA, E. Identification of successional stages and cover changes of tropical forest based on digital surface model analysis. **IEEE Journal of Selected**

Topics in Applied Earth Observations and Remote Sensing, [s. l.], v. 9, n. 12, p. 5385–5397, 2016.

BLAIR, J. B.; DUBAYAH, R.; HOFTON, M. A.; LUTHCKE, S. B.; RABINE, D.; WAKE, S.; COYLE, B.; STYSLEY, P.; SALERNO, C. Full-waveform Airborne and Spaceborne Laser Altimetry for Mapping and Sampling the Earth's Forests, Cryosphere, and Land surfaces. In: AGU FALL MEETING ABSTRACTS 2014, **Anais...** [s.l: s.n.]

BLAIR, J. B.; RABINE, D. L.; HOFTON, M. A. The Laser Vegetation Imaging Sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 54, n. 2–3, p. 115–122, 1999.

BREIMAN, L. Random forests, machine learning 45. **J. Clin. Microbiol**, [s. l.], v. 2, n. 30, p. 199–228, 2001.

BROVELLI, M. A.; CRESPI, M.; FRATARCANGELI, F.; GIANNONE, F.; REALINI, E. Accuracy assessment of high resolution satellite imagery orientation by leave-one-out method. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 63, n. 4, p. 427–440, 2008.

BROWN, C.; BOYD, D. S.; SJÖGERSTEN, S.; CLEWLEY, D.; EVERS, S. L.; APLIN, P. Tropical peatland vegetation structure and biomass: Optimal exploitation of airborne laser scanning. **Remote Sensing**, [s. l.], v. 10, n. 5, p. 671, 2018.

BRUGGISSER, M.; RONCAT, A.; SCHAEPMAN, M. E.; MORSDORF, F. Retrieval of higher order statistical moments from full-waveform LiDAR data for tree species classification. **Remote sensing of environment**, [s. l.], v. 196, p. 28–41, 2017.

CAO, L.; COOPS, N. C.; INNES, J. L.; DAI, J.; RUAN, H.; SHE, G. Tree species classification in subtropical forests using small-footprint full-waveform LiDAR data. **International journal of applied earth observation and geoinformation**, [s. l.], v. 49, p. 39–51, 2016. a.

CAO, L.; COOPS, N.; HERMOSILLA, T.; INNES, J.; DAI, J.; SHE, G. Using small-footprint discrete and full-waveform airborne LiDAR metrics to estimate total biomass and biomass components in subtropical forests. **Remote Sensing**, [s. l.], v. 6, n. 8, p. 7110–7135, 2014.

CAO, L.; GAO, S.; LI, P.; YUN, T.; SHEN, X.; RUAN, H. Aboveground biomass estimation of individual trees in a coastal planted forest using full-waveform airborne laser scanning data. **Remote Sensing**, [s. l.], v. 8, n. 9, p. 729, 2016. b.

CHASE, M. W.; CHRISTENHUSZ, M. J. M.; FAY, M. F.; BYNG, J. W.; JUDD, W. S.; SOLTIS, D. E.; MABBERLEY, D. J.; SENNIKOV, A. N.; SOLTIS, P. S.; STEVENS, P. F. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. **Botanical Journal of the Linnean Society**, [s. l.], v. 181, n. 1, p. 1–20, 2016.

CHAUVE, A.; DURRIEU, S.; BRETAR, F.; PIERROT-DESEILLIGNY, M.; PUECH, W. Processing full-waveform lidar data to extract forest parameters and digital terrain model: Validation in an alpine coniferous forest. In: FORESTSAT CONFERENCE'07 2008, **Anais...** [s.l: s.n.]

CHAUVE, A.; MALLET, C.; BRETAR, F.; DURRIEU, S.; PIERROT-DESEILLIGNY, M.; PUECH, W. Processing full-waveform lidar data: modelling raw signals. In: INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES 2007 2007, **Anais...** [s.l: s.n.]

CRESPO-PEREMARCH, P.; RUIZ, L. A. A full-waveform airborne laser scanning metric extraction tool for Forest structure modelling. Do scan angle and radiometric correction matter? **Remote Sensing**, [s. l.], v. 12, n. 2, p. 292, 2020.

CRESPO-PEREMARCH, P.; RUIZ, L. Á.; BALAGUER-BESER, Á.; ESTORNELL, J. Analyzing the role of pulse density and voxelization parameters on full-waveform LiDAR-derived metrics. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 146, p. 453–464, 2018. a.

CRESPO-PEREMARCH, P.; TOMPALSKI, P.; COOPS, N. C.; RUIZ, L. Á. Characterizing understory vegetation in Mediterranean forests using full-waveform airborne laser scanning data. **Remote sensing of environment**, [s. l.], v. 217, p. 400–413, 2018. b.

DRAKE, J. B.; DUBAYAH, R. O.; CLARK, D. B.; KNOX, R. G.; BLAIR, J. B.; HOFTON, M. A.; CHAZDON, R. L.; WEISHAMPEL, J. F.; PRINCE, S. Estimation of tropical forest structural characteristics using large-footprint lidar. **Remote Sensing of Environment**, [s. l.], v. 79, n. 2–3, p. 305–319, 2002.

DUBAYAH, R.; BLAIR, J. B.; GOETZ, S.; FATOYINBO, L.; HANSEN, M.; HEALEY, S.; HOFTON, M.; HURTT, G.; KELLNER, J.; LUTHCKE, S. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. **Science of remote sensing**, [s. l.], v. 1, p. 100002, 2020.

DUBAYAH, R.; GOETZ, S. J.; BLAIR, J. B.; FATOYINBO, T. E.; HANSEN, M.; HEALEY, S. P.; HOFTON, M. A.; HURTT, G. C.; KELLNER, J.; LUTHCKE, S. B. The global ecosystem dynamics investigation. In: AGU FALL MEETING ABSTRACTS 2014, **Anais...** [s.l: s.n.]

DUCIC, V.; HOLLAUS, M.; ULLRICH, A.; WAGNER, W.; MELZER, T. 3D vegetation mapping and classification using full-waveform laser scanning. In: PROCEEDINGS 2006, Vienna, Austria. **Anais...** . In: WORKSHOP ON 3D REMOTE SENSING IN FORESTRY. EARSEL/ISPRS. Vienna, Austria

FAVORSKAYA, M. N.; JAIN, L. C. Overview of LiDAR technologies and equipment for land cover scanning. In: **Handbook on Advances in Remote Sensing and Geographic Information Systems**. [s.l.] : Springer, 2017. p. 19–68.

FERRAZ, A.; SAATCHI, S.; MALLET, C.; JACQUEMOUD, S.; GONÇALVES, G.; SILVA, C. A.; SOARES, P.; TOMÉ, M.; PEREIRA, L. Airborne lidar estimation of aboveground forest biomass in the absence of field inventory. **Remote Sensing**, [s. l.], v. 8, n. 8, p. 653, 2016.

FRIEDMAN, J.; TIBSHIRANI, R.; HASTIE, T. **The elements of statistical learning. Data mining, inference and prediction**, Springer series in Statistics, 2015.

GARCÍA-GUTIÉRREZ, J.; MARTÍNEZ-ÁLVAREZ, F.; TRONCOSO, A.; RIQUELME, J. C. A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. **Neurocomputing**, [s. l.], v. 167, p. 24–31, 2015.

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

GOBAKKEN, T.; NÆSSET, E. Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. **Scandinavian Journal of Forest Research**, [s. l.], v. 19, n. 6, p. 529–542, 2004.

HERMOSILLA, T.; COOPS, N. C.; RUIZ, L. A.; MOSKAL, L. M. Deriving pseudo-vertical waveforms from small-footprint full-waveform LiDAR data. **Remote sensing letters**, [s. l.], v. 5, n. 4, p. 332–341, 2014.

HERMOSILLA, T.; RUIZ, L. A.; KAZAKOVA, A. N.; COOPS, N. C.; MOSKAL, L. M. Estimation of forest structure and canopy fuel parameters from small-footprint full-waveform LiDAR data. **International journal of wildland fire**, [s. l.], v. 23, n. 2, p. 224–233, 2013.

HOLMGREN, J. Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. **Scandinavian Journal of Forest Research**, [s. l.], v. 19, n. 6, p. 543–553, 2004.

HUG, C.; ULLRICH, A.; GRIMM, A. Litemapper-5600-a waveform-digitizing LiDAR terrain and vegetation mapping system. **International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 36, n. Part 8, p. W2, 2004.

HYYPÄ, J.; HYYPÄ, H.; LECKIE, D.; GOUGEON, F.; YU, X.; MALTAMO, M. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. **International Journal of Remote Sensing**, [s. l.], v. 29, n. 5, p. 1339–1366, 2008.

HYYPÄ, J.; YU, X.; HYYPÄ, H.; VASTARANTA, M.; HOLOPAINEN, M.; KUKKO, A.; KAARTINEN, H.; JAAKKOLA, A.; VAAJA, M.; KOSKINEN, J. Advances in forest inventory using airborne laser scanning. **Remote sensing**, [s. l.], v. 4, n. 5, p. 1190–1207, 2012.

ISENBURG, M. **LAStools-efficient LiDAR processing software**. 2014. Disponível em: <<http://lastools.org/>>. Acesso em: 12 nov. 2020.

LAI, X.; ZHENG, M. A denoising method for LiDAR full-waveform data. **Mathematical Problems in Engineering**, [s. l.], v. 2015, 2015.

LECIGNE, B.; DELAGRANGE, S.; MESSIER, C. Exploring trees in three dimensions: VoxR, a novel voxel-based R package dedicated to analysing the complex arrangement of tree crowns. **Annals of botany**, [s. l.], v. 121, n. 4, p. 589–601, 2018.

LEFSKY, M. A.; HARDING, D. J.; KELLER, M.; COHEN, W. B.; CARABAJAL, C. C.; DEL BOM ESPIRITO-SANTO, F.; HUNTER, M. O.; DE OLIVEIRA JR, R. Estimates of forest canopy height and aboveground biomass using ICESat. **Geophysical research letters**, [s. l.], v. 32, n. 22, 2005.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R news**, [s. l.], v. 2, n. 3, p. 18–22, 2002.

- LIM, K.; TREITZ, P.; WULDER, M.; ST-ONGE, B.; FLOOD, M. LiDAR remote sensing of forest structure. **Progress in physical geography**, [s. l.], v. 27, n. 1, p. 88–106, 2003.
- MALLET, C.; BRETAR, F. Full-waveform topographic lidar: State-of-the-art. **ISPRS Journal of photogrammetry and remote sensing**, [s. l.], v. 64, n. 1, p. 1–16, 2009.
- MALLET, C.; LAFARGE, F.; BRETAR, F.; ROUX, M.; SOERGEL, U.; HEIPKE, C. A stochastic approach for modelling airborne lidar waveforms. **Laserscanning**, [s. l.], p. 201–206, 2009.
- MALTAMO, M.; NÆSSET, E.; VAUHKONEN, J. Forestry applications of airborne laser scanning. **Concepts and case studies. Manag For Ecosys**, [s. l.], v. 27, p. 2014, 2014.
- MALTAMO, M.; SUVANTO, A.; PACKALÉN, P. Comparison of basal area and stem frequency diameter distribution modelling using airborne laser scanner data and calibration estimation. **Forest Ecology and Management**, [s. l.], v. 247, n. 1–3, p. 26–34, 2007.
- MARSELIS, S. M.; ABERNETHY, K.; ALONSO, A.; ARMSTON, J.; BAKER, T. R.; BASTIN, J.-F.; BOGAERT, J.; BOYD, D. S.; BOECKX, P.; BURSLEM, D. F. Evaluating the potential of full-waveform lidar for mapping pan-tropical tree species richness. **Global Ecology and Biogeography**, [s. l.], v. 29, n. 10, p. 1799–1816, 2020.
- MARTINS-NETO, R. P.; TOMMASELLI, A. M. G.; IMAI, N. N.; DAVID, H. C.; MILTIADOU, M.; HONKAVAARA, E. Identification of Significant LiDAR Metrics and Comparison of Machine Learning Approaches for Estimating Stand and Diversity Variables in Heterogeneous Brazilian Atlantic Forest. **Remote Sensing**, [s. l.], p. 2444, 2021. Disponível em: <<https://www.mdpi.com/2072-4292/13/13/2444>>. Acesso em: 28 jun. 2021.
- MCROBERTS, R. E.; NÆSSET, E.; GOBAKKEN, T.; BOLLANDSÅS, O. M. Indirect and direct estimation of forest biomass change using forest inventory and airborne laser scanning data. **Remote Sensing of Environment**, [s. l.], v. 164, p. 36–42, 2015.
- MILTIADOU, M.; AGAPIOU, A.; GONZALEZ ARACIL, S.; HADJIMITSIS, D. G. Detecting Dead Standing Eucalypt Trees from Voxelised Full-Waveform Lidar Using Multi-Scale 3D-Windows for Tackling Height and Size Variations. **Forests**, [s. l.], v. 11, n. 2, p. 161, 2020.
- MILTIADOU, M.; CAMPBELL, N. D.; ARACIL, S. G.; BROWN, T.; GRANT, M. G. Detection of dead standing Eucalyptus camaldulensis without tree delineation for managing biodiversity in native Australian forest. **International journal of applied earth observation and geoinformation**, [s. l.], v. 67, p. 135–147, 2018.
- MILTIADOU, M.; CAMPBELL, N. D.; COSKER, D.; GRANT, M. G. A Comparative Study about Data Structures Used for Efficient Management of Voxelised Full-Waveform Airborne LiDAR Data during 3D Polygonal Model Creation. **Remote Sensing**, [s. l.], v. 13, n. 4, p. 559, 2021.
- MILTIADOU, M.; GRANT, M. G.; CAMPBELL, N. D.; WARREN, M.; CLEWLEY, D.; HADJIMITSIS, D. G. Open source software DASOS: Efficient accumulation, analysis, and visualisation of full-waveform lidar. In: SEVENTH INTERNATIONAL CONFERENCE ON

REMOTE SENSING AND GEOINFORMATION OF THE ENVIRONMENT (RSCY2019) 2019, **Anais...** : International Society for Optics and Photonics, 2019.

MILTADOU, M.; WARREN, M. A.; GRANT, M.; BROWN, M. Alignment of hyperspectral imagery and full-waveform LiDAR data for visualisation and classification purposes. **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 7, p. W3, 2015.

MMA; IBAMA; ICMBIO. **Plano de manejo da Estação Ecológica Mico - Leão - Preto.**, 2007. Disponível em: <http://www.icmbio.gov.br/portal/images/stories/imgs-unidades-coservacao/esec_mico_leao_preto.pdf>. Acesso em: 6 nov. 2019.

MONNET, J.-M.; CHANUSSOT, J.; BERGER, F. Support vector regression for the estimation of forest stand parameters using airborne laser scanning. **IEEE Geoscience and remote sensing letters**, [s. l.], v. 8, n. 3, p. 580–584, 2011.

MORSDORF, F.; FREY, O.; MEIER, E.; ITTEN, K. I.; ALLGÖWER, B. Assessment of the influence of flying altitude and scan angle on biophysical vegetation products derived from airborne laser scanning. **International journal of remote sensing**, [s. l.], v. 29, n. 5, p. 1387–1406, 2008.

NÆSSET, E. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. **Remote sensing of environment**, [s. l.], v. 80, n. 1, p. 88–99, 2002.

NÆSSET, E. Effects of different flying altitudes on biophysical stand properties estimated from canopy height and density measured with a small-footprint airborne scanning laser. **Remote Sensing of Environment**, [s. l.], v. 91, n. 2, p. 243–255, 2004. a.

NÆSSET, E. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. **Scandinavian Journal of Forest Research**, [s. l.], v. 19, n. 2, p. 164–179, 2004. b.

NIE, S.; WANG, C.; XI, X.; LUO, S.; LI, S.; TIAN, J. Estimating the height of wetland vegetation using airborne discrete-return LiDAR data. **Optik**, [s. l.], v. 154, p. 267–274, 2018.

OLIVEIRA, A. A. A. De; CENTENO, J. A. S.; HAINOSZ, F. S. POINT CLOUD GENERATION FROM GAUSSIAN DECOMPOSITION OF THE WAVEFORM LASER SIGNAL WITH GENETIC ALGORITHMS. **Boletim de Ciências Geodésicas**, [s. l.], v. 24, n. 2, p. 270–287, 2018.

PAL, M. Random forest classifier for remote sensing classification. **International Journal of Remote Sensing**, [s. l.], v. 26, n. 1, p. 217–222, 2005.

PERSSON, A.; HOLMGREN, J.; SODERMAN, U. Detecting and measuring individual trees using an airborne laser scanner. **Photogrammetric Engineering and Remote Sensing**, [s. l.], v. 68, n. 9, p. 925–932, 2002.

PERSSON, \AAsa; SÖDERMAN, U.; TÖPEL, J.; AHLBERG, S. Visualization and analysis of full-waveform airborne laser scanner data. **International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 36, n. part 3, p. W19, 2005.

- PIROTTI, F. Analysis of full-waveform LiDAR data for forestry applications: a review of investigations and methods. **iForest-Biogeosciences and Forestry**, [s. l.], v. 4, n. 3, p. 100, 2011.
- PRASAD, A. M.; IVERSON, L. R.; LIAW, A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. **Ecosystems**, [s. l.], v. 9, n. 2, p. 181–199, 2006.
- R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2017.
- REITBERGER, J.; KRZYSZEK, P.; STILLA, U. Analysis of full waveform lidar data for tree species classification. **International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 36, n. Part 3, p. 228–233, 2006.
- REITBERGER, J.; SCHNÖRR, C.; KRZYSZEK, P.; STILLA, U. 3D segmentation of single trees exploiting full waveform LIDAR data. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 64, n. 6, p. 561–574, 2009.
- REX, F. E.; SILVA, C. A.; DALLA CORTE, A. P.; KLAUBERG, C.; MOHAN, M.; CARDIL, A.; SILVA, V. S. Da; ALMEIDA, D. R. A. De; GARCIA, M.; BROADBENT, E. N. Comparison of Statistical Modelling Approaches for Estimating Tropical Forest Aboveground Biomass Stock and Reporting their Changes in Low-intensity Logging Areas using Multi-temporal LiDAR Data. **Remote Sensing**, [s. l.], v. 12, n. 9, p. 1498, 2020.
- RIEGL. **DataSheet LMS-Q680i**. 2012. Disponível em: <http://www.riegl.com/uploads/tx_pxpriegldownloads/10_DataSheet_LMS-Q680i_28-09-2012_01.pdf>. Acesso em: 7 abr. 2021.
- ROUSSEL, J.-R.; AUTY, D.; DE BOISSIEU, F.; MEADOR, A. S. **lidR: Airborne LiDAR data manipulation and visualization for forestry applications**. 2018. Disponível em: <<https://rdr.io/cran/lidR/>>. Acesso em: 21 jan. 2021.
- SHAN, J.; TOTH, C. K. **Topographic laser ranging and scanning: principles and processing**. [s.l.] : CRC press, 2018.
- SHEN, X.; CAO, L.; CHEN, D.; SUN, Y.; WANG, G.; RUAN, H. Prediction of forest structural parameters using airborne full-waveform LiDAR and hyperspectral data in subtropical forests. **Remote Sensing**, [s. l.], v. 10, n. 11, p. 1729, 2018.
- SHENDRYK, I.; BROICH, M.; TULBURE, M. G.; MCGRATH, A.; KEITH, D.; ALEXANDROV, S. V. Mapping individual tree health using full-waveform airborne laser scans and imaging spectroscopy: A case study for a floodplain eucalypt forest. **Remote Sensing of Environment**, [s. l.], v. 187, p. 202–217, 2016.
- SHIN, J.; TEMESGEN, H.; STRUNK, J. L.; HILKER, T. Comparing modeling methods for predicting forest attributes using LiDAR metrics and ground measurements. **Canadian Journal of Remote Sensing**, [s. l.], v. 42, n. 6, p. 739–765, 2016.

SUMNALL, M. J.; HILL, R. A.; HINSLEY, S. A. Comparison of small-footprint discrete return and full waveform airborne LiDAR data for estimating multiple forest variables. **Remote Sensing of Environment**, [s. l.], v. 173, p. 214–223, 2016.

THIEL, K. H.; WEHR, A. Performance capabilities of laser scanners—an overview and measurement principle analysis. **International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 36, n. 8, p. 14–18, 2004.

TOTH, C. K.; ZALETNYIK, P.; LAKY, S.; GREJNER-BRZEZIŃSKA, D. The potential of full-waveform LiDAR in mobile mapping applications. **Archiwum Fotogrametrii, Kartografii i Teledetekcji**, [s. l.], v. 22, 2011.

ULLRICH, A.; PFENNIGBAUER, M. Echo digitization and waveform analysis in airborne and terrestrial laser scanning. In: PHOTOGRAMMETRIC WEEK 2011, **Anais...** [s.l: s.n.]

VAUHKONEN, J.; ENE, L.; GUPTA, S.; HEINZEL, J.; HOLMGREN, J.; PITKÄNEN, J.; SOLBERG, S.; WANG, Y.; WEINACKER, H.; HAUGLIN, K. M. Comparative testing of single-tree detection algorithms under different types of forest. **Forestry**, [s. l.], v. 85, n. 1, p. 27–40, 2012.

WAGNER, W.; ULLRICH, A.; DUCIC, V.; MELZER, T.; STUDNICKA, N. Gaussian decomposition and calibration of a novel small-footprint full-waveform digitising airborne laser scanner. **ISPRS journal of Photogrammetry and Remote Sensing**, [s. l.], v. 60, n. 2, p. 100–112, 2006.

WANG, C.; LUO, S.; XI, X.; NIE, S.; MA, D.; HUANG, Y. Influence of voxel size on forest canopy height estimates using full-waveform airborne LiDAR data. **Forest Ecosystems**, [s. l.], v. 7, n. 1, p. 1–12, 2020.

WANG, H.; GLENNIE, C. Fusion of waveform LiDAR data and hyperspectral imagery for land cover classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 108, p. 1–11, 2015. Disponível em: <<https://doi.org/10.1016/j.isprsjprs.2015.05.012>>

YAO, W.; KRZYSZEK, P.; HEURICH, M. Tree species classification and estimation of stem volume and DBH based on single tree extraction by exploiting airborne full-waveform LiDAR data. **Remote Sensing of Environment**, [s. l.], v. 123, p. 368–380, 2012.

YU, X.; HYYPPÄ, J.; KAARTINEN, H.; MALTAMO, M. Automatic detection of harvested trees and determination of forest growth using airborne laser scanning. **Remote sensing of Environment**, [s. l.], v. 90, n. 4, p. 451–462, 2004.

ZHOU, T.; POPESCU, S. C. Bayesian decomposition of full waveform LiDAR data with uncertainty analysis. **Remote sensing of environment**, [s. l.], v. 200, p. 43–62, 2017.

ZHOU, T.; POPESCU, S. C.; LAWING, A. M.; ERIKSSON, M.; STRIMBU, B. M.; BÜRKNER, P. C. Bayesian and classical machine learning methods: a comparison for tree species classification with LiDAR waveform signatures. **Remote Sensing**, [s. l.], v. 10, n. 1, p. 39, 2018a.

ZHOU, T.; POPESCU, S.; MALAMBO, L.; ZHAO, K.; KRAUSE, K. From LiDAR waveforms to Hyper Point Clouds: A novel data product to characterize vegetation structure. **Remote Sensing**, [s. l.], v. 10, n. 12, p. 1949, 2018b.

CHAPTER 5 – CLASSIFICATION OF TREE SPECIES IN THE UPPER CANOPY OF ATLANTIC FOREST INTEGRATING HYPERSPECTRAL IMAGES AND LIDAR DATA

1 INTRODUCTION

The discrimination of tree species is an essential task in natural forest surveys, as this information serves to monitor biodiversity and invasive species, sustainable forest management, and conservation practices, floristic and phytosociological forest inventory, wildlife habitat mapping, among others (FASSNACHT et al., 2016; PIIRONEN et al., 2017).

Remote Sensing data have great potential for species differentiation and classification, especially with the development of narrowband sensors, such as hyperspectral sensors. These sensors can acquire nearly continuous reflectance spectra in many narrow bands for each pixel. These spectra can be used for detailed quantitative analyses.

However, there is a great diversity of species in tropical forests and heterogeneity in the canopy, making species classification with high-resolution images a big challenge. In addition, other issues as the spectral similarity between different species or between age classes of the same species make the spectra indistinguishable, increasing classification errors (COCHRANE, 2000; FÉRET; ASNER, 2012; ZHANG et al., 2006). Furthermore, hyperspectral sensors are not suitable to derive structural parameters of the forest, such as tree height, canopy volume, or the number of strata of a stand, since only the reflectance of the top of objects is recorded (BUDDENBAUM; SEELING; HILL, 2013). These structural data can provide important information to differentiate species in tropical forests.

On the other hand, active remote sensing using LiDAR (Light Detection and Ranging) systems can provide three-dimensional information about the vegetation, thus allowing a better understanding of the geometry of the vertical structure of forests (KIM et al., 2009; MORSDORF et al., 2010).

The integration of different remote sensing data, like LiDAR (peak or full-waveform returns), with optical images (multispectral or hyperspectral) has been used as complementary sources to obtain abundant high-quality data (LUO et al., 2016) and thus provide more information about the forests. The integration of three-dimensional geometric data from LiDAR with the spectral information of optical images offers a rich range of information that can facilitate and increase the accuracy of tree species classification in complex forests.

Due to the great diversity of species present in tropical forests, monitoring changes in these forests are of great relevance given climate change, rampant deforestation, and changes in flora composition (FÉRET; ASNER, 2012). Thus, the objective of this work was to extract information from LiDAR geometric data (peak and full waveform returns) and spectral data from hyperspectral images obtained from a lightweight camera to classify eight species present in the upper canopy of a Brazilian Atlantic Forest remnant. The three datasets (PR and FWF LiDAR data and hyperspectral images) were used isolated and combined, resulting in five different classification scenarios that were analysed to assess the improvements in the classification accuracy with the geometric data only or combined with spectral data.

2 DATASETS ACQUISITION

2.1 STUDY SITE AND SAMPLES SELECTION

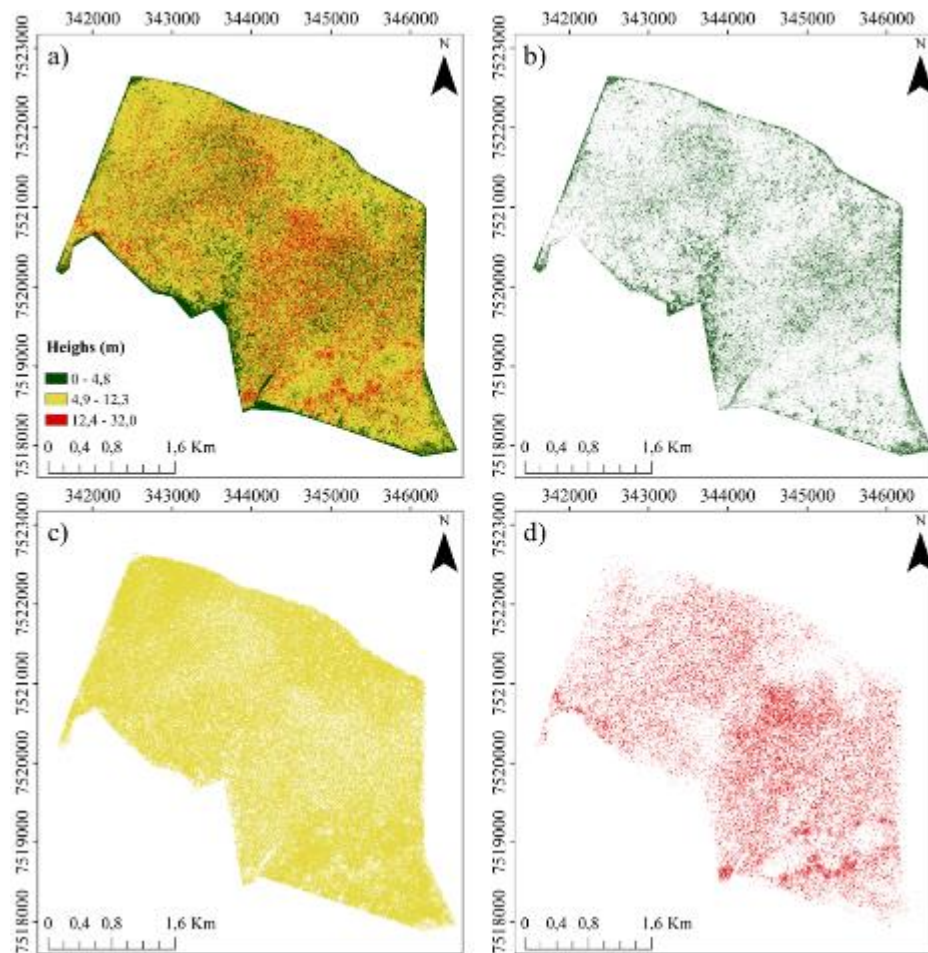
The study area has been described in previous chapters, mainly in Chapter 2, as well as field surveys for the measurement and identification of tree species. The upper stratum trees of the forest were selected for classification from the set of all trees identified in the surveyed plots. The vertical stratification of vegetation was based on Mariscal-Flores, (1993) and Souza et al. (2003), who divided the forest into three strata (lower, middle and upper), based on the average height of the trees and the standard deviation of the heights. The lower stratum comprises trees with total height (H) less than the average height (H_m) minus one standard deviation (1σ); the upper stratum is defined as $H \geq (H_m + 1\sigma)$ and the middle stratum comprises trees with $(H_m - 1\sigma) \leq H < (H_m + 1\sigma)$.

In native forests, the total height variable is difficult to obtain in the field. For the vertical stratification of the forest, we have used the CHM (Canopy Height Model) obtained from the LiDAR survey in the study area. The description of the technique to produce the CHM appears in Chapter 3.

To avoid the effects of ground points on stratification, we considered all points above 1 m as vegetation. As a result, the lower stratum comprises trees below 4.8 m, the middle stratum includes trees with heights between 4.9 m and 12.3 m and the upper stratum comprises trees above 12.4 m (Figure 1). Smaller trees were not selected due to the presence of lianas which can significantly modify the spectral response of these trees (MIYOSHI et al., 2020). In addition, it is difficult to isolate these trees' crowns in aerial images due to their location below larger trees, causing uncertainties in the correct recognition. As most trees have a height

between 6.4 m and 12.8 m (as shown in Figure 4 in Chapter 2) the middle stratum is more prominent, and the height of the upper canopy is not very significant, despite the presence of some tall trees (Above 25 m) in the *Ponte Branca* Forest remnant.

Figure 1 – Canopy height models of *Ponte Branca* Forest remnant. (a) All tree heights. (b) Lower stratum. (c) Middle stratum. (d) Upper stratum.



The selection was based on two criteria: (1) trees higher than 12.4 m (upper stratum) and (2) species containing at least 6 samples. These selection criteria ensure that there would be training and validation samples for the classification algorithm. Based on these conditions, 89 individuals of eight different species were selected. A summary of the selected species is presented in Table 1.

Table 1 – Characteristics of selected species in the field.

ID	Tree species	Family	ITC ¹	Hm(m) ²	Characteristics
AnPe	<i>Anadenanthera peregrina</i>	Fabaceae - Mimosoidae	7	15.7	Evergreen species, with characteristics from pioneer to early secondary. It is fast-growing and its uses include urban afforestation, recovery of degraded areas, and wood for civil construction (CARVALHO, 2003).
ApLe	<i>Apuleia leiocarpa</i>	Fabaceae - Caesalpionideae	7	13.4	Deciduous and slow-growing species, with characteristics from pioneer to early secondary. Its wood is resistant, suitable for construction of external structures. Furthermore, it can be used in urban afforestation, honey production, and riparian forest restoration in areas without flooding (CARVALHO, 2003).
AsPo	<i>Aspidosperma polyneuron</i>	Apocynaceae	10	23.5	Evergreen species, late secondary to climax. Long-lived species with very slow growth. Wood with a high economic value has good mechanical resistance used in the furniture industry, construction, carpentry, and shipbuilding (CARVALHO, 2003).
CoLa	<i>Copaifera langsdorffii</i>	Fabaceae - Caesalpionideae	12	14.5	Semi-deciduous tree, with late secondary to climax characteristics. Species with remarkable plasticity and easy adaptation. Long-lived tree with moderate to slow growth. High durability wood used in civil construction. However, the most significant feature of this species is the extraction of its essential oil, used in the cosmetics, plastics, paints, and resins industry (CARVALHO, 2003).
HeAp	<i>Helietta apiculata</i>	Rutaceae	6	13.7	Evergreen tree, with early and late secondary characteristics. This species is slow-growing, with dense wood, and is very useful for manufacturing pieces that require great durability. In addition, this species has a good development in shallow and rocky soils, indicated for the recovery of degraded areas (CARVALHO, 2008).
HyCo	<i>Hymenaea courbaril</i>	Fabaceae - Caesalpionideae	12	14.8	Long-lived semi-deciduous tree with late secondary to climax characteristics. This species presents moderate to slow growth with high-density wood. The uses are for civil and external construction and carpentry. The resin from this tree is used to manufacture varnishes and medicinal uses. In addition, this species can be used for the production of honey (CARVALHO, 2003).
PtPu	<i>Pterodon pubescens</i>	Fabaceae - Faboideae	12	17.9	Deciduous species, with characteristic of initial secondary. It is fast-growing, and the wood presents high density being used for civil construction. Other uses of this species include honey production, urban afforestation, and recovery of degraded areas (CARVALHO, 2014).
SyRo	<i>Syagrus romanzoffiana</i>	Arecaceae	23	13.2	Palm tree, with a characteristic of pioneer species, early secondary and late secondary. This species has great plasticity, occurring in soils of low and high chemical fertility, drained to flooded. Its growth is slow, and its fruits serve as food for countless animals (CARVALHO, 2006).

¹ Individual tree crown; ² Average height of the species obtained from the CHM.

2.2 REMOTE SENSING DATA

2.2.1 LiDAR data

The acquisition of these data was described in Chapter 3 Section 2.2 and in Chapter 4 Section 3.4. For the classification of tree species, we have used the data obtained at a flight height of 900 m, which was delivered in two forms: peak returns (PR) and full waveform (FWF).

2.2.2 Hyperspectral imagery data

The images were acquired by the hyperspectral frame camera Rikola, model DT-0011, produced by Senop Ltd (MIYOSHI et al., 2018; OLIVEIRA; TOMMASELLI; HONKAVAARA, 2016, 2019). The camera was mounted on an UAV (Unmanned Aerial Vehicle) rotor-wing quadcopter.

The Rikola camera is based on the FPI (Fabry-Perrot Interferometer), which consists of two partially reflective parallel surfaces, separated from each other with an air gap. This separation determines the wavelength transmitted by the interferometer, as the light rays that pass through the surfaces undergo multiple reflections according to the separation distance. Thus, changing the separation distance between the surfaces makes it possible to sensitize the camera sensor at different wavelengths (HONKAVAARA et al., 2013; MÄKELÄINEN et al., 2013; SAARI et al., 2009). In addition, the camera has two CMOS sensors that operate simultaneously, with a pixel size of 5.5 μm generating images with 1017 x 648 pixels. The first sensor collects images at wavelengths between 647 - 900 nm and the second sensor between 500 - 635 nm, with a minimum spectral resolution of 10 nm at the FWHM (full width at half maximum) (MIYOSHI et al., 2018). These features allow a flexible configuration of spectral bands (MIYOSHI et al., 2018). In addition to these sensors, the Rikola camera has an auxiliary sensor that measures the irradiance and a GNSS receiver that provides the camera's latitude and longitude at the time of image acquisition.

The camera was configured to standalone mode, in which images are saved in a memory card. In this mode, the number of spectral bands to be acquired is limited to 25, due to the transfer time of the images acquired by the sensor to the memory card, the acquisition interval between two sequential images and exposure time of each image (MIYOSHI et al., 2018; OLIVEIRA; TOMMASELLI; HONKAVAARA, 2019).

The choice of wavelengths for each band (Table 2) was based on Miyoshi et al. (2018), who selected the best set of wavelengths based on the spectral characterisation of the tree species present in the *Ponte Branca* forest remnant.

Table 2 – Wavelengths (λ) used in each of the Rikola camera bands and their respective FWHM.

Band	λ (nm)	FWHM(nm)	Band	λ (nm)	FWHM(nm)
1	506.22	12.44	11	650.96	14.44
2	519.94	17.38	12	659.72	16.83
3	535.09	16.84	13	669.75	19.80
4	550.39	16.53	14	679.84	20.45
5	565.10	17.26	15	690.28	18.87
6	580.16	15.95	16	700.28	18.94
7	591.90	16.61	17	710.06	19.70
8	609.00	15.08	18	720.17	19.31
9	620.22	16.26	19	729.57	19.01
10	628.75	15.30	20	740.42	17.98
			21	750.16	17.97
			22	769.89	18.72
			23	780.49	17.36
			24	790.30	17.39
			25	819.66	17.84

The integration time was 10 ms, with an interval of 0.061 seconds between adjacent band exposures. Thus, each hyperspectral cube with 25 bands took 0.899 seconds to be acquired. Due to the misalignment between the two sensors of the Rikola camera and the UAV displacement during acquisition, the spectral bands of the hyperspectral cube show a slight difference in orientation and position. These misalignments are corrected with orthorectification of all hyperspectral image bands (MIYOSHI et al., 2018; NÄSI et al., 2015).

Four flight campaigns were needed to obtain images of the 15 surveyed plots in the *Ponte Branca* forest remnant. The flight campaigns were carried out in 2016 and 2017 (Table 3) in the same season (winter) and under the same clear day conditions with few clouds and wind to avoid differences of solar angle and illumination. The same characteristics were also maintained in different years to avoid phenological differences between the various trees of the same species.

On each flight, image blocks were acquired with a longitudinal overlap of at least 70% and lateral overlap of at least 50%.

Table 3 – Summary of flight campaigns for images acquisition.

Plot	Date	Time (UTC – 3h)
P4, P5, P6	09 August 2016	11h46
P1, P3	10 August 2016	13h05
P8 to P15	01 July 2017	10h11
P2, P7	01 July 2017	12h19

In each flight campaign, signalized Ground Control Points (GCP) and radiometric targets were placed close to the study site, for bundle adjustment and radiometric calibration (Figure 2). GCP coordinates were determined with a dual-frequency GNSS receiver. The radiometric targets were produced with EVA (Ethylene Vinyl Acetate), with approximated dimensions of 0.90 m x 0.90 m in three colours: black, dark grey and light grey. On these targets, reflectance measurements were taken with a FieldSpec® Handheld spectroradiometer, manufactured by ASD (ASD, 2002), to transform the images' DN (digital numbers) into physical values.

Figure 2 – Targets located near the overflow area. In red the radiometric targets, and in blue the GCPs targets.



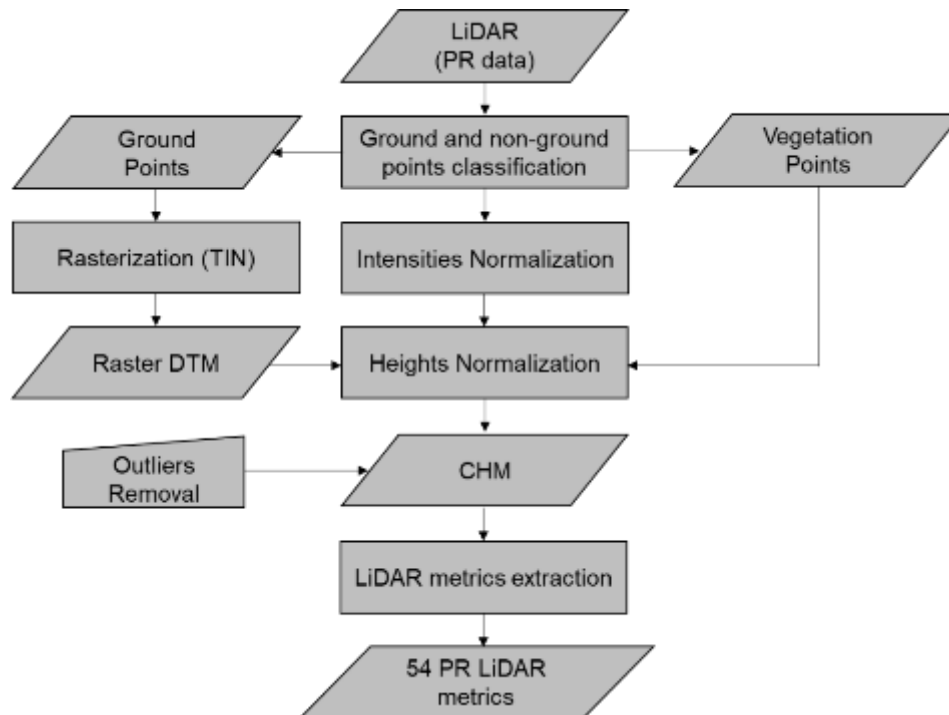
3 METHODOLOGY

3.1 LiDAR DATA PROCESSING

The processing of PR LiDAR peak returns data was described in Chapter 3, Section 2.3. A summary of the processing steps is presented as a flowchart in Figure 3. The point cloud was classified as ground points and vegetation points, using the software *LAStools* (ISENBURG,

2014). The *lasR* package (ROUSSEL et al., 2018) in R environment (R CORE TEAM, 2017) was used to perform the normalisation of intensities, generation of DTM (Digital Terrain Model), generation of the CHM, filtering of noisy points and, finally, extraction of 54 metrics. Digital models and LiDAR metrics were extracted in raster format with a GSD (Ground Sample Distance) of 0.50 m, and are related to the elevation, intensity and pulse returns values. The details of extracted metrics and their meaning is presented in Chapter 3, Table 2.

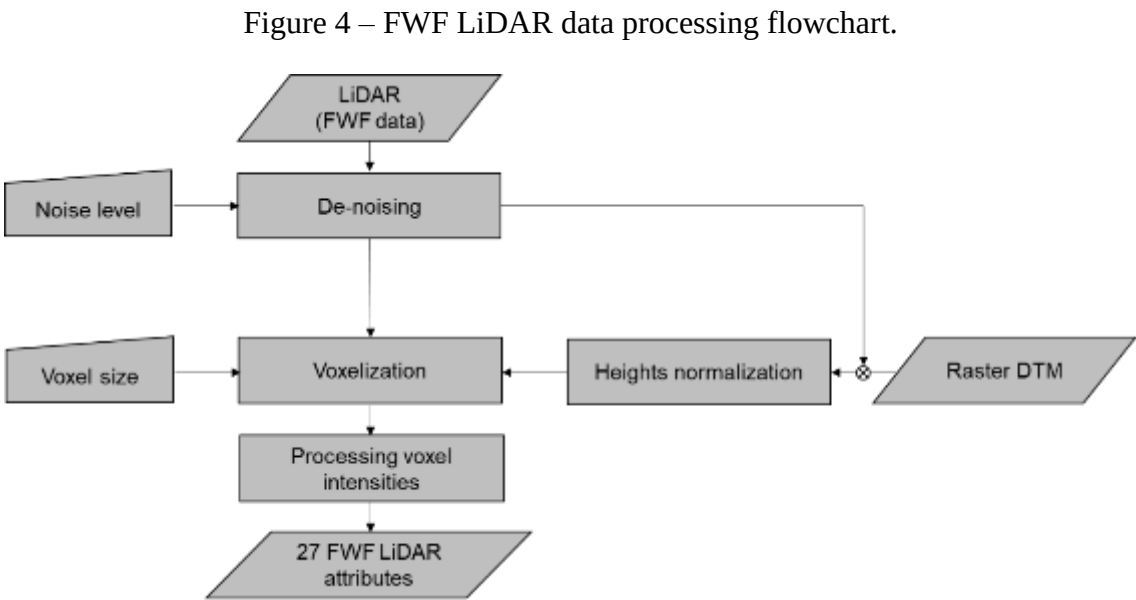
Figure 3 – PR LiDAR data processing flowchart.



In Chapter 4, Section 3.5, there is a description of FWF LiDAR data processing. These data were processed in the open-source software DASOS (MULTIADOU et al., 2019). The first step was de-noising the data through low-level filtering. The filtering threshold adopted was 2 (nl_2), which means that any sample below this amplitude value was eliminated. This value was chosen based on experimental evidence (Chapter 4, Section 4.2), and it was the best threshold for estimating the variable "*number of tree species*", so it was expected that this threshold should also be suitable for classifying tree species.

After de-noising, voxelization was performed, with voxels size of 1 m x 1 m x 1 m. In this step, the normalization of the heights of the voxel columns was performed using the DTM produced with PR data. From the voxelized space, 27 attribute vectors were extracted through a sliding 3D cuboid window, which moves downwards from the top of the voxelized space,

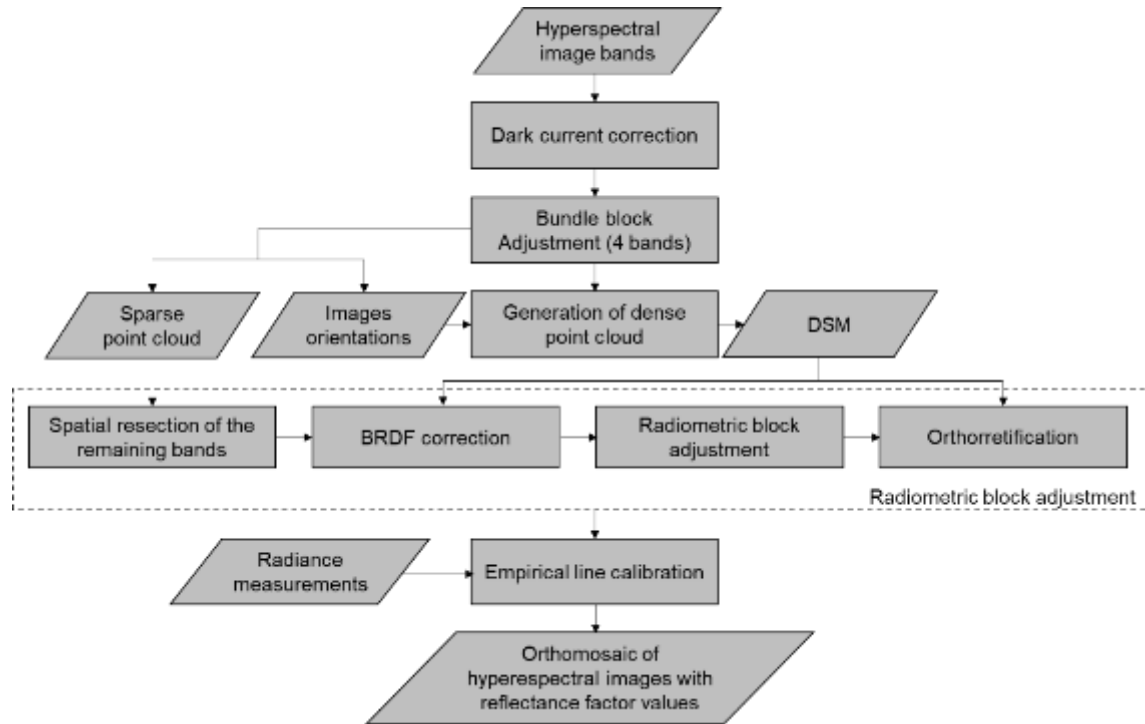
until the first voxel that contains an accumulated intensity value derived from the waveform samples (MILTIADOU et al., 2018, 2020). A detailed description of this process is presented in Chapter 4, Section 3.5.2 and the extracted attributes with their respective explanation are in Table 3. A summary of the processing flow with LiDAR FWF data is shown in Figure 4.



3.2 HYPERSPECTRAL IMAGES PROCESSING

The images obtained by the Rikola hyperspectral camera require some specific processing to correct the anisotropy factor and possible lighting variations during image acquisition (HONKAVAARA et al., 2013; MIYOSHI et al., 2018; MORIYA et al., 2017). The processing flow of hyperspectral images is shown in Figure 5. This process was applied for each flight campaign.

Figure 5 – Hyperspectral images flowchart. Adapted from Moriya et al. (2017) and Näsi et al. (2015).



The processing of hyperspectral images was performed using the methodology used by Honkavaara et al. (2012a, 2012b, 2013); Miyoshi et al. (2018, 2020); Moriya et al. (2017) and Näsi et al. (2015). Firstly, the images were corrected of dark current, using an image with the camera lens obstructed by an opaque low-reflectance object, to remove electronic noise from the camera (MORIYA et al., 2017).

The geometric processing was performed in several steps, the first being with the Agisoft PhotoScan software (Agisoft LLC, St. Petersburg, Russia). To optimize processing time, image orientations were estimated for four bands of the Rikola camera, two from each sensor (band 1 = 506.22 nm and band 10 = 628.75 nm from sensor two; band 11 = 650.96 nm and band 25 = 819.66 nm from sensor one) in a simultaneous bundle adjustment. The IOPs (interior orientation parameters) and EOPs (exterior orientation parameters) were estimated using self-calibrating BBA (bundle block adjustment). The IOPs were estimated for each sensor. The EOPs were estimated using the camera's GNSS positions as initial values and refined in the BBA and with the GCPs. From the generated point cloud, the estimated parameters were optimized with the manual removal of outliers and using gradual selection of tie points to verify the projection error. After these procedures, calibrated IOPs and EOPs and a sparse point cloud were generated; then, there were generated a dense point cloud, a DSM

(digital surface model) with a GSD (ground sample distance) of 0.25 cm GSD and a resampled DSM with GSD of 5 m (MIYOSHI, 2016; MIYOSHI et al., 2020; MORIYA et al., 2017).

After this initial geometric processing, further photogrammetric techniques and radiometric processing were applied based on the methodology developed by Honkavaara et al. (2012a, 2013, 2017). The EOPs for the remaining bands were computed with space resection, using the sparse point cloud as a source of control. However, due to the anisotropic characteristics of vegetation reflectance, quality of the sensor system, stability and atmospheric conditions, lighting changes due to clouds and solar position, the same object does not present the same DN in different images. While the reflectance anisotropy is modelled by the BRDF (bidirectional reflectance distribution function), the relative differences in the overlapped images must be estimated in radiometric block adjustment (HONKAVAARA et al., 2012a). The radiometric block adjustment assumes that the same object must provide a similar DN (digital number) in all the images in which it appears. This method uses DN values of radiometric tie points, obtained from the resampled DSM in the overlapped images. This information determines the parameters of the radiometric model describing the differences between the DNs in the different images using the principle of weighted least squares. The DN values in the radiometric tie points were determined from a search window of predefined size (5 m x 5 m). Relative correction parameters are determined in relation to a reference image obtained in the nadir to correct the differences in illumination between the images, and a linear BRDF model was applied (Equation 1).

$$DN_{jk} = a_{relj}(a_{abs} \cdot R_{jk}(\theta_i, \theta_r, \varphi) + b_{abs}) \quad (1)$$

where DN_{jk} is the digital number of pixel k in image j ; $R_{jk}(\theta_i, \theta_r, \varphi)$ is the reflectance factor with respect to the zenith angle of incident light θ_i and reflected light θ_r , and with the relative azimuthal angle $\varphi(\varphi_r - \varphi_i)$ in relation to the incident φ_i and reflected φ_r azimuthal angle (further details about this model see Honkavaara et al. (2013)); a_{relj} is the relative correction factor of different illumination with respect to the reference image. Miyoshi et al. (2018) working in the same study area, defined the optimal value for a_{relj} as 1; a_{abs} and b_{abs} are the parameters defined by the empirical line.

In radiometric block adjustment, orthorectification of each image band is also performed. A DSM with the same GSD as the mosaic to be produced was used. Furthermore,

the bands' misalignments were also corrected in this process. At the end of this process, an orthomosaic with 25 spectral bands and corrected for illumination and anisotropy was produced.

The empirical line method was applied in the orthomosaics from data obtained with the spectroradiometer in the radiometric reference targets, for each flight campaign. The goal is to calculate the relation between the values obtained in the Rikola images and the reflectance of the targets in the field, from a linear regression. The values of gain and offset were estimated, transforming the DN into physical values of reflectance factor (BAUGH; GROENEVELD, 2008; RICHARDS; JIA, 2006)

As the spectroradiometer used in our work collects information in the range between 325 nm and 1075 nm, it was necessary to adjust the wavelength ranges to match the configuration of the bands in the Rikola camera. Thus, the spectral curves were simulated according to the spectral ranges of the camera bands, adopting the Gaussian curve for spectral sensitivity (MORIYA et al., 2017). This simulation allows evaluating the adherence of the spectral response of the targets obtained with the bands of any camera with that obtained by the spectroradiometer, which has a more refined spectral resolution (MIYOSHI, 2016). With the physical values of reflectance factor, it is possible to characterize the targets spectrally, compare data from different sensors or different times, and obtain vegetation indices (JENSEN, 2009; PONZONI; SHIMABUKURO; KUPLICH, 2012).

3.3 FEATURES EXTRACTION

The delineation of ITC (individual tree crowns) was performed manually through the interactive vectorization. Automatic segmentation methods were not used due to the low number of samples and the difficulty of applying successful techniques in tropical forests due to the complexity and heterogeneity of the canopy structure (ASNER et al., 2002; MILLIKAN et al., 2019; TOCHON et al., 2015).

The vectorization process was performed twice: one for the orthomosaic of the hyperspectral images and another for the CHM obtained by the LiDAR data. This was done, due to the misalignment of the two datasets, caused by the different acquisition times and by the lack of GCP inside the forest which probably affected the estimated orientations with bundle block adjustment. Furthermore, since the study area is a dense forest, it would be challenging to find homologous points to perform suitable co-registration between the two datasets.

Care was taken to avoid vectorizing structures that did not belong exclusively to the tree of interest. However, shadows and structures below the treetops could be inserted in the ITC, thus affecting the computed spectral response of the segmented trees. To cope with these uncertainties, the centroid of each ITC was calculated, and then the average value of a window with 5 x 5 pixels, centred on the centroid, was used to extract features from tree species in both datasets (hyperspectral mosaics and LiDAR).

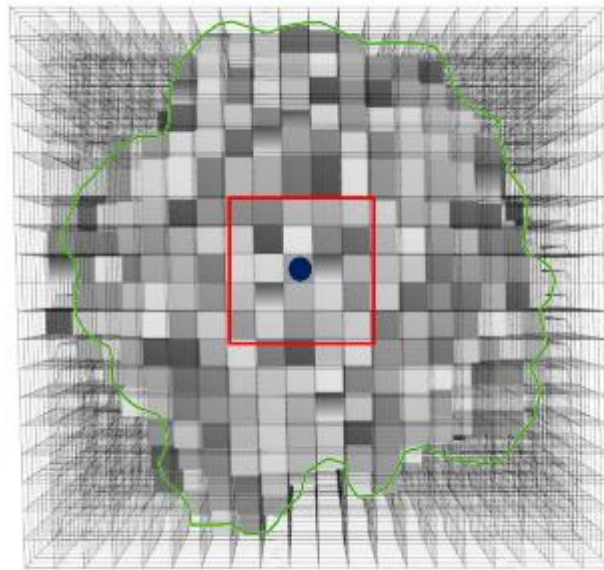
The extracted PR LiDAR metrics were transformed into a new set of variables by PCA (principal component analysis). Martins-Neto et al. (2021) tested different sets of LiDAR metrics were tested to estimate stand and diversity variables in the same area of our study. The PCs (principal components) were the best set of inputs, as it decreases the correlation between the metrics and gathers information of each LiDAR metric on PCs (ABDI; WILLIAMS, 2010; CLIFF, 1988; HONGYU; SANDANIELO; DE OLIVEIRA JUNIOR, 2016; MANLY; ALBERTO, 2016).

The PCA transformation was performed in R environment (R CORE TEAM, 2017) with the *FactMineR* package (LÊ; JOSSE; HUSSON, 2008). Firstly, the LiDAR metrics were normalized to zero mean and variance equal to one, selecting components based on the Kaiser criterion (KAISER, 1958). Components with eigenvalues greater than one explain the variation of the LiDAR metrics. Thus, five PCs were chosen in this process. From each selected tree, 5 PCs were extracted by the transformation of the original LiDAR metrics, and used as input data for classification.

In a previous analysis presented in Chapter 4, Section 3.5.2, attributes extraction from FWF LiDAR data was performed at the plot level. However, our objective in this experiment is to characterize each tree segment for classification. Thus, instead of voxelizing the entire plot, defining the voxel column of interest from the centre of the plot, the column was defined from the centroid of each tree, obtained from the ITCs (Figure 6).

As the defined voxel size was 1 m x 1 m x 1 m, we used a 3D window with cuboid geometry, to extract attributes from the voxelized space, with dimensions of 5 x 5 voxels in X and Y. For the axis Z, the value was related to the tree's height. In this case, care must be taken to include only trees' points, avoiding ground points. Of the 27 attributes that can be extracted from the software DASOS (MILTADOU et al., 2019), we have used 19, as the waveforms did not have their intensities calibrated.

Figure 6 – Top view of a voxelised tree, with the ITC (in green), the centroid calculated from the ITC (in blue) and the 3D window with 5 voxels in X and Y.

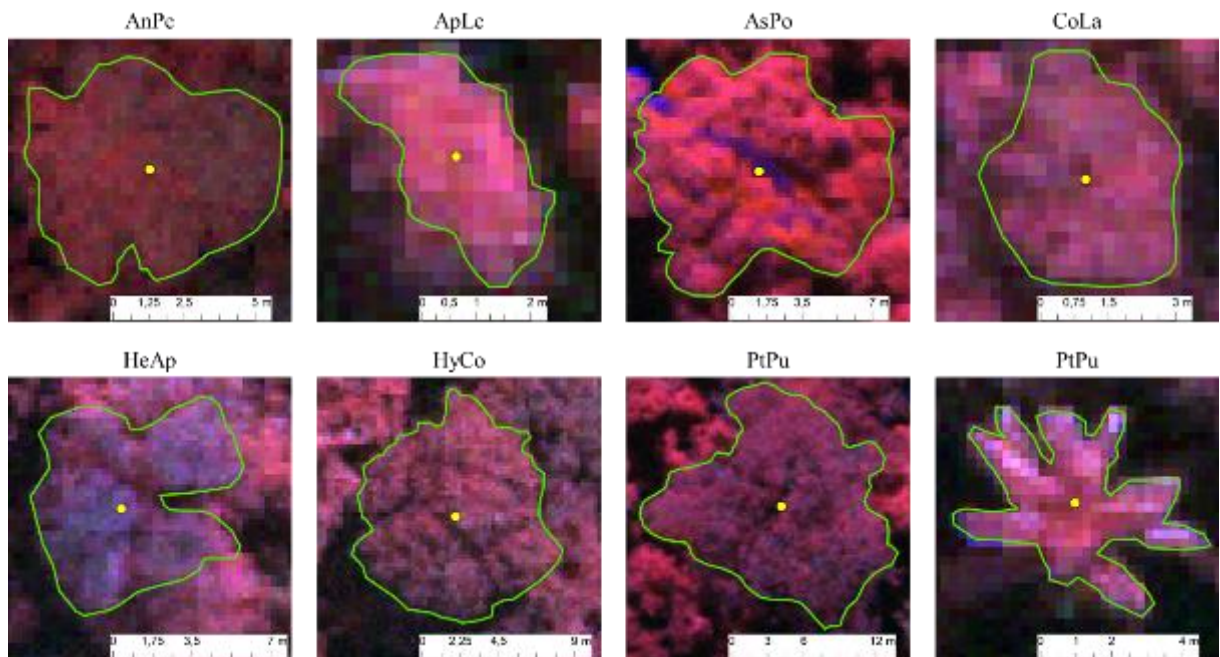


The features extracted from the hyperspectral orthomosaics were the spectral signature of each tree. An average curve was calculated for each one of the tree species. The extracted spectra were visually analyzed, and those wavelengths that best differentiated the species were verified. Then, VI (vegetation indices) related to these wavelengths were computed. Eleven vegetation indices related to structure, leaf pigments, and plant physiology were calculated (Table 3) and used for classification. All VIs were adapted to the spectral bands of the Rikola hyperspectral camera. The spectral bands closest to the wavelength of a specific VI were adopted as a choice criterion. A sample of each species identified in the field and delineated in hyperspectral orthomosaics, using a false-color composition, is shown in Figure 7, as well the centroids calculated from the ITCs.

Table 4 – Vegetation indices (VIs) calculated from hyperspectral orthomosaics.

ID	Vegetation Index	Equation	Rikola Bands	Reference
NDVI	Normalized	$\frac{\rho_{750} - \rho_{650}}{\rho_{750} + \rho_{650}}$	$\frac{B21 - B11}{B21 + B11}$	Rouse et al. (1974)
	Difference			
	Vegetation Index			
ND	Normalized	$\frac{\rho_{682} - \rho_{553}}{\rho_{682} + \rho_{553}}$	$\frac{B14 - B4}{B14 + B4}$	Gandia et al. (2004); Main et al. (2011)
	Difference 682/553			
NDVIh	Normalized	$\frac{\rho_{780} - \rho_{550}}{\rho_{780} + \rho_{550}}$	$\frac{B23 - B4}{B23 + B4}$	Gitelson; Kaufman; Merzlyak, (1996); Le Maire; Francois; Dufrene, (2004)
	Difference 780/550 Green NDVI hyper			
MCARI	Modified Chlorophyll Absorption in Reflectance Index	$(\rho_{700} - \rho_{670}) - 0.2(\rho_{700} - \rho_{550}) \left(\frac{\rho_{700}}{\rho_{670}} \right)$	$(B16 - B13) - 0.2(B16 - B4) \left(\frac{B16}{B13} \right)$	Daughtry et al. (2000)
PRI	Photochemical Reflectance Index	$\frac{\rho_{535} - \rho_{565}}{\rho_{535} + \rho_{565}}$	$\frac{B3 - B5}{B3 + B5}$	Gamon; Serrano; Surfus, (1997)
PSRI	Plant Senescence Reflectance Index	$\frac{\rho_{679} - \rho_{506}}{\rho_{679} + \rho_{506}}$	$\frac{B14 - B1}{B14 + B1}$	
PSSR	Pigment Specific Simple Ratio	$\frac{\rho_{750}}{\rho_{819}}$	$\frac{B21}{B25}$	Blackburn, (1998)
		$\frac{\rho_{679}}{\rho_{670} - \rho_{780}}$	$\frac{B14}{B13 - B23}$	
RE	Red edge	$\frac{\rho_{670} - \rho_{780}}{2}$	$\frac{B13 - B23}{2}$	Clevers, (1994)
REP	Red edge position	$700 + 40 \frac{\rho_{rededge} - \rho_{700}}{\rho_{740} - \rho_{700}}$	$700 + 40 \frac{\rho_{rededge} - B16}{B20 - B16}$	Baranoski; Rokne, (2005); Clevers, (1994); Dawson; Curran, (1998)
RENDVI	Red Edge Normalized	$\frac{\rho_{753} - \rho_{700}}{\rho_{753} + \rho_{700}}$	$\frac{B21 - B16}{B21 + B16}$	Gitelson; Merzlyak; Lichtenthaler, (1996)
	Difference Vegetation Index			
SIPI	Structure Insensitive Pigment	$\frac{\rho_{800} - \rho_{500}}{\rho_{800} + \rho_{680}}$	$\frac{B24 - B1}{B24 + B1}$	Peñuelas; Baret; Filella, (1995)
	reflectance index			

Figure 7 – Individual tree crowns samples (in green) for each species identified in the field and delineated in hyperspectral orthomosaic (R: 780.49 nm; G: 650.96 nm; B: 535.09 nm). The yellow dots are the centroids calculated from the ITCs.



3.4 AUTOMATIC CLASSIFICATION

The three extracted datasets (VIs from hyperspectral orthomosaics, PCs from PR LiDAR data, and voxel-related attributes from FWF LiDAR data) were combined to assess the improvement in the accuracy of tree species classification in tropical forests, using the spectral response of tree species from hyperspectral data, combined or not with structural data from PR and FWF LiDAR data. Thus, five different scenarios were tested with these datasets (Table 4).

Table 5 – Different scenarios tested with the three different datasets with the number of features used in each test.

Scenario	Specification	Number of Features
a	VIs	11
b	PCs PR LiDAR	5
c	Attributes FWF LiDAR	19
d	VIs + PCs PR LiDAR	16
e	VIs + Attributes FWF LiDAR	24

The algorithm used for the classification of tree species was RF (random forest), which consists of several decision trees, with the class of a given sample being determined with the most frequent vote among the decision trees (BREIMAN, 2001). The algorithm randomly creates new training sets, with substitution and resampling of the original data, many times as the number of samples (BELGIU; DRĂGUȚ, 2016). The classification was performed in R environment (R CORE TEAM, 2017) using the *randomForest* package (LIAW; WIENER, 2002). The following parameters were selected: number of trees built (*ntree* = 1000) and number of candidate variables in each tree node (*mtry*), defined as the square root of the number of input data in each of the tested scenarios (Table 4). With RF it is possible to obtain the degree of importance of each one of the features used as input for the classification of tree species. Moreover, this algorithm handles data with high dimensionality in classification problems (PAL, 2005; PRASAD; IVERSON; LIAW, 2006). We avoided using the artificial neural network method (the best estimation method for forest variables, as shown in Chapter 3) since we were unsure whether this method has overestimated the estimations.

Due to the low number of samples, 50% were used for training and 50% for classification validation. This problem can be circumvented with the LOOCV (leave-one-out cross-validation) method. Miyoshi et al. (2020), Nevalainen et al. (2017) and Sothe et al. (2019) mentioned the success of this technique when working with less than ten samples per class or

for an unbalanced number of samples per class. Details on the use of LOOCV are given in Chapter 3, Section 2.6, and Chapter 4, Section 3.7.

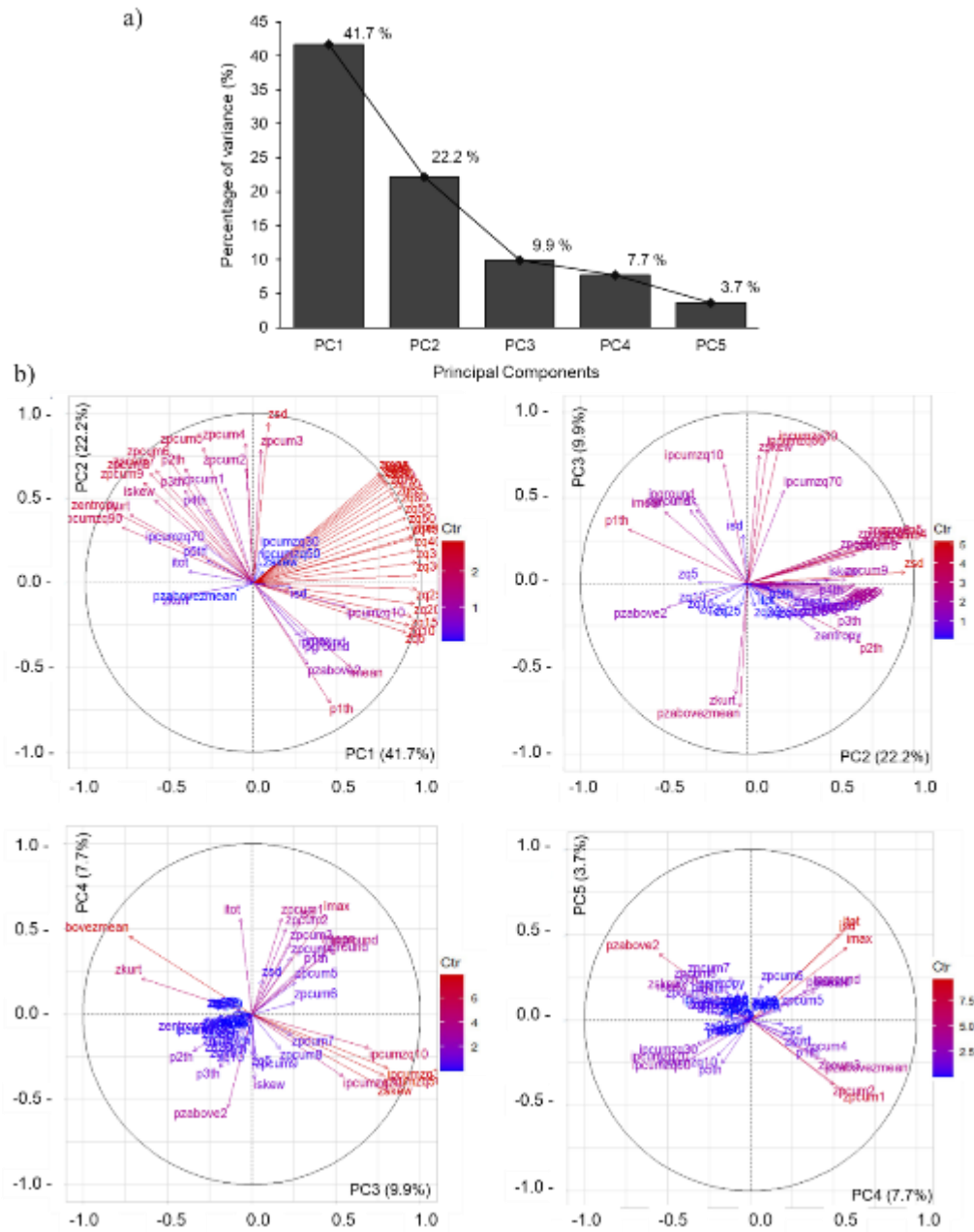
The classification evaluation was performed with the results obtained from the LOOCV process. Then, the following statistics were calculated: overall accuracy (OA), producer's accuracy (PA), user's accuracy (UA), and Cohen's Kappa index (κ) (COHEN, 1960) for each tested scenario. For the best result, the confusion matrix was calculated and the relative importance of the features that better separated the tree species was also analyzed.

4 RESULTS

4.1 PRINCIPAL COMPONENT ANALYSIS AND SPECTRAL FEATURES

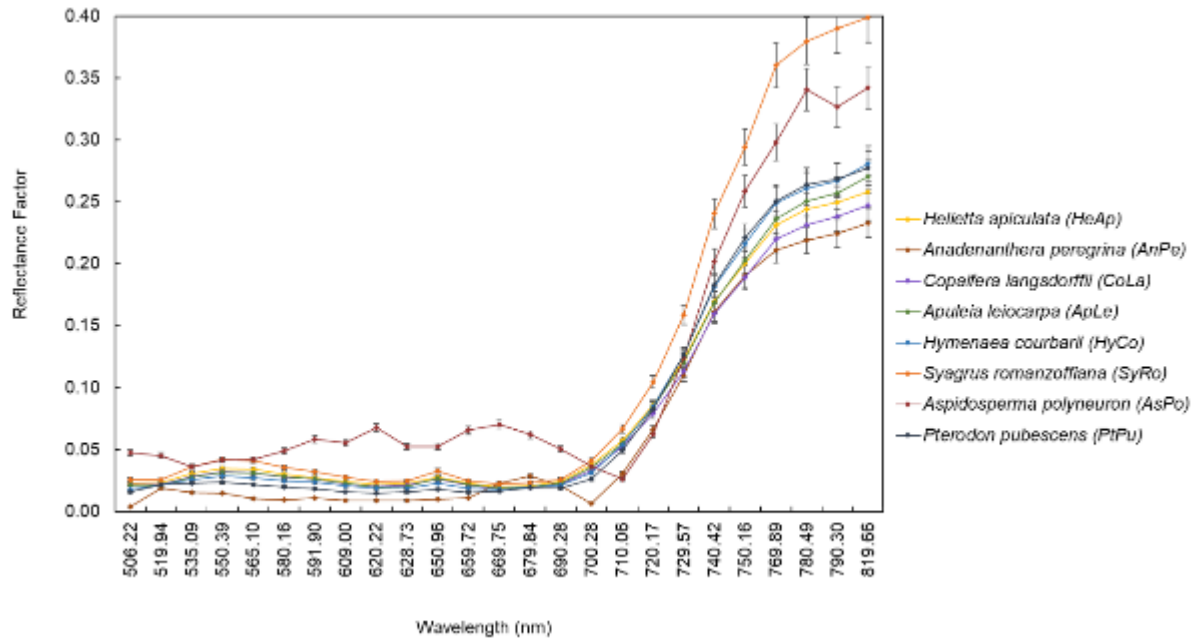
The first five PCs metrics, generated by transforming the PR LiDAR metrics with the PCA, explained 85% of the variance of the LiDAR, according to the Kaiser criterion. Figure 8 shows the percentage of variation that can be explained by each PC and the projection of each LiDAR metric with its respective contribution.

Figure 8 – (a) Percentage of variance explained by each PC. (b) Projection of PR LiDAR metrics in each PCs and their respective contribution.



An average curve was calculated for each species (Figure 9) from the spectra collected in the segments on the hyperspectral mosaic. The regions of the spectrum that best differentiate the species and the spectral characteristics of each species were observed for comparison purposes. The Appendix B (Figure B1) shows the spectra of all trees for each species.

Figure 9 – Mean spectra of each tree species with standard error bars.



It is possible to notice a similar pattern to the average spectrum of the species. A small peak in the green region (550.39 nm) was observed. A particular differentiation between the species is noticed. From 769.89 nm it is possible to observe a greater differentiation between the species spectra. However, some species did not show this pattern. For example, the AsPo species has a higher reflectance in the visible region than others, mainly between 580.16 nm and 669.75 nm. Another species with a different behaviour is AnPe: two reflectance factor peaks are observed in the visible region, at 519.94 nm, corresponding to the wavelength of green, and at 669.75 nm corresponding to the wavelength of red. Furthermore, two absorption reflectance factor peaks (700.28 nm for AnPe and 710.06 nm for AsPo) are observed at the beginning of the near infrared.

Some inferences can be made, such as the tree's level of development, crown transparency, leaf distribution, leaf senescence, St, and whether the tree is under biotic and abiotic stress conditions. Explaining the reasons of the changes in the spectral response of the species, requires further studies. The physiological condition of the trees at the time of images acquisition and during different seasons should be observed and compared. However, this kind of study is troublesome to be performed in tropical forests.

4.2 TREES SPECIES CLASSIFICATION

The classification results for the five tested scenarios (see Table 4) are shown in Table 5. The best classification scenario was "d", the combination of the VIs obtained by the hyperspectral orthomosaics with the PCs of the PR LiDAR metrics, presenting OA of 62.8%, followed by the scenario "a" (VIs) with OA of 55.8%. In the third place, it is scenario "e" (VIs plus LiDAR FWF attributes), and the worst results were achieved with scenario "b" and "c", respectively, obtained with PCs of LiDAR PR metrics (OA of 44.2%) and attributes of FWF LiDAR (OA of 41.9%). The confusion matrix for the best classification scenario is shown in Table 6.

The producer and user accuracies for each species in the different scenarios tested are also presented in Table 5. The HeAp species was not classified with any combined technique. This species showed significant variability between individuals (see Appendix B), even when located in the same plot with the same successional stage. Ferreira et al. (2016) commented that future investigations should be done to understand the spectral variations intra species in Brazilian native forests, showing that this is a common challenge when using spectral information. All trees of the AnPe species were correctly classified in three scenarios ("a", "d", "e") of the five tested, with 100% of producer's and user's accuracy, indicating that there was no confusion in the classification of this species. However, the scenarios using only LiDAR data ("b" and "c"), were not successful in classifying these trees. The ApLe species was only classified using the PCs of the PR LiDAR metrics with a PA of 33% and with a UA of 50%. All trees of the SyRo species were correctly classified in scenario "d", with a PA of 100%, but some species were wrongly classified as SyRo, so the UA was 68.8%.

Attributes extracted from FWF LiDAR data did not enable good results in classification. The only species having a good classification with this data as input was AsPo with a PA of 100%. However, some trees were wrongly classified as AsPo, thus achieving a UA of 67%.

Table 6 – Classification results for each tested scenario. The best results are highlighted.

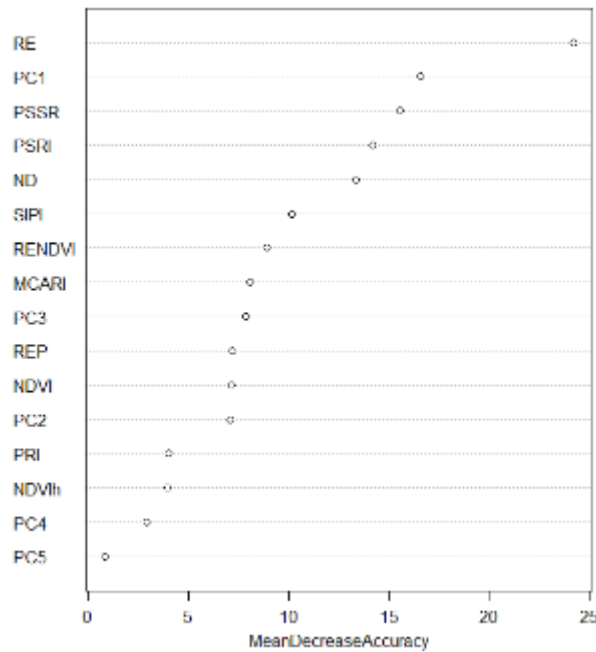
Scenario	Tree Species																OA (%)	κ
	HeAp		AnPe		CoLa		ApLe		HyCo		SyRo		AsPo		PtPu			
	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)		
a	0.0	0.0	100.0	100.0	66.7	66.7	0.0	0.0	33.3	33.3	90.9	83.3	80.0	66.7	16.7	50.0	55.8	0.48
b	0.0	0.0	0.0	0.0	33.3	28.6	33.3	50.0	66.7	44.4	81.8	50.0	40.0	66.7	16.7	33.3	44.2	0.32
c	0.0	0.0	0.0	0.0	28.6	25.0	0.0	0.0	16.7	25.0	75.0	47.4	100.0	66.7	33.3	66.7	41.9	0.28
d	0.0	0.0	100.0	100.0	50.0	75.0	0.0	0.0	66.7	57.1	100.0	68.8	80.0	66.7	33.3	66.7	62.8	0.55
e	0.0	0.0	100.0	100.0	33.3	40.0	0.0	0.0	16.7	12.5	100.0	68.8	80.0	80.0	33.3	66.7	53.5	0.44

Table 7 – Confusion matrix for the classification of the 8 tree species in the best scenario.

Species	Reference								SUM	UA (%)
	HeAp	AnPe	CoLa	ApLe	HyCo	SyRo	AsPo	PtPu		
HeAp	0	0	1	1	0	0	0	0	2	0%
AnPe	0	3	0	0	0	0	0	0	3	100%
CoLa	0	0	3	0	0	0	0	1	4	75%
ApLe	0	0	1	0	0	0	1	0	2	0%
HyCo	1	0	0	1	4	0	0	1	7	57%
SyRo	1	0	0	1	1	11	0	2	16	69%
AsPo	1	0	1	0	0	0	4	0	6	67%
PtPu	0	0	0	0	1	0	0	2	3	67%
SUM	3	3	6	3	6	11	5	6	OA = 62.8% κ = 0.55	
PA (%)	0%	100%	50%	0%	67%	100%	80%	33%		

The importance of the variables obtained from the RF classifier, in terms of MDA (mean decrease accuracy) for scenario "d", is shown in Figure 10. The reflectance factor in the red edge (RE) band, obtained from the hyperspectral orthomosaics, was the variable mainly contributing to tree species' separability. The second most important variable was the first PC, obtained from the PR LiDAR metrics.

Figure 10 – Importance of variables for classification of tree species in scenario "d".



5 DISCUSSION

In this study, we classified eight tree species in an Atlantic Forest remnant with remote sensing data. Three different datasets were tested: hyperspectral images, peak return LiDAR data, full-waveform LiDAR data and some combinations. Many studies have approached trees' species classification in temperate and subtropical forests using these data isolated or combined (FASSNACHT et al., 2016; KOENIG; HÖFLE, 2016; MICHALOWSKA; RAPIŃSKI, 2021; WAN et al., 2021), but few studies were performed to classify tree species in Brazilian tropical forests, mainly due to the difficulty of access, difficulty of obtaining a sufficient number of samples, huge heterogeneity, and diversity of species in these forests.

To classify emerging tree species in the *Ponte Branca* forest remnant (the same study area). Miyoshi et al. (2020) used hyperspectral data from the Rikola camera aboard an UAV,

collected on three dates to understand whether multitemporal data can improve the classification. The variables used were the normalized and non-normalized species spectra, and the algorithm for classification was RF. The same problems regarding low sample sufficiency were reported. The use of temporal spectral information improved the performance in three of the eight species analyzed, and the best result was obtained with the normalized spectra (OA of 50%). Our results using the hyperspectral images was slightly better (55.8%), probably because we have used VIs in the regions of the spectrum where species are better differentiated. In addition, most of our samples were obtained on the same date (either in 2016 or in 2017), so there weren't significant differences for most species. The species *Apuleia leiocarpa* (ApLe) was not classified with the hyperspectral images in Miyoshi et al. (2020), as in our case, probably because this species in winter loses all its leaves, and its spectral response is affected by branches, understory and terrain. The most important features for classification were the wavelength of 628.73 nm and the beginning of the red-edge, which confirms our results, showing that the red-edge band has a fundamental role in species discrimination.

In another inland Atlantic forest remnant in the state of São Paulo - Brazil, Ferreira et al. (2016) classified eight forest species using airborne hyperspectral images obtained with the AisaEAGLE sensor in the VNIR (visible-near infrared) region and AisaHAWK sensor in the SWIR (short-wave infrared) region. Various combinations of wavelengths and vegetation indices were tested with four different classification methods. Species discrimination had the best performance using visible bands (mainly wavelengths located at 550 nm and 650 nm) and SWIR bands. The VIs contributed positively to the classification when integrated with the VNIR features and should be used if the sensor does not obtain data in the SWIR region. The PSRI vegetation index (see Table 3) was one of the most important for species differentiation, and in our case, it had the fourth-highest relative importance (see Figure 10). Among the classifiers tested, the RF presented an OA of 67.7%, and the best classifier was the linear discriminant analysis (LDA) with an OA of 90.1%. The authors reported that the RF could not choose the best features to separate the tree species. The results of this study are better than ours, first due to higher sampling sufficiency (273 samples of 8 species were collected) while we collected 89 samples of 8 species. Furthermore, the sensor used has a higher spectral resolution, acquiring more wavelengths that allowed a better spectral characterization of forest species than the lightweight Rikola camera used in our study.

It was possible to verify that data from narrow band sensors are an important tool in the discrimination of tropical species, since it is possible to obtain specific wavelengths and

vegetation indices in regions of the spectrum where it is possible to differentiate species according to the spectral response. LiDAR geometric data also have potential for discriminating tree species, but its isolated use does not improve classification accuracy. The use of PCs of PR LiDAR metrics and FWF LiDAR data attributes did not show satisfactory results (OA of 44.2% and 41.9% respectively), with kappa index classified as reasonable, according to the classification proposed by Landis; Koch, (1977). Michalowska and Rapiński, (2021) commented that using only geometric features from LiDAR (i.e., height distribution) can decrease the classification accuracy. Considering only the vertical structure of the vegetation, they are not species-specific, as the same species can have different structures, or different species can present morphological similarities. Furthermore, tropical forests have multiple layers with smaller trees below the canopy, and, thus, LiDAR height distributions are ineffective for species separability (MICHALOWSKA; RAPIŃSKI, 2021; YOU et al., 2020).

FWF LiDAR metrics have great potential for classification of tree species. Some authors, such as Cao et al. (2016); Reitberger; Krzystek; Stilla, (2008), used metrics related to the number of waveform peaks, waveform distance, height of median energy, roughness and return of waveform energy. Hollaus et al. (2009) used FWF LiDAR metrics related to echo height distributions, mean and standard deviation of echo widths, mean intensities and backscatter cross-sections. In China's subtropical forests, the OA was 68.6% for classification of six species (CAO et al., 2016). Reitberger, Krzystek and Stilla, (2008) found an OA of 79% for classification of leaf-on trees and OA of 95% for classification of leaf-off trees in Bavarian Forest National Park in Germany. In pre-Alps region of Austria, three species (beech, spruce and larch) were classified with an OA of 85% by Hollaus et al. (2009). None of these studies were performed in tropical forests, and the attributes we have used, extracted from the FWF data, were also different. Further studies should be performed changing the parameters present in the software DASOS (e.g. test different window sizes) to investigate how they affect the results in classification and assess the improvement in the accuracy of tree species classification.

LiDAR geometric data, especially when combined with radiometric data, intensity and spectral data, provides valuable information for the classification of species in complex environments (FASSNACHT et al., 2016; MICHALOWSKA; RAPIŃSKI, 2021). Using heights extracted from LiDAR data, Asner et al., (2008) and Cho et al. (2012) reported improvements in tree species classification in African savannas and Hawaiian forests, respectively, with OA of 79% and 91%. Other studies using LiDAR metrics, related to height

and canopy cover, have reported improvements of approximately 6% in species classification accuracy when combined with hyperspectral data. Such studies were performed in more diverse environments such as subtropical forests in China (SHEN; CAO, 2017) and agroforestry systems in Africa (PIIROINEN et al., 2017).

In our study, the use of PC metrics, extracted from PR LiDAR, combined with VIs from the hyperspectral images, showed 7% of improvement compared to the classification using only hyperspectral images. This result indicates that even in tropical forests, combining geometric data from LiDAR with the spectral information from images provides a better dataset for classifying species in complex forests.

Instead of using the PR LiDAR metrics in their raw format, we have used transformed metrics with PCA. We performed some preliminary studies that showed that using metrics in their raw form, or a previous selection by correlation analysis did not improve the classifications results. Figure 8b shows the contribution of each PR LiDAR metric in the principal components. It can be seen that the metrics related to the percentiles of heights (zq^{th}) and the cumulative return in elevations distributions ($zpcumx$) were the metric with the highest contribution to the first PC, which was the one that best differentiated species (Figure 10). In their raw format, these metrics are highly correlated, but when we transformed them by PCA, the information contained in the metrics was merged and uncorrelated, providing very valuable data for a complex forest, such as the object of our study.

6 CONCLUSIONS

In this study, we investigated the use of three different datasets for the classification of eight forest species present in the upper canopy of a forest remnant from the Brazilian Atlantic Forest. The datasets used were vegetation indices (VIs), extracted from hyperspectral images, peak returns (PR), transformed from principal component analysis (PCA), and attributes extracted from voxelized space with full-waveform (FWF) LiDAR data. These data were used independently or combined as input variables for classification.

Among the tested combinations, the use of isolated PR or FWF LiDAR data was the scenario that classified the species with the lowest overall accuracy (44.2% PR data and 41.9% with FWF data). The best results were achieved with a scenario that used VIs with the PR LiDAR metrics, transformed by PCA, with an overall accuracy of 62.8%. The PCs provided

relevant information, as they gather information from all LiDAR metrics. In addition, this transformation produced less correlated metrics.

Due to the absence of leaves, the species *Apuleia leiocarpa* was not classified in any scenario, except using the PC of the LiDAR PR metrics. The species *Helietta apiculata* was not classified in any scenario, due to its significant spectral variability among individuals.

The reflectance factor in the red-edge (RE) and the Pigment Specific Simple Ratio (PSSR) vegetation index were the attributes from the hyperspectral images that best differentiated the species, together with the first PC of PR LiDAR. For the latter, the most significant contribution were the metrics related to the height distributions.

Further studies should be conducted to investigate more FWF metrics and the use of intensities values, as well as the use of original spectral data, instead of vegetation indices, to improve classification accuracy. In addition, the development and improvement of segmentation techniques for the tropical forest environment to automate the process of tree crown delineation and extraction of features from different datasets.

Despite the difficulties observed in this study, regarding the low sampling sufficiency of trees, and the time lag between the different flight campaigns, the classification results can be considered satisfactory for the complex forest environment studied. These results can serve for management and conservation practices of these forest remnants. The mapping of these species allows the knowledge of the spatial distribution of species with potential for forest restoration, as well as for production and seeds holder.

7 REFERENCES

- ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley interdisciplinary reviews: computational statistics**, [s. l.], v. 2, n. 4, p. 433–459, 2010.
- ASD. FieldSpec UV/VNIR Handheld Spectroradiometer User's Guide. **Boulder CO**, [s. l.], 2002.
- ASNER, G. P.; KNAPP, D. E.; KENNEDY-BOWDOIN, T.; JONES, M. O.; MARTIN, R. E.; BOARDMAN, J.; HUGHES, R. F. Invasive species detection in Hawaiian rainforests using airborne imaging spectroscopy and LiDAR. **Remote sensing of Environment**, [s. l.], v. 112, n. 5, p. 1942–1955, 2008.
- ASNER, G. P.; PALACE, M.; KELLER, M.; PEREIRA JR, R.; SILVA, J. N.; ZWEEDE, J. C. Estimating canopy structure in an Amazon forest from laser range finder and IKONOS satellite observations 1. **Biotropica**, [s. l.], v. 34, n. 4, p. 483–492, 2002.
- BARANOSKI, G. V. G.; ROKNE, J. G. A practical approach for estimating the red edge position of plant leaf reflectance. **International Journal of Remote Sensing**, [s. l.], v. 26, n. 3, p. 503–521, 2005.
- BAUGH, W. M.; GROENEVELD, D. P. Empirical proof of the empirical line. **International Journal of Remote Sensing**, [s. l.], v. 29, n. 3, p. 665–672, 2008.
- BELGIU, M.; DRĂGUȚ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS journal of photogrammetry and remote sensing**, [s. l.], v. 114, p. 24–31, 2016.
- BLACKBURN, G. A. Spectral indices for estimating photosynthetic pigment concentrations: a test using senescent tree leaves. **International Journal of remote sensing**, [s. l.], v. 19, n. 4, p. 657–675, 1998.
- BREIMAN, L. Random forests, machine learning 45. **J. Clin. Microbiol**, [s. l.], v. 2, n. 30, p. 199–228, 2001.
- BUDDENBAUM, H.; SEELING, S.; HILL, J. Fusion of full-waveform lidar and imaging spectroscopy remote sensing data for the characterization of forest stands. **International Journal of Remote Sensing**, [s. l.], v. 34, n. 13, p. 4511–4524, 2013.
- CAO, L.; COOPS, N. C.; INNES, J. L.; DAI, J.; RUAN, H.; SHE, G. Tree species classification in subtropical forests using small-footprint full-waveform LiDAR data. **International journal of applied earth observation and geoinformation**, [s. l.], v. 49, p. 39–51, 2016.
- CARVALHO, P. E. R. **Espécies arbóreas brasileiras**. Brasília, DF: Embrapa Informação Tecnológica Brasília, 2003. v. 1
- CARVALHO, P. E. R. **Espécies arbóreas brasileiras**. Brasília, DF: Embrapa Informação Tecnológica Brasília, 2006. v. 2

CARVALHO, P. E. R. **Espécies arbóreas brasileiras**. Brasília, DF: Embrapa Informação Tecnológica Brasília, 2008. v. 3

CARVALHO, P. E. R. **Espécies arbóreas brasileiras**. Brasília, DF: Embrapa Informação Tecnológica Brasília, 2014. v. 5

CHO, M. A.; MATHIEU, R.; ASNER, G. P.; NAIDOO, L.; VAN AARDT, J. A. N.; RAMOELO, A.; DEBBA, P.; WESSELS, K.; MAIN, R.; SMIT, I. P. Mapping tree species composition in South African savannas using an integrated airborne spectral and LiDAR system. **Remote Sensing of Environment**, [s. l.], v. 125, p. 214–226, 2012.

CLEVERS, J. G. Imaging spectrometry in agriculture-plant vitality and yield indicators. In: **Imaging spectrometry—A tool for environmental observations**. [s.l.] : Springer, 1994. p. 193–219.

CLIFF, N. The eigenvalues-greater-than-one rule and the reliability of components. **Psychological bulletin**, [s. l.], v. 103, n. 2, p. 276, 1988.

COCHRANE, M. A. Using vegetation reflectance variability for species level classification of hyperspectral data. **International journal of remote sensing**, [s. l.], v. 21, n. 10, p. 2075–2087, 2000.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, [s. l.], v. 20, n. 1, p. 37–46, 1960.

DAUGHTRY, C. S.; WALTHALL, C. L.; KIM, M. S.; DE COLSTOUN, E. B.; MCMURTREY III, J. E. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. **Remote sensing of Environment**, [s. l.], v. 74, n. 2, p. 229–239, 2000.

DAWSON, T. P.; CURRAN, P. J. Technical note A new technique for interpolating the reflectance red edge position. [s. l.], 1998.

FASSNACHT, F. E.; LATIFI, H.; STERENĆZAK, K.; MODZELEWSKA, A.; LEFSKY, M.; WASER, L. T.; STRAUB, C.; GHOSH, A. Review of studies on tree species classification from remotely sensed data. **Remote Sensing of Environment**, [s. l.], v. 186, p. 64–87, 2016.

FÉRET, J.-B.; ASNER, G. P. Tree species discrimination in tropical forests using airborne imaging spectroscopy. **IEEE Transactions on Geoscience and Remote Sensing**, [s. l.], v. 51, n. 1, p. 73–84, 2012.

FERREIRA, M. P.; ZORTEA, M.; ZANOTTA, D. C.; SHIMABUKURO, Y. E.; DE SOUZA FILHO, C. R. Mapping tree species in tropical seasonal semi-deciduous forests with hyperspectral and multispectral data. **Remote Sensing of Environment**, [s. l.], v. 179, p. 66–78, 2016.

GAMON, Ja.; SERRANO, L.; SURFUS, J. S. The photochemical reflectance index: an optical indicator of photosynthetic radiation use efficiency across species, functional types, and nutrient levels. **Oecologia**, [s. l.], v. 112, n. 4, p. 492–501, 1997.

GANDIA, S.; FERNÁNDEZ, G.; GARCÍA, J. C.; MORENO, J. Retrieval of vegetation biophysical variables from CHRIS/PROBA data in the SPARC campaign. **Esa Sp**, [s. l.], v. 578, p. 40–48, 2004.

GITELSON, A. A.; KAUFMAN, Y. J.; MERZLYAK, M. N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. **Remote sensing of Environment**, [s. l.], v. 58, n. 3, p. 289–298, 1996.

GITELSON, A. A.; MERZLYAK, M. N.; LICHTENTHALER, H. K. Detection of red edge position and chlorophyll content by reflectance measurements near 700 nm. **Journal of plant physiology**, [s. l.], v. 148, n. 3–4, p. 501–508, 1996.

HOLLAUS, M.; MÜCKE, W.; HÖFLE, B.; DORIGO, W.; PFEIFER, N.; WAGNER, W.; BAUERHANS, C.; REGNER, B. Tree species classification based on full-waveform airborne laser scanning data. **Proceedings of SILVILASER**, [s. l.], p. 14–16, 2009.

HONGYU, K.; SANDANIELO, V. L. M.; DE OLIVEIRA JUNIOR, G. J. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and Science**, [s. l.], v. 5, n. 1, p. 83–90, 2016.

HONKAVAARA, E.; HAKALA, T.; SAARI, H.; MARKELIN, L.; MÄKYNEN, J.; ROSNELL, T. A process for radiometric correction of UAV image blocks. **Photogrammetrie, Fernerkundung, Geoinformation**, [s. l.], n. 2, p. 115–127, 2012. a.

HONKAVAARA, E.; KAIVOSOJA, J.; MÄKYNEN, J.; PELLIKKA, I.; PESONEN, L.; SAARI, H.; SALO, H.; HAKALA, T.; MARKLELIN, L.; ROSNELL, T. Hyperspectral reflectance signatures and point clouds for precision agriculture by light weight UAV imaging system. **ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci**, [s. l.], v. 7, p. 353–358, 2012. b.

HONKAVAARA, E.; ROSNELL, T.; OLIVEIRA, R.; TOMMASELLI, A. Band registration of tuneable frame format hyperspectral UAV imagers in complex scenes. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 134, p. 96–109, 2017.

HONKAVAARA, E.; SAARI, H.; KAIVOSOJA, J.; PÖLÖNEN, I.; HAKALA, T.; LITKEY, P.; MÄKYNEN, J.; PESONEN, L. Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight UAV spectral camera for precision agriculture. **Remote Sensing**, [s. l.], v. 5, n. 10, p. 5006–5039, 2013.

ISENBURG, M. **LAStools-efficient LiDAR processing software**. 2014. Disponível em: <<http://lastools.org/>>. Acesso em: 12 nov. 2020.

JENSEN, J. R. **Remote sensing of the environment: An earth resource perspective 2/e**. [s.l.] : Pearson Education India, 2009.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, [s. l.], v. 23, n. 3, p. 187–200, 1958.

KIM, S.; MCGAUGHEY, R. J.; ANDERSEN, H.-E.; SCHREUDER, G. Tree species differentiation using intensity data derived from leaf-on and leaf-off airborne laser scanner data. **Remote Sensing of Environment**, [s. l.], v. 113, n. 8, p. 1575–1586, 2009.

- KOENIG, K.; HÖFLE, B. Full-Waveform airborne laser scanning in vegetation studies—A review of point cloud and waveform features for tree species classification. **Forests**, [s. l.], v. 7, n. 9, p. 198, 2016.
- LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, [s. l.], p. 159–174, 1977.
- LE MAIRE, G.; FRANCOIS, C.; DUFRENE, E. Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements. **Remote sensing of environment**, [s. l.], v. 89, n. 1, p. 1–28, 2004.
- LÊ, S.; JOSSE, J.; HUSSON, F. FactoMineR: an R package for multivariate analysis. **Journal of statistical software**, [s. l.], v. 25, n. 1, p. 1–18, 2008.
- LIAW, A.; WIENER, M. Classification and regression by randomForest. **R news**, [s. l.], v. 2, n. 3, p. 18–22, 2002.
- LUO, S.; WANG, C.; XI, X.; ZENG, H.; LI, D.; XIA, S.; WANG, P. Fusion of airborne discrete-return LiDAR and hyperspectral data for land cover classification. **Remote Sensing**, [s. l.], v. 8, n. 1, p. 3, 2016.
- MAIN, R.; CHO, M. A.; MATHIEU, R.; O’KENNEDY, M. M.; RAMOELO, A.; KOCH, S. An investigation into robust spectral indices for leaf chlorophyll estimation. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 66, n. 6, p. 751–761, 2011.
- MÄKELÄINEN, A.; SAARI, H.; HIPPI, I.; SARKEALA, J.; SOUKKAMÄKI, J. 2D Hyperspectral Frame Imager Camera Data in Photogrammetric Mosaicking. **ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], p. 263–267, 2013.
- MANLY, B. F.; ALBERTO, J. A. N. **Multivariate statistical methods: a primer**. [s.l.] : CRC press, 2016.
- MARISCAL-FLORES, E. J. **Potencial produtivo e alternativas de manejo sustentável de um fragmento de Mata Atlântica secundária, Município de Viçosa, Minas Gerais**. 1993. Master thesis (Master’s in Forest Science) - Universidade Federal de Viçosa., Viçosa - MG, 1993.
- MARTINS-NETO, R. P.; TOMMASELLI, A. M. G.; IMAI, N. N.; DAVID, H. C.; MILTIADOU, M.; HONKAVAARA, E. Identification of Significant LiDAR Metrics and Comparison of Machine Learning Approaches for Estimating Stand and Diversity Variables in Heterogeneous Brazilian Atlantic Forest. **Remote Sensing**, [s. l.], v. 13, n. 13, p. 2444, 2021.
- MERZLYAK, M. N.; GITELSON, A. A.; CHIVKUNOVA, O. B.; RAKITIN, V. Y. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. **Physiologia plantarum**, [s. l.], v. 106, n. 1, p. 135–141, 1999.
- MICHALOWSKA, M.; RAPIŃSKI, J. A Review of Tree Species Classification Based on Airborne LiDAR Data and Applied Classifiers. **Remote Sensing**, [s. l.], v. 13, n. 3, p. 353, 2021.

MILLIKAN, P. H. K.; SILVA, C. A.; RODRIGUEZ, L. C. E.; OLIVEIRA, T. M. De; CARVALHO, M. P. de L. C.; CARVALHO, S. de P. C. Automated individual tree detection in amazon tropical forest from airborne laser scanning data. **Cerne**, [s. l.], v. 25, p. 273–282, 2019.

MILTIADOU, M.; AGAPIOU, A.; GONZALEZ ARACIL, S.; HADJIMITSIS, D. G. Detecting Dead Standing Eucalypt Trees from Voxelised Full-Waveform Lidar Using Multi-Scale 3D-Windows for Tackling Height and Size Variations. **Forests**, [s. l.], v. 11, n. 2, p. 161, 2020.

MILTIADOU, M.; CAMPBELL, N. D.; ARACIL, S. G.; BROWN, T.; GRANT, M. G. Detection of dead standing Eucalyptus camaldulensis without tree delineation for managing biodiversity in native Australian forest. **International journal of applied earth observation and geoinformation**, [s. l.], v. 67, p. 135–147, 2018.

MILTIADOU, M.; GRANT, M. G.; CAMPBELL, N. D.; WARREN, M.; CLEWLEY, D.; HADJIMITSIS, D. G. Open source software DASOS: Efficient accumulation, analysis, and visualisation of full-waveform lidar. In: SEVENTH INTERNATIONAL CONFERENCE ON REMOTE SENSING AND GEOINFORMATION OF THE ENVIRONMENT (RSCY2019) 2019, **Anais...** : International Society for Optics and Photonics, 2019.

MIYOSHI, G. T. **Caracterização espectral de espécies de Mata Atlântica de Interior em nível foliar e de copa**. 2016. Master thesis (Master's in Cartographic Sciences) - Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia, Presidente Prudente - SP, 2016.

MIYOSHI, G. T.; IMAI, N. N.; GARCIA TOMMASELLI, A. M.; ANTUNES DE MORAES, M. V.; HONKAVAARA, E. Evaluation of hyperspectral multitemporal information to improve tree species identification in the highly diverse atlantic forest. **Remote Sensing**, [s. l.], v. 12, n. 2, p. 244, 2020.

MIYOSHI, G. T.; IMAI, N. N.; TOMMASELLI, A. M. G.; HONKAVAARA, E.; NÄSI, R.; MORIYA, É. A. S. Radiometric block adjustment of hyperspectral image blocks in the Brazilian environment. **International journal of remote sensing**, [s. l.], v. 39, n. 15–16, p. 4910–4930, 2018.

MORIYA, E. A. S.; IMAI, N. N.; TOMMASELLI, A. M. G.; MIYOSHI, G. T. Mapping mosaic virus in sugarcane based on hyperspectral images. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, [s. l.], v. 10, n. 2, p. 740–748, 2017.

MORSDORF, F.; MÅRELL, A.; KOETZ, B.; CASSAGNE, N.; PIMONT, F.; RIGOLOT, E.; ALLGÖWER, B. Discrimination of vegetation strata in a multi-layered Mediterranean forest ecosystem using height and intensity information derived from airborne laser scanning. **Remote Sensing of Environment**, [s. l.], v. 114, n. 7, p. 1403–1415, 2010.

NÄSI, R.; HONKAVAARA, E.; LYYTIKÄINEN-SAARENMAA, P.; BLOMQVIST, M.; LITKEY, P.; HAKALA, T.; VILJANEN, N.; KANTOLA, T.; TANHUANPÄÄ, T.; HOLOPAINEN, M. Using UAV-based photogrammetry and hyperspectral imaging for mapping bark beetle damage at tree-level. **Remote Sensing**, [s. l.], v. 7, n. 11, p. 15467–15493, 2015.

NEVALAINEN, O.; HONKAVAARA, E.; TUOMINEN, S.; VILJANEN, N.; HAKALA, T.; YU, X.; HYYPPÄ, J.; SAARI, H.; PÖLÖNEN, I.; IMAI, N. N. Individual tree detection and classification with UAV-based photogrammetric point clouds and hyperspectral imaging. **Remote Sensing**, [s. l.], v. 9, n. 3, p. 185, 2017.

OLIVEIRA, R. A.; TOMMASELLI, A. M.; HONKAVAARA, E. Geometric calibration of a hyperspectral frame camera. **The Photogrammetric Record**, [s. l.], v. 31, n. 155, p. 325–347, 2016.

OLIVEIRA, R. A.; TOMMASELLI, A. M.; HONKAVAARA, E. Generating a hyperspectral digital surface model using a hyperspectral 2D frame camera. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 147, p. 345–360, 2019.

PAL, M. Random forest classifier for remote sensing classification. **International Journal of Remote Sensing**, [s. l.], v. 26, n. 1, p. 217–222, 2005.

PEÑUELAS, J.; BARET, F.; FILELLA, I. Semi-empirical indices to assess carotenoids/chlorophyll a ratio from leaf spectral reflectance. **Photosynthetica**, [s. l.], v. 31, n. 2, p. 221–230, 1995.

PIIROINEN, R.; HEISKANEN, J.; MAEDA, E.; VIINIKKA, A.; PELLIKKA, P. Classification of tree species in a diverse African agroforestry landscape using imaging spectroscopy and laser scanning. **Remote Sensing**, [s. l.], v. 9, n. 9, p. 875, 2017.

PONZONI, F. J.; SHIMABUKURO, Y. E.; KUPLICH, T. M. Sensoriamento remoto da vegetação (Remote sensing of vegetation). [s. l.], 2012.

PRASAD, A. M.; IVERSON, L. R.; LIAW, A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. **Ecosystems**, [s. l.], v. 9, n. 2, p. 181–199, 2006.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2017.

REITBERGER, J.; KRZYSZEK, P.; STILLA, U. Analysis of full waveform LIDAR data for the classification of deciduous and coniferous trees. **International journal of remote sensing**, [s. l.], v. 29, n. 5, p. 1407–1431, 2008.

RICHARDS, J. A.; JIA, X. **Remote sensing digital image analysis**. 4. ed. Berlin, Germany: Springer, 2006.

ROUSE, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the Great Plains with ERTS. **NASA special publication**, [s. l.], v. 351, n. 1974, p. 309, 1974.

ROUSSEL, J.-R.; AUTY, D.; DE BOISSIEU, F.; MEADOR, A. S. **lidR: Airborne LiDAR data manipulation and visualization for forestry applications**. 2018. Disponível em: <<https://rdr.io/cran/lidR/>>. Acesso em: 21 jan. 2021.

SAARI, H.; AALLOS, V.-V.; AKUJÄRVI, A.; ANTILA, T.; HOLMLUND, C.; KANTOJÄRVI, U.; MÄKYNEN, J.; OLLILA, J. Novel miniaturized hyperspectral sensor for

UAV and space applications. In: SENSORS, SYSTEMS, AND NEXT-GENERATION SATELLITES XIII 2009, **Anais...** : International Society for Optics and Photonics, 2009.

SHEN, X.; CAO, L. Tree-species classification in subtropical forests using airborne hyperspectral and LiDAR data. **Remote Sensing**, [s. l.], v. 9, n. 11, p. 1180, 2017.

SOTHE, C.; DALPONTE, M.; ALMEIDA, C. M. De; SCHIMALSKI, M. B.; LIMA, C. L.; LIESENBERG, V.; MIYOSHI, G. T.; TOMMASELLI, A. M. G. Tree Species Classification in a Highly Diverse Subtropical Forest Integrating UAV-Based Photogrammetric Point Cloud and Hyperspectral Data. **Remote Sensing**, [s. l.], v. 11, n. 11, p. 1338, 2019.

SOUZA, D. R. De; SOUZA, A. L. De; GAMA, J. R. V.; LEITE, H. G. Emprego de análise multivariada para estratificação vertical de florestas inequidâneas. **Revista Árvore**, [s. l.], v. 27, p. 59–63, 2003.

TOCHON, G.; FERET, J.-B.; VALERO, S.; MARTIN, R. E.; KNAPP, D. E.; SALEMBIER, P.; CHANUSSOT, J.; ASNER, G. P. On the use of binary partition trees for the tree crown segmentation of tropical rainforest hyperspectral images. **Remote sensing of environment**, [s. l.], v. 159, p. 318–331, 2015.

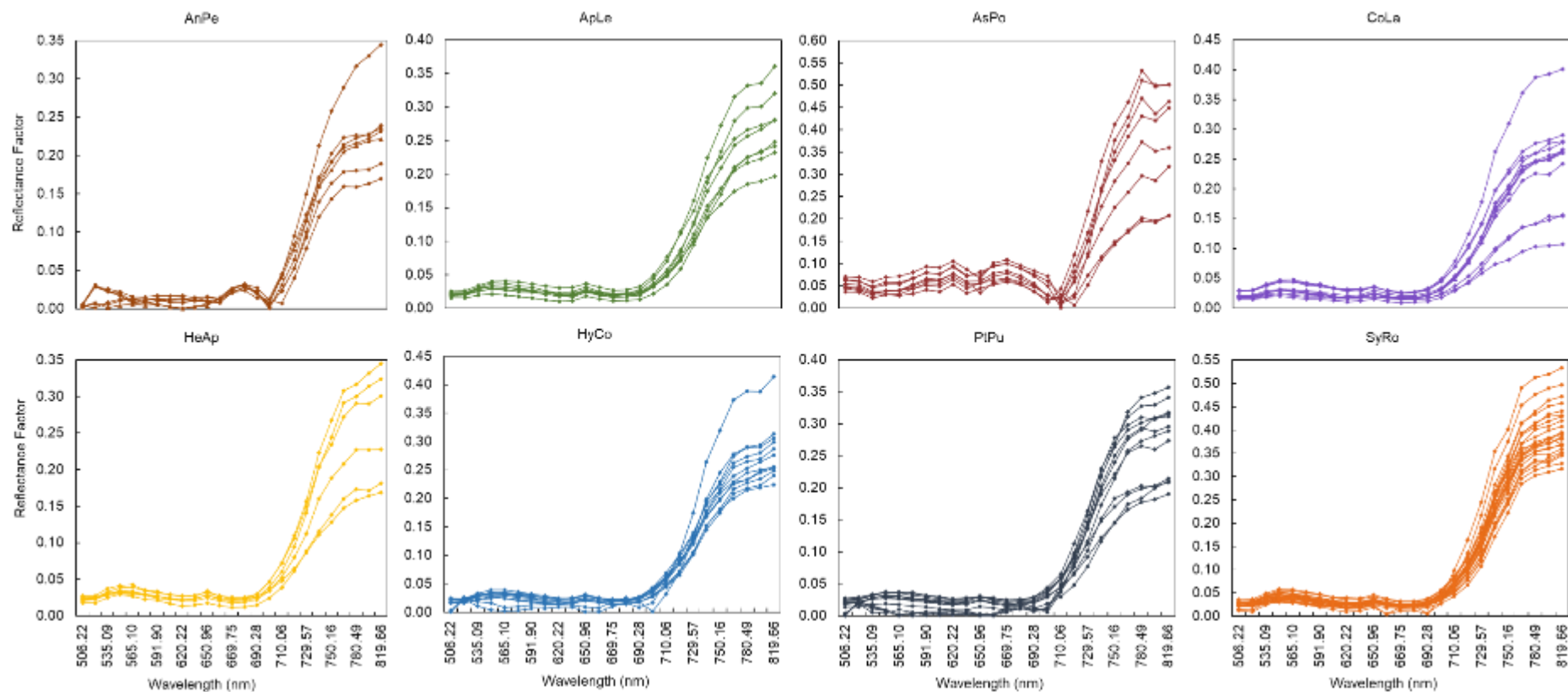
WAN, H.; TANG, Y.; JING, L.; LI, H.; QIU, F.; WU, W. Tree species classification of forest stands using multisource remote sensing data. **Remote Sensing**, [s. l.], v. 13, n. 1, p. 144, 2021.

YOU, H. T.; LEI, P.; LI, M. S.; RUAN, F. Q. Forest Species Classification Based on Three-dimensional Coordinate and Intensity Information of Airborne LiDAR Data with Random Forest Method. **The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [s. l.], v. 42, p. 117–123, 2020.

ZHANG, J.; RIVARD, B.; SÁNCHEZ-AZOFEIFA, A.; CASTRO-ESAU, K. Intra-and inter-class spectral variability of tropical tree species at La Selva, Costa Rica: Implications for species identification using HYDICE imagery. **Remote Sensing of Environment**, [s. l.], v. 105, n. 2, p. 129–141, 2006.

APPENDIX B – CHAPTER 5

Figure B1 – Spectra obtained from each tree for each species, from hyperspectral orthomosaics.



CHAPTER 6 – FINAL REMARKS

Knowledge of the structure of the few remnants of the Atlantic Forest is necessary for decision making and the development of conservation and sustainable management plans to mitigate the effects of degradation over time. This work shows the potential of using high-resolution Remote Sensing to estimate structural variables in a small remnant of inland Atlantic Forest. The performance and effectiveness of peak return and full waveform LiDAR data for estimation of stand and diversity variables and the combination of these data with hyperspectral images for classification of tree species were assessed. This final chapter discusses the highlights, open problems, and recommendations for future studies based on the results achieved.

(1) The study area, the *Ponte Branca* forest remnant, was characterized in Chapter 2, and is a transition area between the Atlantic Forest and Cerrado biomes (with the presence of tree species typical of the Cerrado). The species accumulation curve presents an almost stability with 15 plots, that is, increasing the number of plots does not significantly increase the number of sampled species. Even though a larger number of surveyed plots is indicated, mainly for applications using machine learning approaches for variables estimation, as presented in the other chapters. Instead of using a few plots with large dimensions, decreasing the plot size and increasing the number of samples is more appropriate. Furthermore, the rectangular shape better captures the heterogeneity of the area than the square shape. For studies that do not aim at repeated measurements over time, the use of circular plots or variable area sampling methods (e.g., quadrant method) is also indicated.

(2) For the use of LiDAR metrics as input data in models to estimate forest variables, a pre-selection of metrics is necessary due to the high correlation and to reduce the dimensionality of the data. The transformation of the 54 peak returns LiDAR metrics related to statistics of heights, intensities and pulse returns by Principal Component Analysis (PCA) was effective for estimating seven stand and diversity variables (Chapter 3). Each Principal Component (PC) was able to synthesize relevant information from the set of LiDAR metrics. Among the estimation methods tested, artificial neural networks were the method that produced the best results. However, to achieve these results, we used complex architectures (with two and three hidden layers) that can result in overfitting. The tested architectures must be used with a larger database to check for overfitting. Furthermore, even with the best results, the variables basal area (BA), density (DEN) and the number of tree species (NTS) showed high estimation

errors due to the great variability and difficulty in estimating these variables especially in tropical forests.

(3) A voxel-based approach was used to verify a possible improvement in estimations of the BA, DEN, NTS and aboveground biomass (AGB) variables (Chapter 4) not only for the peak return LiDAR also for full waveform data. Two different flight heights (900 m and 2000 m) were also tested to understand how they influence the estimations. Low-level filtering is needed to eliminate noise signals in FWF data. The chosen low-level filtering threshold is site-specific and depends on the sensor and flight height used. Variable estimations using attributes extracted from waveforms were less biased than PR estimates. There were no significant differences regarding flight height, but lower flight heights require higher filtering thresholds as the sensor records more noisy signals. The results indicate a potential for the use of FWF data in tropical forests and the voxel-based approach available in the DASOS software, as there was an improvement in the errors in the estimates of the four variables, when compared to traditional peak return metrics. For future studies, combine the metrics related to the distribution of heights, intensity, and pulse return from peak return LiDAR data with the attributes extracted from voxelized space of FWF data, to verify improvements in estimations.

(4) The classification of tree species is a complicated task, especially in tropical forests due to the great heterogeneity even among species. In Chapter 5 we tested the combinations of peak return and full waveform LiDAR data with hyperspectral images to understand how geometric information from LiDAR and spectral information from hyperspectral images improve classification accuracy. Among the tested scenarios, the combination of vegetation indices extracted from hyperspectral images with peak return LiDAR metrics, transformed by PCA, resulted in the best classification accuracy. The delineation of the tree crowns was performed manually as the great variability in the tree crowns makes the automatic segmentation process difficult and challenging. For future studies, adapt and develop techniques for automatic segmentation of trees in tropical forests to automate the process of extracting features from individual trees.

(5) Unlike what is found in literature, full waveform LiDAR data used isolated or combined with hyperspectral images were not effective in the classification of tree species, with low accuracy (Chapter 5). We have used different sets of full waveform metrics than other studies, so new tests must be performed by changing the parameters of the software DASOS, like the size of 3D window for extracting attributes and normalizing waveform samples intensities, to verify improvements in classification.

(6) As the forest remnants of the Atlantic Forest are very fragmented and the areas are small, many of these areas have several successional stages, from areas of initial regeneration to areas of mature forest. Therefore, it is recommended to perform field campaigns simultaneously or as close as possible, to the flight campaigns, mainly due to the tree species' phenology and the forest dynamics. The temporal lag between the different collection dates, can result in both geometric and spectral differences.

(7) Finally, we suggest to apply the methods used in this thesis to other forest remnants and larger areas of tropical forest (e.g. Amazon forest) and verify if the accuracy obtained in the estimates and classification is maintained in different forests typologies of different sizes. We also encourage the application of the presented methods and algorithms to data acquired by orbital sensors and medium spatial resolution sensors to understand if they are suitable to a wider spectrum of remote sensing data available for the estimation of structural parameters of forest as well as the classification of tree species.