

Large-area forest assessment and monitoring using disparate lidar datasets

Ranjith Gopalakrishnan

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Forestry

Valerie A. Thomas, Co-Chair
Randolph H. Wynne, Co-Chair
John W. Coulston
Thomas R. Fox

February 3rd, 2017
Blacksburg, Virginia

Keywords: Forest inventory, forestry, remote sensing, lidar, canopy heights, understory, shrub,
productivity, site index, wall-to-wall mapping, FIA

Copyright 2017

Large-area forest assessment and monitoring using disparate lidar datasets

Ranjith Gopalakrishnan

(ABSTRACT)

In the past 15 years, a large amount of public-domain lidar data has been collected over the Southeastern United States. Most of these acquisitions were undertaken by government agencies, primarily for non-forestry purposes. That is, they were collected mostly to aid in the creation of digital terrain models and to support hydrological and engineering assessments. Such data is not ideal for forestry purposes mainly due to the low pulse density per square meter, the high scan angles and low swath overlaps associated with these acquisitions. Nevertheless, the large area of coverage involved motivated this work.

In this dissertation, I first look at how such lidar data (from non-forestry acquisitions) can be combined with National Forest Inventory tree height data to generate a large-area canopy height model. A simple linear regression model was developed using two lidar-based metrics as predictors: the 85th percentile of heights of canopy first returns (h_{85}) and the coefficient of variation of the heights of canopy first returns (cv_canPts). This model had good predictive ability over 76 disparate lidar projects ($R^2 = 0.74$; RMSE=3.0 meters), covering an area of approximately 297,000 square kilometers between them. Factors leading to the residual lack-of-fit of the model were also analyzed and quantified. For example, predictive ability was found to be better for softwood forests, forests with more homogeneous vegetation structure and for terrains with gentler slopes. Given that as much as 30% of the US is covered by public domain non-forestry lidar acquisitions, this is a first step for constructing a national wall-to-wall vertical vegetation structure map, which can then be used to ask important questions regarding forest inventories, carbon sequestration, wildlife habitat suitability and fire risk mitigation.

Then, I examined whether such lidar data could be further used to predict understory shrub presence over disparate forest types. The predictability of classification model was low (accuracy = 62%, kappa = 0.23). Canopy occlusion factors and the heterogeneity of the understory layer were implicated as the main reasons for this poor performance. An analysis of the metrics chosen by the modeling framework highlighted the importance of non-understory metrics (metrics related to canopy openness and topographic aspect) in influencing shrub presence. As the proposed set of metrics were developed over a wide range of temperate forest types and topographic conditions of Southeastern US, it is expected that it will be useful for more localized future studies.

Lastly, I explored the possibility of combining lidar-derived canopy height maps with Landsat-derived stand-age maps to predict plantation pine site index over large areas (site index is a measure of forest productivity). The model performance was assessed using a Monte Carlo technique (RMSE = 3.8 meters, relative RMSE = 19%). A sample site index map for large areas of Virginia and South Carolina was generated (map coverage area: 832 sq. km) and implications were discussed. Analysis of the resulting map revealed the following: (1) there is an increase in site index in most areas, compared to the 1970s, and (2) approximately 83% of the area surveyed had low levels of productivity (defined as site index <22.0 meters for base age of 25 years). This work highlights the efficacy of combining lidar-based canopy height maps with other similar remote sensing based datasets to understand aspects of forest productivity over large areas, and to help make policy-relevant recommendations.

Large-area forest assessment and monitoring using disparate lidar datasets

Ranjith Gopalakrishnan

(GENERAL AUDIENCE ABSTRACT)

Remote sensing, in the context of forestry and forest resource management, involves the acquisition of data over large forested areas by sensors situated at a distance. A good example is a high resolution satellite image over several hundred square kilometers allowing us to identify (say) patches of deforestation, reduced forest productivity, or species diversity.

Lidar (which stands for Light Detection and Ranging) is a relatively new remote sensing technology in which the time it takes for a laser pulse to travel to a feature and return back to the sensor is used to measure how far away the feature is from the sensor. In forests, data from airborne laser scanners enable the measurement of both horizontal and vertical canopy structure (such as tree height and canopy cover).

Data from airborne laser scanners have been collected over a large area of the US (roughly 30%). However, the sensors and acquisition parameters are optimized for the inexpensive collection of the data needed for topographic mapping, and not for forest measurement. Moreover, the lidar data were collected in disparate and dissimilar projects, making the production of maps over large areas technically challenging. A systematic study is required looking at whether lidar data from such dissimilar projects can be used together to generate robust forest parameter maps over large areas. This dissertation details such a study.

Airborne laser scanner data collected for topographic mapping across many disparate projects can be used to estimate several important characteristics about forests. My conclusions are as follows:

- Lidar data can be combined effectively with field measurement data to produce high quality, wall-to-wall tree height maps over a large area.
- These lidar data can be used to map understory shrub presence, albeit with less accuracy, since fewer laser pulses penetrate the canopy.
- Forest age, as estimated using multi temporal earth resource satellite data, can be combined with lidar-derived tree heights to estimate site index (a way to know how fast trees grow on a site) for pine plantations. Most sites in the study area (Eastern Virginia and Central South Carolina) are not particularly productive (site index <22 meters), but they are more productive on the whole than they were in the 1970s.

Overall, the work outlined in this dissertation highlights the efficacy of using lidar data from disparate non-forestry projects along with other datasets to monitor useful forest parameters over large areas, and to help make policy-relevant recommendations.

Acknowledgments

First and foremost, I would like to thank my advisers Dr. Valerie Thomas and Dr. Randolph Wynne for their technical support and guidance for this work. Much more than that, they provided such great role model scientists for me to look up and aspire to. For example, it was always a treat to walk into Randy's room when he was not too busy: I would spot some interesting new book, magazine or catalog on the table or bookshelf and he would launch into a delightful related anecdote or talk about various aspects of it for quite sometime. I always left the room astonished by the breadth and depth of his knowledge. Valerie always has a good solution or suggestion for every technical problem or issue that I hit upon, she was a great help.

I would like to thank the rest of my committee members, Dr. John Coulston and Dr. Thomas Fox. Special thanks is due to Dr. Coulston who had to take time off from his busy schedule to help me with the FIA data (which was key to this work). Dr. Fox was instrumental in helping me understand and interpret my results in the context of Southern pine plantation forests. Dr. Fox was also the institutional principal investigator (PI) from Virginia Tech for the PINEMAP project (<http://www.pinemap.org/>) which was the source of funding for this work. Hence, I am deeply indebted to him for continued funding support and for giving me the opportunity to interact with and learn from a great set of scientists and students (the PINEMAP team).

I would like to thank Virginia Tech's LISA statistics team for their help with the statistical modeling of chapter 3. Chapter 4 was made possible solely due to the forest stand age map provided by Dr. Joby Kauffman. Joby had to extend himself quite a bit to get the South Carolina datasets ready in time: thanks so much Joby! I am also indebted to Dr. Matthew Fagan of the University of Maryland (Baltimore) for sharing his planted pine map work with me. Special things is also due to Dr. Evan Brooks for providing the SSURGO datasets, and for answering all the questions I had during the formulation and design of my last chapter.

I would also like to thank my friends and colleagues at the Interdisciplinary Graduate Education Program in Remote Sensing (IGEP-RS). Through interactions with them, I realized that the "why" of remote sensing is as much fraught with issues and interesting questions as the "what" and the "how".

Les Fuller, the system administration for the CEARS lab was quite accommodating with my myriad software-related requests. I also appreciate the hard work and helpfulness of Stacey Kuhar, Mary Williams and Sue Snow and thank them for all their administrative help.

I also thank my family back home in India for their continuous support through my career. Last and certainly very far from the least, I would like to thank my wife Remya Rajendran - she deserves the biggest paragraph of this section. The sacrifices and inconveniences that she went through during the five years of our time at Blacksburg are innumerable. Her support has been vital in the completion of this work and for my career in general. Thank you so much, Remya!

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Important terms	2
1.3	Forest Inventory and Analysis dataset	3
1.4	Objectives	4
2	Prediction of Canopy Heights over a Large Region Using Heterogeneous Lidar Datasets: Efficacy and Challenges	8
2.1	Introduction	10
2.2	Methods	12
2.3	Results	20
2.4	Discussion	25
2.5	Conclusions	29
3	Shrub detection using disparate airborne laser scanning acquisitions over varied forest cover types	34
3.1	Introduction	35
3.2	Methods	37
3.3	Results	46
3.4	Discussion	51
3.5	Conclusions	55
4	Predicting forest site index for managed pine in Southeastern USA using disparate lidar datasets and Landsat time-series maps	60
4.1	Introduction	61
4.2	Methods	62

4.2.1	Study area	62
4.2.2	Remote sensing datasets	62
4.2.3	Canopy height model	63
4.2.4	Generation of a sample site index map from lidar and Landsat data	65
4.2.5	Accuracy assessment of site index predictions	66
4.2.6	Change in site index	68
4.3	Results	68
4.3.1	Canopy height model	68
4.3.2	Accuracy assessment of site index map	68
4.4	Discussion	75
4.5	Conclusions	80
5	Conclusions	85
5.1	Future work	85
	Appendix A: Code and data repository	87

List of Figures

1.1	FIA plot design. Adapted from [32]	4
2.1	Study region: 13 states of the Southeastern United States	13
2.2	(a) Plot layout of a standard FIA plot; (b) The major steps involved in the lidar data processing	14
2.3	Oblique views of two lidar plots as point-clouds	15
2.4	(a) Correlation between the 85th percentile of lidar first return heights and tree heights; (b) Correlation between the coefficient of variation of the heights of canopy first returns and the FIA measured average dominant tree height	20
2.5	The effect of point density and plot homogeneity on the residuals of the fit	21
2.6	The effect of species grouping (softwoods versus hardwoods) and average stand height on the residuals of the fit	22
2.7	The effect of slope of FIA plot and lidar scan angle on the residuals of the fit	23
2.8	Scatterplot showing the agreement between subplot level dominant height and the height predicted	24
2.9	(a) An aerial photo of the area used for model verification	25
3.1	Our study region which potentially includes all forested areas with lidar coverage in 13 states of the southeast USA	38
3.2	Size and position of the lidar plot with respect to the FIA plot	39
3.3	The distribution of various lidar acquisition parameters for the 153 lidar plots used	43
3.4	(a) Proportion of various forest types represented (as reported by FIA) in the 153 plots used; (b) The physiographic class of the landform as denoted by FIA	44
3.5	(a) Confusion matrix of the classification expressed as percent of total test plots used during the repeated 10-fold cross validation (n=15300); (b) Trends in model classification accuracy for varying percent shrub cover	47
3.6	Correlation matrix of the selected variables long with the dependent variable (continuous variant)	48
3.7	Box-plots depicting the distribution of the four most important variables selected for under- story condition discrimination (absent versus present)	49

3.8	Example lidar point cloud profiles of forest plots which illustrates the efficacy of the the p_{hb2} lidar metric in distinguishing between understory absent and present conditions	51
3.9	Prominent reasons for incorrect model class predictions are illustrated here	55
4.1	Lidar data from 38 projects spread across eight states were used in this study. Each project is shown in a different color. This region is dominated by loblolly pine, with some slash pine (<i>Pinus elliotii</i>) in the extreme south	64
4.2	The lidar plot (green outline) in reference to the FIA plot (black markings). Figure adapted from [25].	65
4.3	The extend of the two lidar projects over which the site index mapping was done are shown in blue shading	66
4.4	Major steps involved in the generation of the site index map	67
4.5	Scatterplot showing the plot-level agreement between the 70th percentile of lidar first return heights and the FIA-measured average dominant tree height ($n = 217$ plots). The fitted linear model is indicated by the blue line	69
4.6	A part of the site index map generated is shown here at a scale where forest stands are visible (Newberry County, South Carolina). A very small percent of pixels (less than 0.1%) where site index values more than 30.0 meters were predicted are not shown (as they were mostly model over-predictions). Most stands register reasonable to good site index values (greater than 20.0 meters) while some under-perform (less than 10.0). A fairly large variation of site index can be seen over the landscape, too.	70
4.7	The distribution of site index values over the surveyed landscape. More than 50% of the area has a site index value less than 20.0 meters (65.6 feet)	71
4.8	The average site index for all stands that fall in the hydrological unit is shown. Only hydrological units with at least 4.5 ha of surveyed forest stand area (50 pixels) are shown. The thick black lines demarcate EPA level 2 ecoregions (Omernik 1987), see text for more details. The areas of decreased productivity (less than 18.0 m) near the transition zone between these two ecoregions roughly correspond to the South Carolina sandhills region, where poor soils are documented [36]	72
4.9	The difference between lidar-predicted site index (as of 2008 and 2010) and SSURGO-estimated site index (around the 1970s).	73
4.10	The average site index difference (between current and historical SSURGO estimates) for stands that fall in the hydrological unit is shown. Only hydrological units with at least 4.5 ha of surveyed forest stand area (50 pixels) are considered. The connotation of the thick black line is same as that in figure 4.8.	74
4.11	(a) Variation of site index for a uniformly-aged stand (age = 26 years) in the Isle of Wight county, Virginia, centered at (latitude, longitude) = (36.862, -76.711) decimal degrees, (b) Variation of site index inside a uniformly-aged stand (age = 21 years) in Williamsburg county, South Carolina, centered at (latitude, longitude) = (33.406, -79.781) decimal degrees. The scale here is the same as (a). Site index varies by as much as 12 meters (40 feet) within the stand.	77

4.12	Correspondence between forest stand ages predicted by eVCT and estimated from other sources (all ages are in years). The 1:1 line is shown in blue. Only plots that were indicated by eVCT to be either 15 years of age or above were examined. (a) Age agreement, using planting record data from the Appomattox-Buckingham State Forest plot set (23 plots, see section 4.4.3). One extreme outlier point at $(x, y) = (10, 72)$ is excluded. It can be seen that most points are near the 1:1 line. The bias is -1.1 years. Around 87% of plots have ± 3 years age agreement while 91.3% of plots have ± 5 years, (b) Age agreement, using the set of 75 FIA plots. The pronounced "vertical line" of points corresponding to eVCT age = 24 is most likely due to VCT being less reliable at the beginning of the time series, see text for more details. The overall agreement is much weaker in this case. The bias is rather small (1.1 years). Around 65.3% of plots have ± 3 years age agreement, 73.3% of plots have ± 5 years and 85.3 have ± 8 years.	78
4.13	Correspondence between site index predictions from lidar and estimates from FIA for 19 planted pine plots. The absolute difference in age (years) predicted by eVCT and estimated by FIA is also indicated. The 1:1 line is shown in blue	79

List of Tables

2.1	The variable importance values of the random forest model generated	24
2.2	Goodness of fit statistics of models based on sets of more homogeneous plots	27
3.1	Metrics used as predictor variables for shrub presence/absence in our modeling effort	45
3.2	Important predictor variables identified for modeling understory along with short descriptions (see table and text for more details)	50
4.1	The proportion of area surveyed falling in two site index classes. Only 17.4% of the area is in a medium-to-high productivity class	70

Chapter 1

Introduction

"The real voyage of discovery consists not in seeking new landscapes, but in having new eyes" said Marcel Proust, the 19th century French writer [1]. The past few decades have seen the field of forest conservation and management gaining such "new eyes", with substantial integration of remote sensing technologies. In this context, *remote sensing* refers to the acquisition of spatial information (such as images) over large areas (such as forested regions) using satellite or aircraft based sensors (for a good introduction to remote sensing concepts, refer to Campbell and Wynne, 2012 [2]). Remote sensing, coupled with effective policy interventions can have substantial positive environmental impacts: The implementation of the Real Time System for Detection of Deforestation (DETER), a satellite-based system that enabled frequent and quick identification of deforestation hot spots, was identified as the main factor for the drastic reduction in Amazon deforestation rates from 2005 to 2011 [3]. The advent of lidar (an acronym for "light detection and ranging" - a technology of generating measures of distance based on return time of emitted laser pulse) and its unique ability to characterize the three-dimensional structure of the forest canopy and understory gives yet another pair of promising new eyes to the forester. Aircraft-mounted lidar systems, commonly known as airborne laser scanning (ALS) systems can be an economically effective way of gathering high-quality structural data over vast forested areas. On a related note, the exponential increase in computing power and software capabilities in the last three decades has made possible a *big data* approach in forest sciences. This involves the computational analysis of large data sets (associated with large areas) to *understand the data* and unveil interesting spatiotemporal patterns, associations and trends. This work is about exploring how large lidar datasets ("big data lidar") can be used effectively to analyze certain key ecological parameters of Southeastern US forests.

1.1 Motivation

Forests of the Southeastern US region have high societal value, both nationally and globally. They are a vital component in the United States carbon budget (a large sink), accounting for around 36% of the total amount of sequestered carbon nationally [4]. Their economic value is also quite significant: they produce around 16% of the global industrial roundwood, which is a bigger proportion than *any other country* [5]. The following societal concerns motivate efforts to better understand and monitor structural information of these forests using large-area ALS data:

1. Forest fires: Wildfires, especially larger ones ($> 10,000$ ha), are becoming a near-nationwide concern the United States recently, notably in the drier western parts. But the Southeast is the leading region

regarding the number of wildfires per year [6]. Recently, there have been a number of large wildfires there too; further, "wildfire potential is likely to increase over the next 50 years in response to forecasted reductions in precipitation and climate driven changes in growing seasons" [7]. ALS systems have been used extensively to create spatially explicit map of forest fuel loads and associated parameters such as canopy base height and canopy bulk density [8]–[12]. These systems can effectively map canopy height, which is highly correlated to canopy fuel load [9]. Lidar data also holds the promise of sensing of the understory component, which is usually hidden from the view of traditional optical sensors by the crowns of larger trees [13]–[15].

2. **Carbon sequestration:** Southeastern forests are huge sinks of carbon, estimated at around 76 Tg C/year [16]. However, these carbon accretion rates could be highly diminished in the face of future climate change, according to some climate-coupled model outputs [17]. Monitoring of this huge carbon pool for such adverse long-term trends is hence crucial. Many previous studies have established the usefulness of lidar-derived structural information to help estimate the biomass of forests [18]–[21], thus motivating our work.
3. **Wildlife habitat suitability:** Many taxa, especially mobile animals, are quite sensitive to vegetation vertical structure (such as understory shrub cover) and lidar has been shown to be useful by many previous studies to describe such vertical characteristics that might influence animal diversity [22]–[26]. Bird species richness has also been correlated with straight-forward forest structural factors such as canopy height, vegetation cover, density and foliage height diversity [27].

Although there have been many efforts using lidar data to help address the above societal concerns in this region, this study is the first known effort to do so by integrating the US national forest inventory data (explained in detail below) with extensive non-forestry lidar data to aid in the generation of standardized forest parameter maps (related to vertical structure) over large areas of the southeast USA.

1.2 Important terms

The following definitions are important for this work:

1. **Non-forestry lidar:** This is lidar data that was collected such that neither the sensor used nor the acquisition parameters were optimized for forestry purposes (e.g., see [28]). Most data used in this dissertation were collected to create digital elevation models (DEMs). The data is then made available in the public domain, possibly via coordinating agencies such as USGS (United States Geological Survey) or NOAA (National Oceanic and Atmospheric Administration).
2. **Forest canopy:** This is the foliar cover in a forest stand, consisting of one or several layers [29]. This is also termed as the overstory layer [14]. Most of the vegetation is in the form of tree crowns, in this layer.
3. **Dominant trees:** These are trees with crowns extending above the general level of the crown cover and receiving full light from above and partly from the side; larger than the average trees in the stand, with crowns well developed but possibly somewhat crowded on the sides [30].
4. **Codominant trees:** These are trees with crowns forming the general level of the crown cover and receiving full light from above but comparatively little from the sides; usually with medium-sized crowns more or less crowded on the sides [30].
5. **Dominant canopy height:** This is the mean height of dominant and codominant trees that have always been dominant or codominant over the life of the stand [31].

6. **Forest understory:** This is the lower part of the forest vegetation cover [14], classified here using a generalized height bin of 3.0 meters or below (relative to the ground). Common components in this layer are shrubs, seedlings, saplings, herbaceous cover, tree stumps and coarse woody debris. This layer is very important in determining fire behavior: the understory has the potential to act as both surface fuels (by which fires can propagate along the ground, in the shrub layer) or as ladder fuels (by which shrub fires can be propagated to the canopy).
7. **Shrub cover:** Shrubs are woody, perennial plants (with a persistent woody stem). They differ from trees by their lower stature and their general absence of a well-defined main stem [29]. Shrub cover is defined as the proportion of ground covered by a vertical projection of the outermost perimeter of the natural spread of the shrub, including small openings [29].
8. **Site index:** Site index of a forest site is defined as the average total height that dominant and co-dominant trees are projected to reach for a given base age (usually 25 or 50 years) [30]. To determine this for a given site by field measurements, the average total height and age are determined from *site trees* (dominant and co-dominant trees with no signs of damage). Then, these values are fitted on a *site index curve*, which basically projects the height of the trees forward till the base age.
9. **Forest stand:** Forest stands are areas of forests with uniform species composition, age and density of planting (trees per hectare) which can hence be treated as homogeneous units [2]. Such stands are the basic unit of forest management.

1.3 Forest Inventory and Analysis dataset

Almost all field data for this effort came from the Forest Inventory and Analysis (FIA) division of the United States Forest Service [32]. This is the statistical wing of the US Forest Service and is tasked with collecting, analyzing and reporting on status and trends of US forests, regardless of ownership. All FIA plots follow a nationally-standardized plot design, which consists of a cluster of four circular subplots arranged in a fixed pattern (see figure 1.1). The sampling intensity is quite high: there is approximately one field sample site for every 2400 ha (6000 ac) of land. Most forested plots in the Southeast are remeasured every five years. Tree and canopy-level measurements (such as tree height) are taken only at the four subplots (total subplot area is 675 m^2). Understory measurements are done of five components (dead herbs, live herbs, dead shrubs, live shrubs, and litter coverage) at the four microplots (total microplot area is 54 m^2). Field crews record the cover (ocularly estimated) and maximum height of each of the four shrub/herb component.

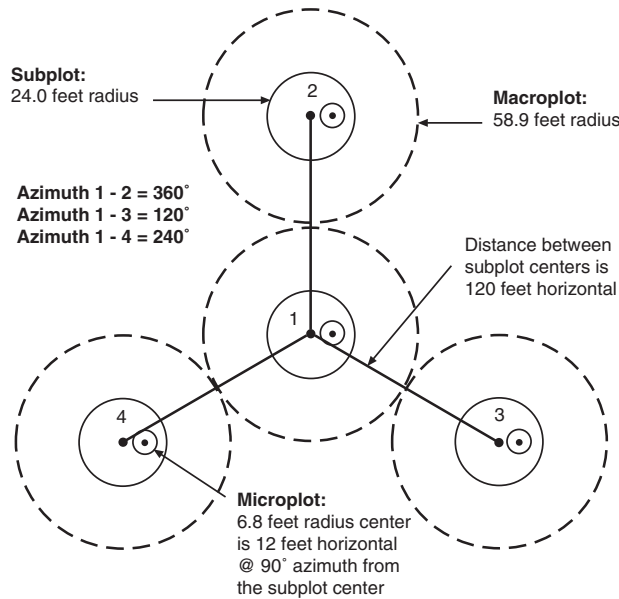


Figure 1.1: FIA plot design. Adapted from [32]

1.4 Objectives

As there is no single national-level lidar data acquisition project in the US, unlike in some smaller countries, we use disparate datasets for achieving large-area coverage. Hence, the objective of this work consists of answering the following questions:

- **Question 1:** Can we predict with reasonable accuracy dominant canopy height over large areas using disparate, medium-scale regional lidar acquisitions? What are the factors that affect the efficacy of such predictions?
- **Question 2:** Can such disparate lidar data be used for understory shrub presence mapping over varied forest types?
- **Question 3:** Can we combine dominant canopy height maps (addressed in question 1) with Landsat-based stand age maps to estimate site index over large areas of pine plantations in the Southeast? Using these maps, what proportion of such area is estimated as having low productivity?

Each question is addressed in a separate chapter in this dissertation. Thus, I try to illustrate the efficacy of combining such large-area lidar datasets with appropriate field data to help tackle issues related to forest science and management.

References

- [1] M. Proust, C. K. Scott-Moncrieff, and S. Hudson, *Remembrance of things past*. Chatto & Windus, 1970.
- [2] J. B. Campbell and R. H. Wynne, *Introduction to Remote Sensing, Fifth Edition*, en. Guilford Press, Feb. 2012, ISBN: 9781609181772.

- [3] J. Assunção, C. Gandour, and R. Rocha, “DETERring Deforestation in the Brazilian Amazon: Environmental Monitoring and Law Enforcement,” *Climate Policy Initiative Report, PUC-Rio, May*, 2013. [Online]. Available: http://www.econ.puc-rio.br/uploads/adm/trabalhos/files/Command_and_Control.pdf.
- [4] D. P. Turner, G. J. Koerper, M. E. Harmon, and J. J. Lee, “A carbon budget for forests of the conterminous United States,” *Ecological Applications*, vol. 5, no. 2, pp. 421–436, 1995. [Online]. Available: <http://onlinelibrary.wiley.com/doi/10.2307/1942033/full>.
- [5] J. P. Prestemon and R. C. Abt, “Timber products supply and demand,” 2002. [Online]. Available: <http://www.treesearch.fs.fed.us/pubs/42386>.
- [6] A. Andreu and L. A. Hermansen-Baez, “Fire in the South 2: The southern wildfire risk assessment,” 2008. [Online]. Available: <http://www.treesearch.fs.fed.us/pubs/34087>.
- [7] D. Wear and J. Greis, “The Southern Forest Futures Project: Summary Report,” Tech. Rep., 2011. [Online]. Available: <http://www.srs.fs.usda.gov/futures/>.
- [8] D. Riano, E. Meier, B. Allgower, E. Chuvieco, and S. L. Ustin, “Modeling airborne laser scanning data for the spatial generation of critical forest parameters in fire behavior modeling,” *Remote Sensing of Environment*, vol. 86, no. 2, pp. 177–186, 2003.
- [9] T. L. Erdody and L. M. Moskal, “Fusion of LiDAR and imagery for estimating forest canopy fuels,” *Remote Sensing of Environment*, vol. 114, no. 4, pp. 725–737, 2010.
- [10] M. Mutlu, S. C. Popescu, C. Stripling, and T. Spencer, “Mapping surface fuel models using lidar and multispectral data fusion for fire behavior,” *Remote Sensing of Environment*, vol. 112, no. 1, pp. 274–285, 2008.
- [11] M. Mutlu, S. C. Popescu, and K. Zhao, “Sensitivity analysis of fire behavior modeling with LIDAR-derived surface fuel maps,” *Forest Ecology and Management*, vol. 256, no. 3, pp. 289–294, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378112708003332>.
- [12] L. A. Arroyo, C. Pascual, and J. A. Manzanera, “Fire models and methods to map fuel types: The role of remote sensing,” *Forest Ecology and Management*, vol. 256, no. 6, pp. 1239–1252, 2008.
- [13] S. Martinuzzi, L. A. Vierling, W. A. Gould, M. J. Falkowski, J. S. Evans, A. T. Hudak, and K. T. Vierling, “Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability,” *Remote Sensing of Environment*, vol. 113, no. 12, pp. 2533–2546, 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425709002132>.
- [14] B. M. Wing, M. W. Ritchie, K. Boston, W. B. Cohen, and M. J. Olsen, “Individual snag detection using neighborhood attribute filtered airborne lidar data,” *Remote Sensing of Environment*, vol. 163, pp. 165–179, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425715001091>.
- [15] F. Morsdorf, A. M\ a arell, B. Koetz, N. Cassagne, F. Pimont, E. Rigolot, and B. Allgöwer, “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*, vol. 114, no. 7, pp. 1403–1415, 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425710000568>.
- [16] F. X. Han, M. J. Plodinec, Y. Su, D. L. Monts, and Z. Li, “Terrestrial carbon pools in southeast and south-central United States,” *Climatic Change*, vol. 84, no. 2, pp. 191–202, 2007.
- [17] U. D.o.A.U. F. Service, *Future of America’s forest and rangelands: Forest Service 2010 Resources Planning Act Assessment. General Technical Report GTR-WO-87*. USDA Forest Service Washington, DC, 2012.

- [18] M. A. Lefsky, W. B. Cohen, D. J. Harding, G. G. Parker, S. A. Acker, and S. T. Gower, "Lidar remote sensing of above-ground biomass in three biomes," *Global ecology and biogeography*, vol. 11, no. 5, pp. 393–399, 2002. [Online]. Available: <http://onlinelibrary.wiley.com/doi/10.1046/j.1466-822x.2002.00303.x/full>.
- [19] M. A. Lefsky, D. J. Harding, M. Keller, W. B. Cohen, C. C. Carabajal, F. Del Bom Espirito-Santo, M. O. Hunter, and R. de Oliveira, "Estimates of forest canopy height and aboveground biomass using ICE-Sat," *Geophysical Research Letters*, vol. 32, no. 22, 2005. [Online]. Available: <http://onlinelibrary.wiley.com/doi/10.1029/2005GL023971/full>.
- [20] G. P. Asner, J. Mascaro, H. C. Muller-Landau, G. Vieilledent, R. Vaudry, M. Rasamoelina, J. S. Hall, and M. van Breugel, "A universal airborne LiDAR approach for tropical forest carbon mapping," *Oecologia*, vol. 168, no. 4, pp. 1147–1160, 2012. [Online]. Available: <http://link.springer.com/article/10.1007/s00442-011-2165-z>.
- [21] G. P. Asner, R. F. Hughes, J. Mascaro, A. L. Uowolo, D. E. Knapp, J. Jacobson, T. Kennedy-Bowdoin, and J. K. Clark, "High-resolution carbon mapping on the million-hectare Island of Hawaii," *Frontiers in Ecology and the Environment*, vol. 9, no. 8, pp. 434–439, 2011. [Online]. Available: <http://www.esajournals.org/doi/abs/10.1890/100179>.
- [22] J. Müller and K. Vierling, "Assessing biodiversity by airborne laser scanning," in *Forestry Applications of Airborne Laser Scanning*, Springer, 2014, pp. 357–374. [Online]. Available: http://link.springer.com/chapter/10.1007/978-94-017-8663-8_18.
- [23] G. J. Newnham, A. S. Siggins, R. M. Blanche, D. S. Culvenor, J. E. Leonard, and J. S. Mashford, "Exploiting three dimensional vegetation structure to map wildland extent," *Remote sensing of environment*, vol. 123, pp. 155–162, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425712001198>.
- [24] P. Hyde, R. Dubayah, W. Walker, J. B. Blair, M. Hofton, and C. Hunsaker, "Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM+, Quickbird) synergy," *Remote Sensing of Environment*, vol. 102, no. 1, pp. 63–73, 2006. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425706000551>.
- [25] C. Bässler, J. Stadler, J. Müller, B. Förster, A. Göttlein, and R. Brandl, "LiDAR as a rapid tool to predict forest habitat types in Natura 2000 networks," *Biodiversity and Conservation*, vol. 20, no. 3, pp. 465–481, 2011. [Online]. Available: <http://link.springer.com/article/10.1007/s10531-010-9959-x>.
- [26] R. Clawges, K. Vierling, L. Vierling, and E. Rowell, "The use of airborne lidar to assess avian species diversity, density, and occurrence in a pine/aspen forest," *Remote Sensing of Environment*, vol. 112, no. 5, pp. 2064–2073, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425708000291>.
- [27] S. Goetz, D. Steinberg, R. Dubayah, and B. Blair, "Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA," *Remote Sensing of Environment*, vol. 108, no. 3, pp. 254–263, 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425706004676>.
- [28] J. C. White, M. A. Wulder, A. Varhola, M. Vastaranta, N. C. Coops, B. D. Cook, D. Pitt, and M. Woods, "A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach," *The Forestry Chronicle*, vol. 89, no. 6, pp. 722–723, 2013. [Online]. Available: <http://pubs.cif-ifc.org/doi/abs/10.5558/tfc2013-132>.
- [29] J. A. Helms and others, *The dictionary of forestry*. CAB INTERNATIONAL, 1998. [Online]. Available: <http://www.cabdirect.org/abstracts/19980616323.html>.
- [30] T. E. Avery and H. E. Burkhart, *Forest measurements 5th ed.* Boston, MA: McGraw Hill, 2001.

- [31] U. Diéguez-Aranda, H. E. Burkhart, and R. L. Amateis, “Dynamic site model for loblolly pine (*Pinus taeda* L.) plantations in the United States,” *Forest Science*, vol. 52, no. 3, pp. 262–272, 2006. [Online]. Available: <http://www.ingentaconnect.com/content/saf/fs/2006/00000052/00000003/art00006>.
- [32] W. A. Bechtold and P. L. Patterson, *The enhanced forest inventory and analysis program: National sampling design and estimation procedures*. US Department of Agriculture Forest Service, Southern Research Station, 2005. [Online]. Available: http://www.srs.fs.usda.gov/pubs/gtr/gtr_srs080/gtr_srs080.

Chapter 2

Prediction of Canopy Heights over a Large Region Using Heterogeneous Lidar Datasets: Efficacy and Challenges

Published in *Remote Sensing*, Volume 7, Issue 9 (September 2015), Pages 11036-11060. available via <http://www.mdpi.com/2072-4292/7/9/11036>

This is an Open Access article. Non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly attributed, cited, and is not altered, transformed or built upon in any way, is permitted. The moral rights of the named author(s) have been asserted.

Attribution: Ranjith Gopalakrishnan contributed to the research design, collected the data, designed and implemented the software for lidar-FIA data intersection, and led the interpretation of results and manuscript writing. Valerie Thomas and Randolph Wynne secured funding, proposed and developed the research design, helped in the interpretation of results and in manuscript writing and revisions. John Coulston helped with the research design, FIA data management and the intersection efforts, and in the interpretation of results.

Article

Prediction of Canopy Heights over a Large Region Using Heterogeneous Lidar Datasets: Efficacy and Challenges

Ranjith Gopalakrishnan ^{1,*}, Valerie A. Thomas ¹, John W. Coulston ² and Randolph H. Wynne ¹

¹ Dept. of Forest Resources and Environment Conservation, Virginia Tech, Blacksburg, VA 24061, USA; E-Mails: thomasv@vt.edu (V.A.T.); wynne@vt.edu (R.H.W.)

² USDA Forest Service (Southern Research Station), Knoxville, TN 37919, USA; E-Mail: jcoulston@fs.fed.us

* Author to whom correspondence should be addressed; E-Mail: ranjith7@vt.edu; Tel.: +1-540-231-7681; Fax: +1-540-231-3698.

Academic Editors: Nicolas Baghdadi and Prasad S. Thenkabail

Received: 20 March 2015 / Accepted: 12 August 2015 / Published: 27 August 2015

Abstract: Generating accurate and unbiased wall-to-wall canopy height maps from airborne lidar data for large regions is useful to forest scientists and natural resource managers. However, mapping large areas often involves using lidar data from different projects, with varying acquisition parameters. In this work, we address the important question of whether one can accurately model canopy heights over large areas of the Southeastern US using a very heterogeneous dataset of small-footprint, discrete-return airborne lidar data (with 76 separate lidar projects). A unique aspect of this effort is the use of nationally uniform and extensive field data (~1800 forested plots) from the Forest Inventory and Analysis (FIA) program of the US Forest Service. Preliminary results are quite promising: Over all lidar projects, we observe a good correlation between the 85th percentile of lidar heights and field-measured height ($r = 0.85$). We construct a linear regression model to predict subplot-level dominant tree heights from distributional lidar metrics ($R^2 = 0.74$, RMSE = 3.0 m, $n = 1755$). We also identify and quantify the importance of several factors (like heterogeneity of vegetation, point density, the predominance of hardwoods or softwoods, the average height of the forest stand, slope of the plot, and average scan angle of lidar acquisition) that influence the efficacy of predicting canopy heights from lidar data. For example, a subset of plots (coefficient of variation of vegetation heights <0.2) significantly reduces the RMSE of our model from 3.0–2.4 m (~20% reduction). We conclude that when all these elements are factored

into consideration, combining data from disparate lidar projects does not preclude robust estimation of canopy heights.

Keywords: forest inventory, forestry, forest mensuration, lidar, canopy heights, wall-to-wall mapping, co-registration, FIA

1. Introduction

Generating unbiased, accurate and spatially explicit maps of the forest canopy height on a planetary scale is important for the following reasons (part of the text in this section appeared in [1]): First, canopy height is a good proxy for forest health and forest biomass, and provides an effective way to monitor the effect of global change on forests. Indeed, forest vertical structure and canopy height have been shown as good predictors for leaf area index [2] and even biodiversity in some forest types [3]. Second, the canopy height (used along with the species type) is an important input for forest fire models. In other words, it is a good indicator of wildfire risk, especially for catastrophic canopy fires. In the US, wildfires are getting larger and causing more damage: A recent paper by a prominent non-profit research group noted that the annual federal spending on wildfires increased (adjusted for inflation) from \$1.4 billion (from 1991 to 1999) to \$3.5 billion (2002–2012) [4]. The report attributed this to mainly historical management practices (such as fire suppression), climate change and more residential development near forested areas. Third, the United Nations Framework Convention on Climate Change (UNFCCC) has championed periodic monitoring of forest biomass via the Reducing Emissions from Deforestation and forest Degradation initiative (REDD; see http://unfccc.int/methods_science/redd/items/4531.php) effort. Keeping this international context in mind, the development of methods to effectively measure canopy heights (a good surrogate for biomass) becomes crucial. The southern US region is no exception regarding these motivations. Currently, wildfires in the South of US are relatively small and easily contained [5]. But a recent analysis of the possible future changes noted that “wildfire potential is likely to increase over the next 50 years in response to forecasted reductions in precipitation and climate driven changes in growing seasons” [6]. Pockets of high wildfire potential are present in Northern Florida, Eastern Texas and the Southern part of North Carolina [7]. Southern forests are also a significant portion of the U.S. carbon budget, accounting for approximately 36% of the sequestered forest carbon in the conterminous United States [8] because of their large area and productivity. Monitoring of this large biomass pool for unfavorable long-term trends is hence very important.

One approach to generate lidar-based maps of canopy height (and other forest biophysical variables) is the *area-based approach*, where statistical regression models are developed between plot-level lidar-derived distributional metrics and field-measured height values, and then used for area-wide estimates. The advantages of such an approach compared with traditional stand-level forest inventories include having complete spatial knowledge of predicted variables, more precise prediction of certain forest variables, and the ability to calculate confidence intervals for estimates [9,10]. A similar globally-scalable approach for producing wall-to-wall canopy height maps over large regions involves using spaceborne lidar. Data from the Geoscience Laser Altimeter System (GLAS) aboard ICESat (Ice, Cloud, and land Elevation Satellite) has been used to generate global canopy height maps, with an RMSE of 6.1 m (the validation

used 66 FLUXNET sites) [11]. Another study used similar data from the same sensor, which was corrected utilizing the Shuttle Radar Topography Mission (SRTM) data [12]. They reported an RMSE of 9.6 m, for the case of using all data from their 66 plots. Though these efforts are useful in a global context, the RMSEs of involved are too high to be useful for more regional forest monitoring. Also, ICESat has been retired since 2010, and there are no spaceborne lidar sensors currently operating (as of August 2015).

There have been some recent efforts to implement an area-based approach with lidar data (both scanning and profiling sensors) in estimating forest biophysical parameters (such as canopy height) over large regions. A recent publication broadly surveyed the work done in this context, and concluded that such sample-based approaches that used ground-plot data were quite well-established for large area (including national) inventories [13]. An effort to estimate national forest metrics over the entire country of Denmark is described in [14]. However, the ALS (airborne lidar system) data there were from a single survey (hence homogeneous with respect to acquisition parameters), which is impractical for other large regions. For larger countries, lidar data are usually collected by multiple agencies with varying levels of expertise, and often for non-forestry purposes (such as terrain mapping). In this case, acquisition parameters vary by project, and “there is a notable lack of standards regarding processing, deliverables, and data quality” [15]. Pooling and using such data requires understanding of how various lidar collection parameters (such as point density and footprint) affect the accuracy of the estimates.

As such, there have been several efforts to assess the impact of lidar sensor specifications and settings and flight settings on the distributional metrics of lidar returns acquired over forest patches. A typical procedure in many prior studies has been to have multiple lidar acquisition flights (with varying acquisition parameters) over the same forest area, and study whether differences in the distributional metrics and their effect on biophysical parameter estimation are statistically significant. A study in southeast Norway considered the effect of varying sensors, flying altitudes, and pulse repetition frequencies on forest canopy metrics, for a small footprint lidar [16]. With respect to sensor difference, five of the nine distributional metrics changed significantly, even though the two sensors were from the same manufacturer (Optech’s ALTM1233 and ALTM3100 series sensors). For example, the 50th percentile of height above ground (h_{50}) changed by as much as 0.3 m, on average. Increasing the flying altitude from 1100 to 2000 m also makes a difference: For example, h_{50} changed by 0.18 m. The importance of pulse density has been noted in other studies as well [17]. Another study looked at the effect of canopy conditions (leaf-on vs. leaf-off) on lidar derived metrics, in a mixed forest [18]. Statistically significant differences were found: h_{50} differed by 0.76 m for first returns and 2.44 m for last returns. These were noted to be significant, as one expects distributional metrics to be relatively stable, especially to conditions such as flying altitude and point density. A recent effort on integrating different lidar datasets used 28 of them together, and analyzed whether the variation in height (captured by lidar) is captured by other satellite-image based classification schemes [19]. The authors concluded that there were significant differences in heights between classes (such as those in the National Land Cover classes of 2006), which speaks to the robustness of these classification schemes. The study shows the utility of mosaicking and analyzing disparate lidar datasets (as we do) and motivates the utility of having canopy height maps to aid better land cover classification. But a drawback of the work is that they did not attempt to bring in field-measured heights into their map construction.

Another challenge for developing statistical regression models of canopy heights is to obtain a reasonably large set of well-designed field plots randomly distributed over the area of interest. In this context, an interesting avenue that has been relatively unexplored for the United States is the use of national forest inventory plots as the field plots. We use field data from such inventory plots; namely plots of the Forest Inventory and Analysis (FIA) program of the United States Forest Service (USFS). The FIA sampling design is statistically robust and their methodology has a legacy of more than two decades. The spatial extension of such point-data using a powerful tool such as lidar opens the possibility of having much more robust and synoptic wall-to-wall estimates of forest parameters. A major challenge in using FIA field data is the uncertainty in the accuracy of plot locations. We designed our lidar plots to factor in this uncertainty, and we further quantified the effect of this design on the accuracy of generated canopy heights.

In this paper, our hypothesis is that ALS data collected by a large number of disparate, medium-scale regional acquisition efforts (over 110, in our case) primarily for non-forestry purposes (such as terrain mapping and digital terrain model generation) can be effectively combined with national forest inventory field measurements (such as those from the FIA) to generate reasonable canopy height maps over large regions, using an area-based approach. Disparate lidar acquisitions are needed because there are no large-scale, national-level lidar data acquisition projects in the US. Such a combination of heterogeneous datasets brings forth various issues such as the effect of lidar pulse density, the land-use patterns of the ground plots and the species groups (softwood vs. hardwood) involved, and questions on how they affect the accuracy of the maps generated. We try to address these questions by quantifying how each such issue may affect the efficacy of the models involved.

2. Methods

2.1. Study Area and Lidar Data Used

Our study area includes forested regions with lidar coverage for the 13 southern states of the US (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Texas, Tennessee, and Virginia). The forests in this region are quite heterogeneous, and may be broadleaf (e.g., oak-hickory and maple-beech-birch forests), or conifer forests (e.g., planted pines) or even mixed forests (e.g., oak-pine, oak-gumcypress).

We obtained regional public-domain lidar datasets archived at three agencies, as follows: (1) The U.S. Geological Survey (USGS); (2) National Oceanic and Atmospheric Administration (NOAA); and (3) Natural Resources Conservation Service (Alabama) (Figure 1). The first two agencies are the coordinating members of large data sharing partnerships, by which they store and disseminate lidar data collected by various partner agencies. For example, NOAA is the lead agency for “Digital Coast”, an initiative by which several organizations have entered into a data sharing initiative [20,21]. Many of these datasets are generated at the county-level. That is, to support engineering, hydrological and environmental studies, the public works department (or a similar department) of a given county initiates contracts with private land survey firms. These contracts are to create digital terrain models of significant parts of the county. This usually involves collecting ALS-based multiple return lidar data over the terrain. This data is then shared in the public domain, via coordinating agencies such as NOAA. Other datasets

were collected by bigger agencies such as the United States Army Corps of Engineers (USACE) and the Bureau of Land Management (BLM). This heterogeneity in collection agencies is reflected in the data in terms of lidar acquisition parameters and the amount of post-acquisition quality checking done.

The lidar data were in the form of geo-referenced point clouds from 76 separate Lidar projects, with data acquisition years ranging from 2002–2012. The size of the data were around 7 terabytes (compressed). The lidar data acquired in all of these projects had multiple returns, with the maximum number of returns being four.

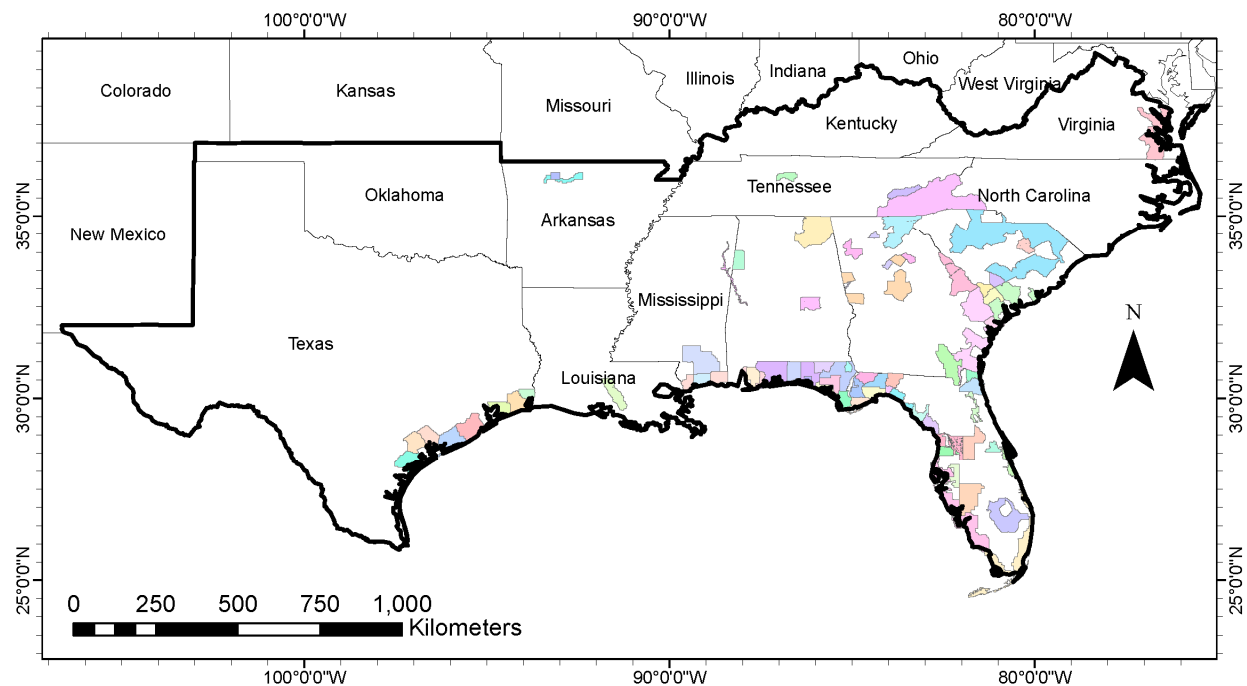


Figure 1. Study region—13 states of the Southeastern United States (outlined by thick black border). Lidar coverage of various projects (76 in all) is shown in different colors. The total area of lidar coverage is ~297,000 square kilometers. A total of 1755 Forest Inventory and Analysis (FIA) plots were used from this area.

2.2. Forest Inventory Data

The Forest Inventory and Analysis (FIA) program of the United States Forest Service (USFS) is responsible for collecting nation-wide forest inventories in the US. It has been collecting field data on a nationally-standardized sampling plot since the 1990s [22]. Such plots are known as phase 2 (P2) plots, as they are used in the second phase of the three-phased inventory scheme. The quality assurance protocol used for data collection is quite robust (which will be discussed in detail later) which contributes to the usability of this dataset. The FIA P2 plot (henceforth FIA plot) layout is shown in Figure 2a. The plot location is recorded with GPS devices of variable accuracies, so co-registration with lidar data may not be always accurate. The height of all standing trees greater than or equal to 12.7 cm (5.0 inches) in diameter in the *four subplots* are measured; hence, plot size = ~675 m². As this plot size is relatively large and includes as much as 120–130 trees in a well-stocked area, there is a good sampling of the local canopy height by these FIA subplots. The field heights of trees were measured using clinometers.

Because of practical limitations of such devices in closed canopy stands, the stated accuracy of field-based height measurements was $\pm 10\%$.

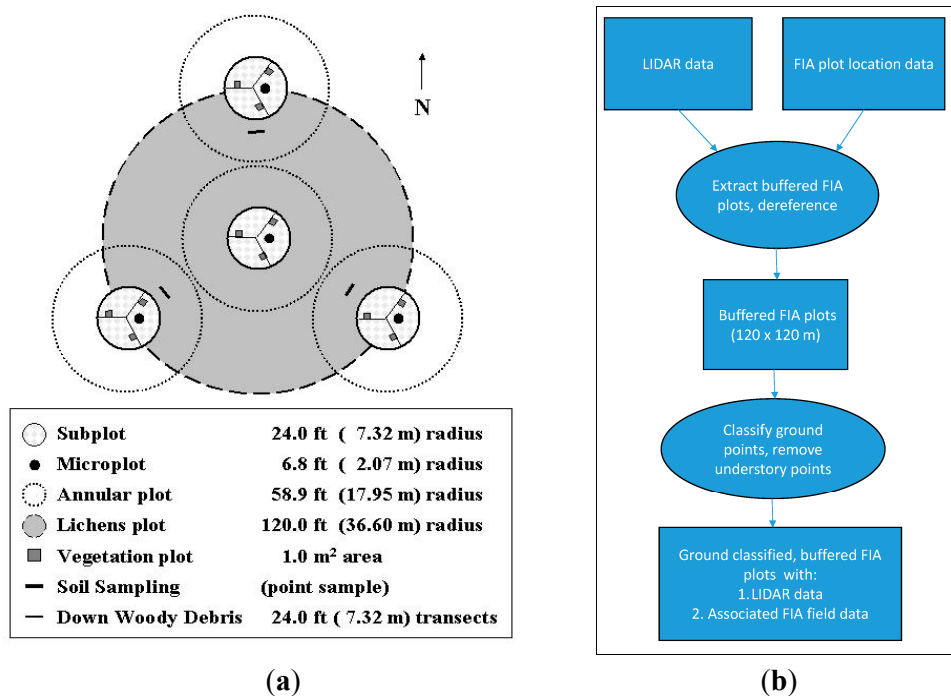


Figure 2. (a) Plot layout of a standard FIA plot [23]. The height of all trees with diameter at breast height (dbh) greater or equal to 5.0 inches is measured in the four subplots shown [22]; (b) The major steps involved in the lidar data processing.

2.3. Lidar Plot Size

For the area based approach, the sample unit is a square grid cell, and forest parameters are calculated at this resolution. It is important that the area of the ground plot and the grid cell (and hence the lidar plot) should be as similar as possible [10]. Given the spatial design of the FIA plots (see Figure 2a), it would take a square of side of at least 87.8 m to be centered on the FIA plot to cover all the four subplots. Another consideration that has to be taken into account is the error in the GPS locations of the plots, as measured by the FIA. Hence, these two criteria translate into the following:

- (1) The size of the lidar plot should be as close to that needed for the FIA plot (87.8 m).
- (2) Given the uncertainty in the spatial location of the FIA plot, the lidar plot should have a good probability of encompassing the entire FIA plot.

As for the spatial location uncertainty, an estimated probability density function of GPS errors used has been reported by scientists at the USFS FIA program [24]. This function was based upon quite extensive field trials and other considerations [25]. From this probability density function [24], we gathered that a vast majority of GPS location errors are within 15 m. Hence, we decided to use a buffer of 15 m on all sides, for meeting criteria 2 above. Hence, a good lidar plot is a square of size is 87.8 m (criteria 1) + 30 m (15 m on both sides, criteria 2) = 117.8 m. We decided to use a round number of 120.0 m, for simplicity.

2.4. Lidar Data Processing and Metrics Computed

The major steps involved in the processing of lidar data are shown in Figure 2b, and are elaborated below:

- (1) For each FIA plot location, we checked whether there was a spatially corresponding lidar acquisition within ± 2 years of plot measurement. If so, lidar data corresponding to a plot encompassing the FIA plot was extracted. This was a north-south oriented square plot of length 120 m, centered at the given location of the FIA plot center. These square plots will be henceforth called the *buffered FIA plots* (see Figure 3). A total of 3337 such plots (corresponding to FIA plots) were cut out from the lidar point cloud data.
- (2) Ground classification on the buffered FIA plots was done using the method of progressive TIN (triangulated irregular network) densification [26], as implemented in *lasground*, which is part of the open-source toolset *lastools* (see <http://lastools.org>).
- (3) Understory removal: Several height thresholds have been used in the literature to remove possible ground and understory points; these range from 0.9 m [27] to 2.0 m [10,28]. We decided on the threshold of 3.0 m for the following two reasons: (a) After manual inspection of several plots for understory height, we decided that a threshold of 3.0 m was better; (b) We looked at the FIA phase 3 plots for the region, where understory heights were also measured. There, we noticed that a good proportion of shrub heights were recorded above 3 m. Hence, all points in the height bin of 0.0–3.0 m (above ground) were considered non-canopy points, and were not considered for calculation of canopy height metrics.

The resulting preprocessed data consisted of a set of plots for which we have lidar data and associated FIA individual tree heights (for the subplots). All data analysis was then performed on this set of paired elements.

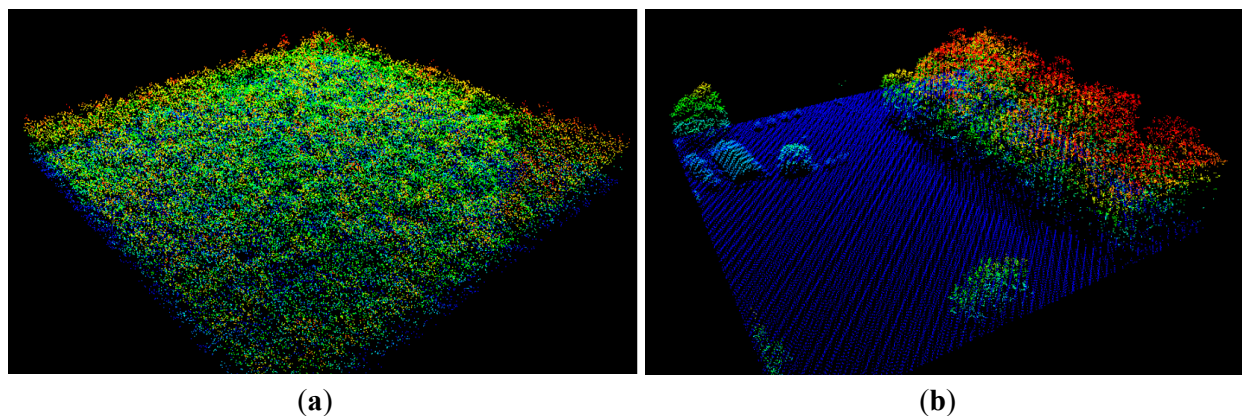


Figure 3. (a) An oblique view of a buffered FIA lidar plot (120×120 m). The lidar point cloud is colored by elevation. Blue represents low-elevation (ground) points, while red represents higher elevations. The four circular FIA subplots of Figure 2a cover only a fraction of this area. This plot is highly homogeneous, with respect to vegetation height; (b) A highly non-homogeneous buffered FIA lidar plot, with trees clustered on the right side.

Several metrics (descriptive lidar metrics and otherwise) were computed over the buffered FIA plots, and for the FIA data:

- (1) Lidar distributional metrics for canopy height: Height percentiles have been shown to be significant canopy height predictors in previous related efforts [10,27]. Hence, the major percentiles (5th, 10th, 15th, 20th ...) of the heights (above ground) of canopy first returns (*i.e.*, greater than 3.0 m from the ground) over the buffered FIA plots were calculated. These are denoted as h_5 , h_{10} , h_{15} , *etc.* We also calculated the coefficient of variation of the heights of canopy first return points, as it has shown to be significant in such models [27]. This metric is denoted as cv_canPts .
- (2) Plot homogeneity: This quantifies the homogeneity of distribution of vegetation height in the area around the FIA plots. The importance of this metric will be explained in the next section. For computing it, we divide the buffered FIA plot into 144 square units, each of area 100 m². This was done by using a regular grid pattern. Then, we calculate the 85th percentile of heights of all returns (including understory) for each of those 144 square units. Finally, we calculate the coefficient of variation of these 144 values (henceforth CV), which is a normalized quantification of the amount of dispersion of vegetation heights over the plot. Hence, given that $(h_{85})_i$ is the 85th percentile height of all lidar returns for the i th square unit, the CV at the plot level is calculated as:

$$CV = \frac{\left(\frac{1}{143} \sum_{i=1}^n ((h_{85})_i - \overline{h_{85}})^2 \right)^{1/2}}{\overline{h_{85}}} \quad (1)$$

$$\text{where } \overline{h_{85}} = \frac{\sum_{i=1}^{144} (h_{85})_i}{144} \quad (2)$$

We include the understory points in this calculation because we hypothesize that the amount of homogeneity of the entire vegetation structure (including the understory) on the 120 m square plot is instrumental to good correlations between lidar and field measured metrics. This is because the presence of understory vegetation plays a role in determining how well ground detection algorithms work.

- (1) Scan angle: The average scan angle, in degrees, as recorded in the lidar metadata. Additionally, we flagged plots that had lidar data with either no scan angle recorded or had scan angles improperly recorded. Also, we flagged plots where the lidar data were from multiple flight lines (hence, averaged scan angles are not representative). These were marked with a “no good scan angle information” flag.
- (2) Slope of the buffered FIA lidar plot: This is the average slope, expressed in degrees, measured in the field by the FIA crew.
- (3) The dominant tree height of all the trees measured by FIA on the subplots. The dominant tree height (henceforth “dominant height”) is the average height of the five tallest trees in the four FIA subplots. Note that this definition differs from the more common definition used in forestry. We used this definition to maintain consistency across all plots.

- (4) We also estimated the effect of broad species grouping (softwood vs. hardwood) on the models. Softwood trees include evergreen conifers such as pines, spruces and cedars, while hardwoods include deciduous trees such as oaks, maples and birch. The FIA field crew also records species data for all trees measured. We used that information to estimate the percentage (by basal area) of softwoods in the four FIA subplots. That is, we defined and calculated *percent_softwood* using the following formula:

$$\text{percent_softwood} = \frac{\left(\sum_{i=1}^m (BA)_i \right) \times 100.0}{\left(\sum_{i=1}^n (BA)_i \right)} \quad (3)$$

where $(BA)_i$ = the basal area of the i th tree; m = number of softwood trees on the plots; (summation over all softwood trees); n = total number of trees on the plot; (summation over all trees)

2.5. Accounting for Plots with Multiple Land Use Conditions

For a land use parcel to be classified as forest by the FIA, it must be at least 10 percent stocked with tree species, at least 1 acre in size, and at least 120 feet wide [22]. An important aspect of the FIA plot design (fully randomized plot locations) is that plots can potentially straddle multiple land use types (called “conditions” in the FIA literature). For example, one could have a plot where 40% (by area) is forested, 40% is nonforested (agricultural land) and 20% is water (rivers, lakes, *etc.*). Such randomized locations are needed for truly unbiased estimates with no domain misclassifications [22]. Forest land also includes transition zones, such as areas between forest and nonforest lands that meet the minimal tree stocking requirement. Unimproved roads and trails, streams, and clearings in forest areas are also classified as forest if they are less than 36.6 m (120 feet) wide or less than 0.4 hectares (an acre) in size. But as our modeling effort was focused on forested regions, we selected a subset of FIA plots with predominant forested land use at the time of measurement, defined as follows: (a) Over 90% of the land use was classified as “forested”; (b) There were at least three trees on each of the four FIA subplots. These two conditions resulted in the selection of 1755 FIA plots (from the 3337 plots discussed in Section 2.4.), and all subsequent analysis and modeling were restricted to data from these plots.

2.6. Main Model Specification

We developed an OLS (ordinary least squares) based bivariate linear regression model based on data from the 1755 plots mentioned above. All variables were estimated at the buffered FIA plot level. We considered the dominant tree height of all the trees measured by FIA on the subplots as the response variable. The following predictor variables were considered:

- (1) A height percentile from lidar canopy first returns (from the set of h_5 , h_{10} , h_{15} , h_{20} , *etc.*). The percentile chosen was the one most correlated to the dominant height.
- (2) *cv_canPts*, the coefficient of variation of the heights of canopy first returns.

We chose a simple model with no data transformations as our primary interest was in the regression residual analysis for factor importance. Also, a simple model is less prone to overfitting [29], which is

essential when one plans to extend the model over a large area. Before constructing the model, we assessed the assumptions of linear regression (near-linear trend, normality of distribution of variables, homogeneity of variances) with both the explanatory and response variables to ensure that the model choice was appropriate. Then, the model was developed and standard goodness-of-fit metrics like the coefficient of determination (R^2) and root mean square error (RMSE) were calculated.

2.7. Factors Affecting Efficacy of Prediction

The objective of this stage of analysis was to understand the relative importance of several factors in affecting the efficacy of the main model. We considered the following factors:

- (a) Point density of lidar returns (all returns), over the buffered FIA plot. This is expressed as (number of returns)/m².
- (b) Plot homogeneity, estimated from lidar returns (quantified by CV).
- (c) Percent softwood (estimated from FIA field data).
- (d) Average height of trees (estimated from FIA field data).
- (e) Slope of the plot terrain (estimated from FIA field data).
- (f) Average lidar pulse scan angle (estimated from lidar metadata).

Such an analysis can have three potential uses: (1) Understand the importance of lidar acquisition parameters (like point density) on the robustness of the models involved; (2) Understand how the study area characteristics affect model performance; (3) Help in the stratification of the study area, for better confidence intervals of parameter estimates. It is important to recognize that the above list is not exhaustive with respect to factors that could possibly affect model performance. Nevertheless, we were constrained by factors that were consistently available across all 76 projects involved. We used regression residual analysis to analyze the relative importance of the above factors in effecting deviations from a linear fit. Such analysis is useful in understanding the factors that affect the main model [30,31]. We made standard boxplots of the residuals for various ranges of the factors involved. If the spread (variance) of the residuals were affected by the level of the factor at hand, it indicates that the factor is important to model canopy heights. That is, (for example) if plots with low point density had a bigger spread of residuals than plots with higher point density, it indicates the importance of point density in effecting lower residuals. We used the Brown-Forsythe test, a formal nonparametric test for homogeneity of variances to test the significance of our measured difference in variance, or spread [32].

For another independent assessment of the relative importance of factors above, we also used the Random Forest (RF) algorithm, a novel and powerful extension of classification tree techniques. Variable importance is determined in this technique by changing the data for a given variable and noting the increase in prediction error. The random forest algorithm has been shown to produce excellent results in classifications of remotely sensed and ecological data [33]. We used the RF package (version 4.6–10) [34] in *R* (www.r-project.org). The predictor variables were those used in the main model plus five of the six factors listed above (the average height of trees was excluded, explained below). The response variable was dominant height, same as that used in the main model. The average height of trees was excluded, as it is highly correlated with the response variable. A subset consisting of a random selection of 70% of the plots was used for training, and the rest (30%) of the plots used for testing. This process was repeated

10 times and the results were averaged. First, the robustness of the model was verified by comparing the RMSE (generated using the testing subset) with the RMSE of the main model. Then, the relative importance of each of the six factors above were recorded. For this exercise, we had to exclude plots that had the “no good scan angle information” flag set. Hence, we could use only the 1153 plots where all required independent variables were available.

2.8. Generation of Sample Canopy Height Map

Recent papers have detailed the area-based approach, a methodology to derive wall-to-wall maps of forest parameter estimates from Lidar data [9]. There have been several efforts that have successfully generated such maps [35]. There are two essential stages involved in this approach. During the first stage, ALS data is acquired over the region of interest, along with vegetation measurement data from sample ground plots. Then, predictive models are developed between the ALS data and ground measurements. In the second stage, the ALS metrics used for model construction are calculated over the entire wall-to-wall area. Here, the sample unit would be a grid cell, roughly corresponding to the size of plots used for the models. Then, the predictive model is applied at each sample unit to get the map of forest parameter estimates. A wall-to-wall forest canopy height map of a large portion of the southeastern US generated at the end of this stage would be very useful to forest scientists and the forest management community. However, we realized that implementing the second stage described above was non-trivial for us. That is, designing, implementing, and testing the software system to process seven terabytes of lidar data, then validating the end products, was beyond the scope of this study. Also, there would be significant computation time involved. Hence, this paper concentrates on the first stage (developing predictive models). We also decided to generate a canopy height map for a relatively small area (9.2×9.2 square kilometers), to test out the predictive model developed. Specifically, an area straddling Darlington and Marlboro counties in the state of South Carolina was chosen. This area had the following advantages: (1) The lidar coverage over that area was from a project that was quite representative of our lidar dataset; the acquisition was done in 2008, with a point density of 2–3 points/m²; (2) We had access to high-quality aerial orthophotography for that area from the summer of 2009; (3) Due to the presence of various land-cover features such as agriculture, managed tree plantations, natural forests, a river and some lakes, one could expect a good range of forest stand heights; (4) At such small scales, landscape level patterns are quite discernible. Hence, we could compare broad features of our map to high-resolution aerial orthographic photos.

The map was made with the following steps:

Step 1: Divide the area into grid-cells of size 120×120 m.

Step 2: Identify non-forest grid cells (urban, agriculture, water) and flag them as “non-forested”, with all other grid cells labelled “forested”. This was done by visual inspection of the associated high-resolution image. But, on a larger scale, this can be done by using existing land-classification products such as the National Land Cover Dataset (NLCD) of the United States [36].

Step 3: Calculate descriptive lidar metrics for all the forested grids, as described in Section 2.4 above. Then, compute the grid-level predicted dominant height by using the main model. Note that we use only lidar first-returns above a threshold of 3.0 m for computation of lidar metrics (to remove understory points). There might be some grid-cells in our region that have no points above this threshold. We do

not use the predictive model for such points, and denote their canopy height as zero. In other words, these are marked as “pure understory” grid cells.

3. Results

3.1. Main Model

After examining the correlations between plot-level lidar distributional metrics and dominant tree heights, the 85th percentile (h_{85}) of lidar heights were selected. Hence, dominant tree height was used as the dependent variable (*dom_tree_ht* below), while h_{85} and *cv_canPts* were tried out as candidates independent ones. The final OLS based linear regression model has the form:

$$dom_tree_ht = 5.658 + 0.859 \times h_{85} + 3.776 \times cv_canPts \quad (4)$$

The R^2 of the fit was 0.74 and the RMSE (root mean square error) was 3.0 m. Both independent variables were highly significant ($p < 0.0001$). Their individual correlation with dominant height can be seen in Figure 4. As the independent variables had the ranges of $3.0 < h_{85} < 42.0$ and $0.09 < cv_canPts < 0.85$ for model calibration, the model is only valid for these variable ranges.

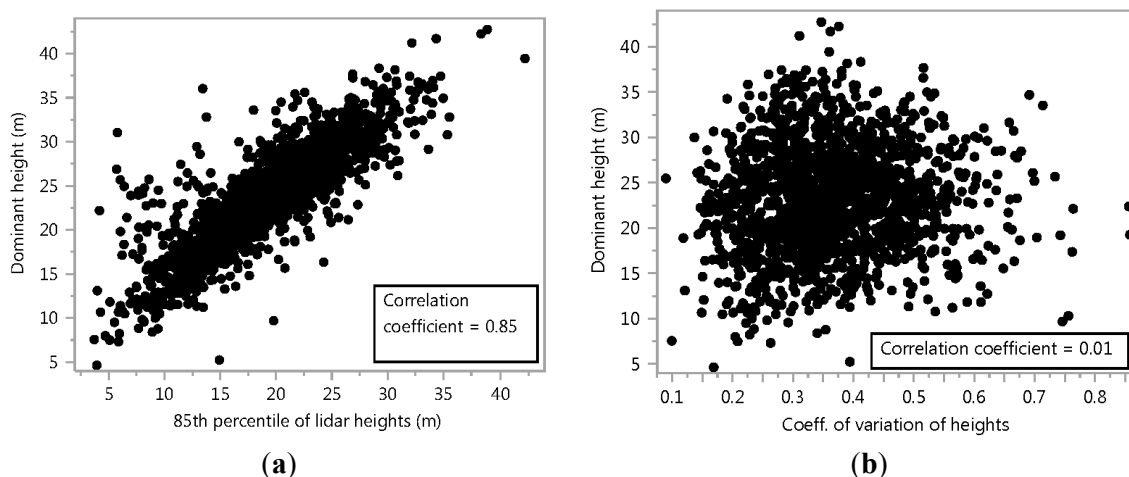


Figure 4. (a) Correlation between the 85th percentile of lidar first return heights and the FIA-measured average dominant tree height ($n = 1755$ plots) (b) Correlation between the coefficient of variation of the heights of canopy first returns (*cv_canPts*) and the FIA-measured average dominant tree height ($n = 1755$ plots).

3.2. Variable Importance for Goodness of Fit

As mentioned in Section 2.8, we analyzed the residuals of the linear fit above with respect to various factors. To study the effect of point density on the residuals, we made a standard boxplot of the residuals for various ranges of point density values (Figure 5a). The slight effect of better point density to affect better fits can be seen here, albeit quite muted. That is, the spread of residuals (measured by the difference in quantiles) changes slightly from 6.9 to 6.5 m, when one moves from low to high point density. Also, we do not detect significant differences in the variance of the residuals at various levels used, using the Brown–Forsythe test (test statistic = 2.174, p -value = 0.089). Nevertheless, the importance

of point density has been noted in various previous papers: The authors in [17] report that point density had even greater influence than footprint size on canopy height estimates. The relative unimportance of this variable for us may be due to our large plot size, which guarantees a sufficiently large number of returns even with low point density. The effect of plot homogeneity can be seen in Figure 5b. Plots with low homogeneity (such as the one in Figure 3b) have higher spread of residuals, whereas more homogeneous plots (Figure 3a) are well-modelled by the linear fit. The quartile range decreases by as much as 6.7 m (from 12.1 to 5.4) between the two extremes, in this case. This highlights the considerable importance of plot homogeneity in explaining the lack-of-fit of several plots. Also, the Brown-Forsythe test points to significant differences in the variances (test statistic = 55.05, p -value < 0.001). Hence, we note that the effect of plot-level vegetation heterogeneity is pronounced in this set of plots.

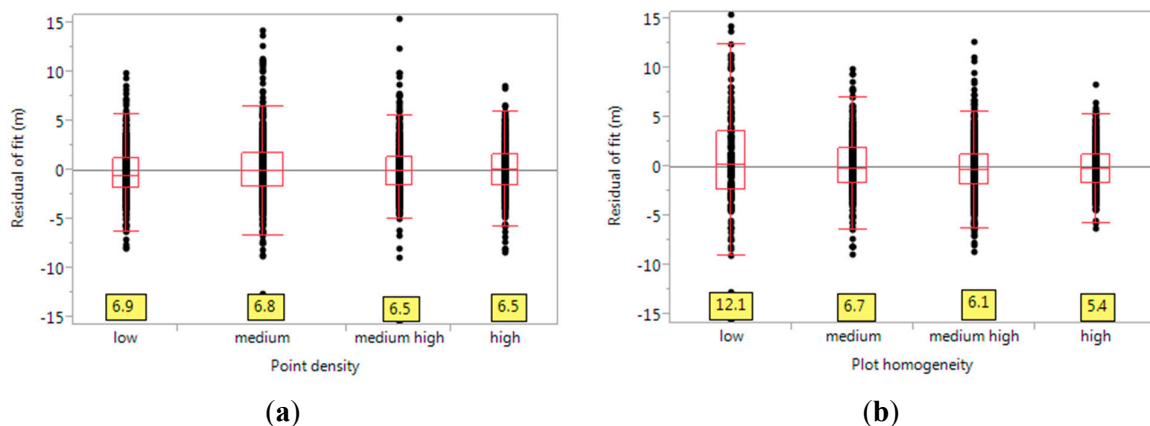


Figure 5. (a) The effect of point density on the residuals of the linear fit. Here, “low” denotes point densities of 0.25–0.5 points/m², “medium” is 0.5–2.0, “medium high” is 2.0–4.0, and “high” is above 4.0. The values at the bottom (yellow boxes) are the quantile ranges, between the 10th and the 90th quantiles (this is same for other figures). They represent the spread of the residuals, for various ranges of point density. Higher point density has the effect of reducing the range of residuals, giving a better fit; (b) Effect of plot homogeneity on the residuals. We use coefficient of variation (CV) to quantify plot homogeneity, where higher CV values denote lower plot homogeneity. Here, the label of “low” homogeneity denotes CV values above 0.5, “medium” is 0.3–0.5, medium high is 0.15–0.3, and high is CV below 0.15. The significant shrinking of the quantile range (from 12.1 to 5.4 m) for increasing plot homogeneity is notable.

We also calculated the effect of species grouping (softwood vs. hardwood) by estimating *percent_softwood*, the percentage (by basal area) of softwoods in the stand. These values were then binned and corresponding boxplots were generated (see Figure 6a). For example, the bin of *low* denotes that the softwood basal area is less than 25% of the total basal area, or that the stand is dominated by hardwoods. Likewise, *medium* is for basal area of 25%–50%, *medium high* is for 50%–75% and *high* is for 75%–100% (mostly softwoods). Here, it can be seen that hardwood stands are associated with higher quantile range than softwoods. This is because all our lidar acquisitions were done in a leaf-off vegetation state. On the other hand, when the stands were dominated by softwoods, the quantile range decreased by 1.7 m (from 7.2 to

5.5). These results are similar to those of [37], where it was reported that percentile methods have problems in estimating height in deciduous compound canopies during leaf-off conditions. Again, it was reported that model predictions improve when the study area is stratified into softwoods and hardwoods, with softwoods generating better R^2 values [14]. Also, the Brown-Forsythe test shows significant differences in variances (test statistic = 7.281, p -value < 0.001).

Another factor of importance is the average height of the trees in these stands (which is an indicator of the age class of the trees). This is explored in Figure 6b. For a stand with short trees, we over-predict height whereas we under-predict for tall trees. We attribute this to two reasons: (1) Tall trees indicate mature stands with more shrubs and understory, increasing the chance of errors in ground classification; (2) When the trees measured on the subplots are tall, they have higher probability to be the dominant trees in the buffered FIA plot (and to be associated with the 85th percentile of lidar returns). This also suggests that a better model may be a piecewise linear one, with a segment for the low tree height, another for the medium and medium high tree heights and one for the high tree heights. In general, we can see that there is very less difference between the variances for various levels of the factor. This is confirmed by the Brown-Forsythe (test statistic = 0.461, p -value = 0.709), indicating that average height of trees is not a significant factor regarding our model's inaccuracy.

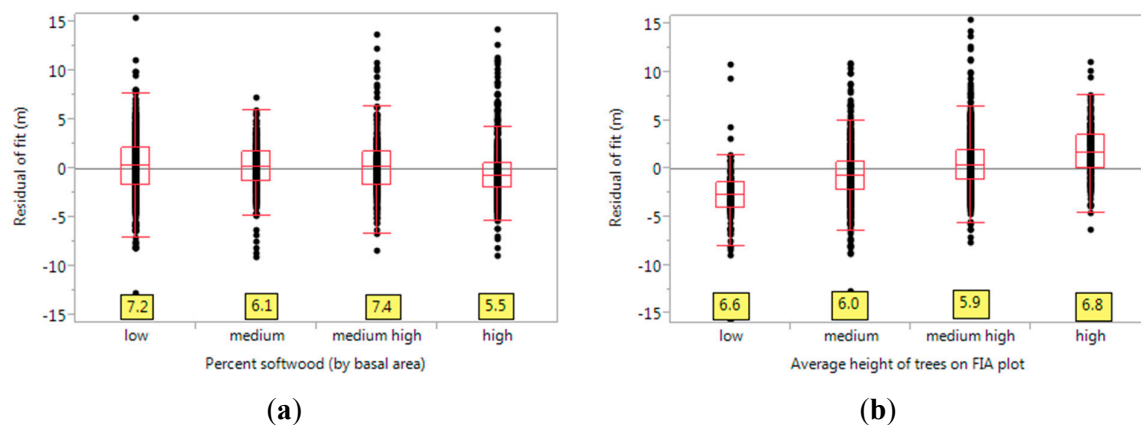


Figure 6. (a) The effect of species groups (softwood vs. hardwoods) on the residuals. The values at the bottom (yellow boxes) are the quantile ranges, between the 10th and the 90th quantiles. Here, softwood basal area is used for the categories. That is, “low” indicates a softwood basal area of 0%–25%, medium is 25%–50%, medium high is 50%–75% and high is more than 75%. The quantile range is much greater in the hardwoods case; (b) Effect of height of the stand on the residuals. The x-axis represents the average height of all trees measured by the FIA (in the four subplots). Here, “low” denotes stand height of 0–10 m, “medium” is 10–15 m, “medium high” is 15–20 m and “high” is more than 20 m.

We also calculated the effect of slope of the FIA plot (see Figure 7a). In this case, the Brown-Forsythe test only shows slight differences in variances (test statistic = 0.668, p -value < 0.572). Our model over-predicts heights on steep slopes. This may be due to the fact that these plots correspond to mountainous terrain, where there is usually more softwoods, for which our model over-predicts (see Figure 6a). Another reason may be the relative lack of understory on these drier sites. The effect of scan angle on the residuals is seen in Figure 7b. The relative lack of effect of both these factors on the spread

is mostly due to the large plot size used. This guarantees that the distributions are fairly stable over such changes.

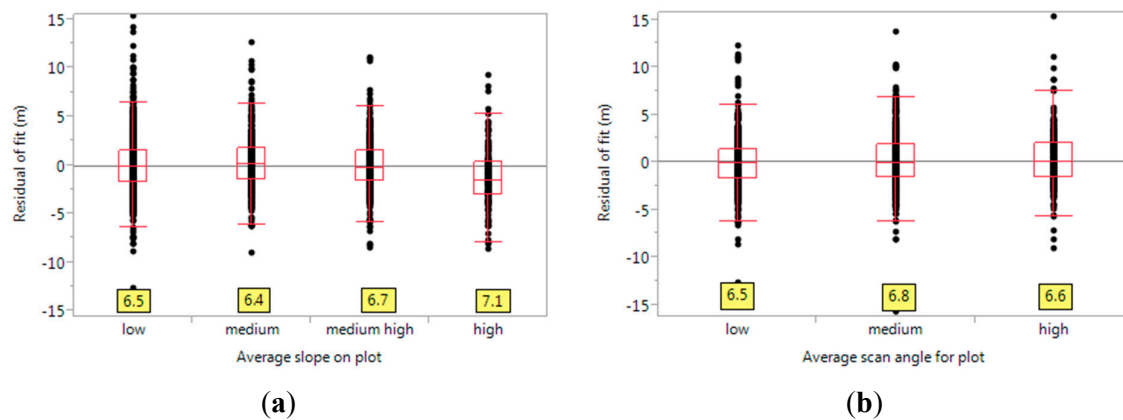


Figure 7. (a) The effect of average slope of the FIA plot the residuals. The values at the bottom (yellow boxes) are the quantile ranges, between the 10th and the 90th quantiles. Here, “low” indicates a slope of 0 or 1 degrees, medium is a slope of 1–10 degrees, medium high is 10–25 degrees and high is more than 25. The quantile ranges are similar, but higher for steeper plots; (b) Effect of lidar scan angle on the residuals. The x-axis represents the average scan angle, as recorded in the lidar metadata. Here, “low” denotes scan angles from 0 to 5 degrees, medium from 5 to 12 degrees, and high above 12 degrees.

3.3. Variable Importance from the Random Forest Model

We generated a random forest model using 1153 plots (see Section 2.7.). The model was quite robust, yielding an RMSE of 2.87 m with independent test data. The variable importance values are given in Table 1. One can note that the relative variable importance values differs from those from the main model. This is mostly because of the additional variables available to the random forest model. For example, it used the “percent softwood” variable and that variable was able to explain some of the variance in CV (for example, mixed stands of evergreen and deciduous trees can be expected to have higher CV). But the relative trends are the same. That is, plot homogeneity (as quantified by CV) and percent softwood are important variables for modeling stand heights, followed by slope. Scan angle and point density are not important at this plot size. The relative importance of these factors is useful to keep in mind when one tries to improve the precision of height estimates by stratification techniques. Plot slope, in general, has been found to be important previously too: For individual tree heights, it has been reported that underestimates of tree height were larger on lower elevations (−0.85 m) than on higher elevations (0.17 m) [38]. The importance of slope for the accuracy of Lidar-derived tree detection has also been studied [39]. It was reported that the accuracy of tree detection is 74% for steep slopes but is 86% for gentle slopes.

As plot homogeneity was established in the above analysis as an important variable, we examined its efficacy in explaining residuals of our main model in Figure 8. It can be seen that points with relatively high land-use homogeneity cluster nearer to the 1:1 line, and are hence better modeled. Again, most of the points far from the 1:1 line (*i.e.*, not well modeled by us) have low plot homogeneity.

Table 1. The variable importance values of the random forest model generated. The importance of a variable is estimated by seeing how much the prediction error increases when that variable is permuted, keeping all else unchanged [34]. The variables at the top of the table are more important. The Brown–Forsythe test statistic (which quantifies the spread of residuals of our main model) is also given for reference. NA: Not Applicable.

Variable	Increase in Error (%)	Brown-Forsythe Test Statistic
h_{85}	113.2	NA
Percent softwood	22.3	7.28
CV (measure of plot homogeneity)	15.0	55.05
cv_canPts	14.3	NA
Slope	9.0	0.67
Point density (PD)	4.2	2.17
Scan angle	3.3	0.41

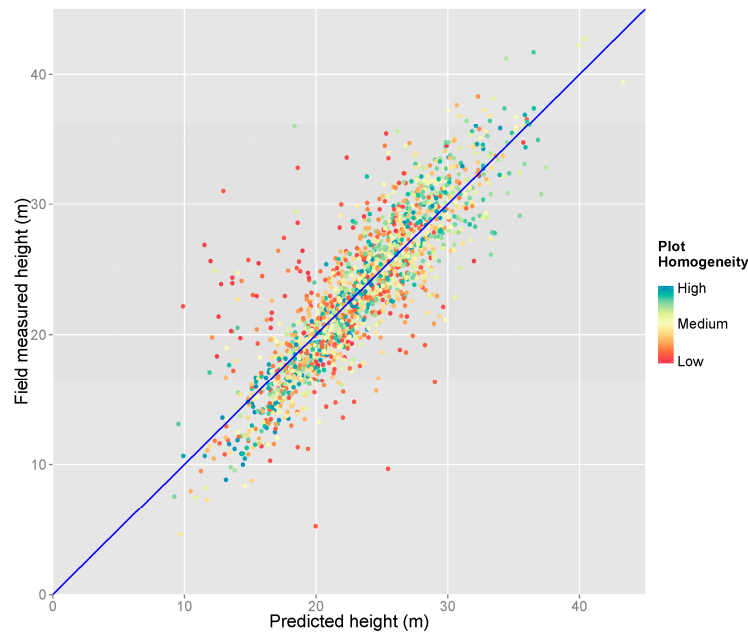


Figure 8. Scatterplot showing the agreement between subplot-level dominant height and the height predicted by our main model ($n = 1755$ plots). The color of each point indicates the land-use homogeneity of the field plot associated with it (measured by CV). The dark blue line is the 1:1 line between predicted and measured heights.

3.4. Sample Canopy Height Map

A sample canopy height map was generated over a 9.2×9.2 square kilometers area, as described in Section 2.8. (Figure 9). The range of dominant canopy heights that we generate (37 m) is reasonable for the area at hand: Loblolly pine is known to grow to a height of ~ 40 m in that region. Also, the spatial variation in canopy height seems to follow expected patterns on the landscape. The heterogeneity in vegetation heights along the river is picked up by the CV (dark green) pixels, as expected. These results indicate that a wall-to-wall canopy height model with bounded uncertainty can be generated from our lidar dataset.

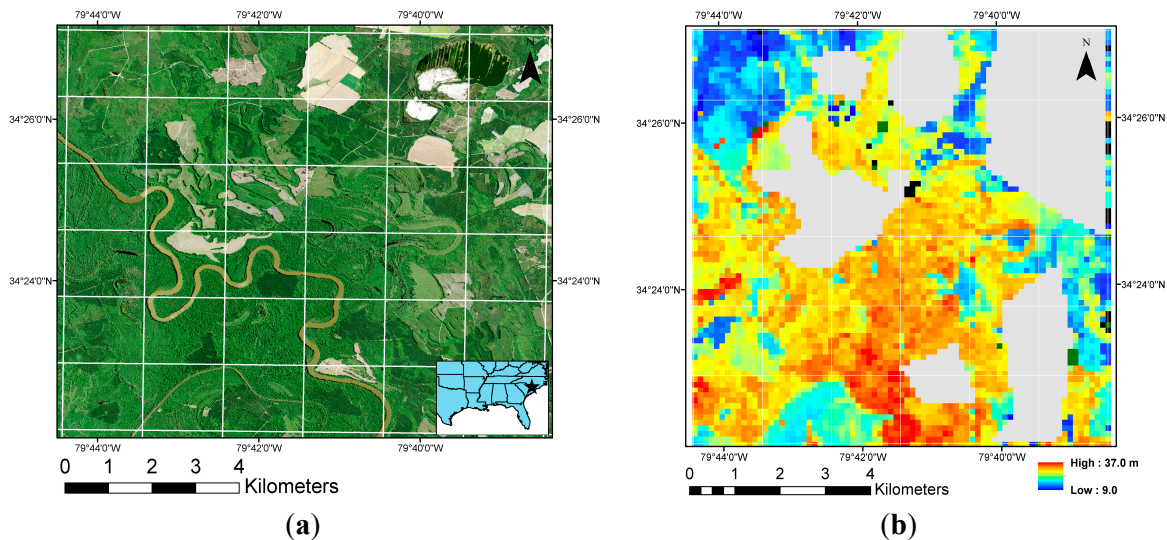


Figure 9. (a) A high-resolution aerial photo of the area used for verification of the model. The image is from the National Agriculture Imagery Program (NAIP) of the United States Department of Agriculture, acquired during the summer of 2009. The inset shows the location of the area in southeast United States. The coordinates of the center of the area are 34.410016N, 79.690276W (latitude, longitude; using WGS 84 datum). The white lines demarcate lidar data “tiles”, each of ~1.8 km wide; (b) Canopy height map for the area generated using our main model. The height is denoted in m, and the faint white lines correspond to the lidar data tiles, same as shown in Figure 9a. The grey area is non-forested land, the black pixels are “pure understory”, and the dark green pixels are where lidar metrics are otherwise beyond the range of our model (hence, the model is not applicable).

4. Discussion

4.1. The Influence of Co-Registration Issues

Good geographical co-registration of lidar and field plot data is essential for accurate prediction of forest stand biophysical properties from a lidar point cloud [40,41]. The location of FIA field plots are determined by GPS instruments whose positional accuracies are determined by a host of factors such as equipment, user, satellite, atmospheric and environmental [42]. Moreover, GPS signals are easily blocked or scattered by forest canopies (which often leads to multipathing) and hence it is problematic to get horizontal accuracies of less than 5 m on plot locations [43,44]. Hence, the error in our model is due to two reasons: (1) The inability of lidar to fully characterize the canopy profile. This may be due to factors such as incorrect estimation of the ground, undersampling of tall trees, lidar returns disproportionately reflected from the lower branches, and sensor errors; (2) The mismatch between the FIA and the lidar plots. The shape and size of the FIA and the lidar plots are different. The lidar plot needed to be designed this way to make sure that it had a good probability of containing the FIA plot. A consequence of this is that these plots could be covering very different patches of forests. For example, the FIA subplots may be over a patch of very tall pine trees, while there may be mostly very short beech trees over the lidar plots.

For understanding the effect of co-registration errors (and augmented plot size required), consider any of the 1755 lidar plots used for our main models. The total error of our estimate of canopy height for that plot has the following three components:

$$E_{\text{total}} = E_{\text{model}} + E_{\text{non-representative-subplots}} + E_{\text{FIA-measurements}}$$

where:

- E_{total} is the total error, which is the difference between the actual dominant height of the forest patches on that lidar plot, and that predicted by our models.
- E_{model} is the error in modelling *the dominant height of the trees on the FIA subplots* from the lidar data, on the lidar plot. This is captured by the RMSEs of our models, and may be due to several factors, as discussed in Section 2.7.
- $E_{\text{non-representative-subplots}}$ is the error caused by the fact that the dominant height of the trees on the FIA subplots (that we model) is not the same as the dominant height of the trees over the entire lidar plot. The difference can sometimes be large, may be due to the fact that the FIA subplots may be sampling a patch of forest with very different characteristics (see above) or that most of the area of the subplots could be outside the lidar plot (due to very high co-registration errors). But in general, this error term is expected to be lesser in areas where vegetation is more homogeneous. There have been recent reports that LiDAR metrics are less sensitive to co-registration errors in dense, spatially homogeneous stands than sparse, heterogeneous stands [40]. However, further work is needed to quantify this term.
- $E_{\text{FIA-measurements}}$ is the error in the FIA measurement, recording and processing of tree heights. We assume this to be relatively so small as to be negligible.

Given the adequate sampling of the area of interest by the large number of plots, it can be expected that the error for any grid cell of generated maps (such as those in Figure 9a) would be the same as E_{total} above. Also, in the absence of significant co-registration issues, the lidar point cloud over just the FIA subplots could have been extracted (instead of extracting a bigger buffered plot). In this case, the term $E_{\text{non-representative-subplots}}$ would be zero.

In a recent paper, it has been pointed out that in spite of the increasing acceptance of airborne lidar in forest research and operational forest management, there have been only a few published studies examining the impact of co-registration error on the prediction accuracy of regression estimators [40]. In that study, the authors examined various factors (mainly plot size and co-registration errors) contributing to the successful development of spatially extendable regression models. They used a large dataset ($n = 179$) of computer-generated coniferous forest canopies to conclude that regression models using larger plots (area of plot = 707 m²) were substantially more robust to the ill-effects of co-registration errors. Goodness-of-fit values like R^2 and RMSE (of the biomass, in this case) were shown to be less affected when the plot size was increased. For example, when the plot size was increased from 314–1964 m², the RMSE of forest biomass prediction expressed as a proportion of average biomass went down from 47% to 34%. In this context, FIA's relatively large plot size works to our advantage. There have also been recent attempts to correct FIA plot locations using high-density lidar data (which is assumed to have minimal registration errors) [45]. This is done by identifying and matching individual

tree crowns between the two datasets. But such efforts are not scalable, as the spatial extent of high-density lidar data (in which distinct individual trees can be identified) remains small.

4.2. Lidar-Based Grid Cell Homogeneity

A significant finding of this work was the efficacy of a lidar-based homogeneity metric in explaining the errors of our model. The CV metric was found to be important both in the main model and in the random forest model. To further study this, we considered two subsets of plots, which are more homogeneous: (a) $CV < 0.5$; (b) $CV < 0.2$. We developed OLS based linear regression models for these plots with *dom_tree_ht* as the dependent variable and *h85*, *cv_canPts* as the independent ones. The results of this exercise are summarized in Table 2. Hence, if one considers only areas of higher homogeneity (such as $CV < 0.2$), the RMSEs involved can be much lower, such as 2.44 m. These RMSEs only represent the E_{model} term (see Section 4.1., above). But it can be reasonably expected that the $E_{\text{non-representative-subplots}}$ term is also lower in homogeneous forest areas. If one assumes that $E_{\text{non-representative-subplots}}$ term is nearly zero when $CV < 0.2$, the E_{total} of the model is close to 2.44 m. This is comparable to recent estimates for much smaller areas: Erdody and Moskal reported an RMSE of 1.9 m for tree heights using lidar ($n = 57$) [27]. Also, the percent of original number of plots in the new subset gives an indication of the percent of the grid cells for which this new “homogeneous” model can be used.

Table 2. Goodness of fit statistics of models based on sets of more homogeneous plots.

CV Threshold	Number of Plots	% of Original Num. Plots	R^2	RMSE (m)
$CV < 0.5$	1573	89.6	0.81	2.60
$CV < 0.2$	729	41.5	0.84	2.44

4.3. The Advantage of Using National-Level Forest Inventory Plots

A major advantage of using FIA plots for constructing prediction models is the relatively large size of the plots involved ($\sim 675 \text{ m}^2$ or $\sim 0.07 \text{ ha}$). The four-point cluster plot design was specifically chosen to capture more spatial variability and hence reduce between-plot variance [22]. Large plot sizes have the following advantages: (1) Capture of more spatial variability: This is true for FIA plots because of their size and the four-point cluster design; (2) Minimization of the edge effect: The edge effect is an important aspect in the design of field plots. This is caused by vegetation at the edge of the plot being “counted in” in the field survey, but not in the remote sensing data (or vice-versa). Circular plots with large area minimize the edge effect [40]; (3) Reduce the effect of co-registration inaccuracies: It has been shown that larger plots have the potential to nullify the ill-effects of co-registration errors [40]. This is because the size of the plot increases the spatial overlap between the lidar and the ground plot, for unit co-registration error.

When regression models are used to predict forest biophysical parameters over the entire lidar coverage area, it is assumed that the set of reference plots captures the full spectrum of variability for both the predictor and response variables used [46]. This is so that the probability and magnitude of extrapolation errors (caused by vegetation structures not represented in the ground plot set) are decreased. In this context, the fact that we have a large number of randomly distributed ground plots works to our advantage. And, there is potential for significant improvement: We are confident that we

can use data from that entire set of FIA ground-reference plots of the study area, with better lidar coverage in the future. But a disadvantage of using FIA data also stems from their extensive coverage of the land area. The FIA design uses fully randomized plot designs, where plots are established, independent of the land use and the land use heterogeneity. Such randomized locations are essential for unbiased estimates of forest inventory parameters, which is the main focus of the FIA program. But these plots are not ideal for calibrating models based on remote sensed data, due to the problem of “mixed pixels” (plots where there are multiple land uses).

The third advantage of using FIA’s dataset is the robust quality assurance and management protocol in place [47]. A subset of the measured FIA plots are chosen to be re-measured by an independent field crew, as a “blind check”. This crew had no knowledge of the results of the original measurements—thus the term “blind check”. Measurement repeatability is the overall key objective and is defined in terms of a *tolerance* and a *measurement quality objective* (MQO); these are defined for each type of FIA measurement. The tolerance value is the acceptable range of variation allowed and a minimum proportion of the original measurements (MQO) are expected to be within tolerance. For example, for tree height, the quality objective is that the recorded height should be within $\pm 10\%$ of the true height (as determined by the blind check) at least 90% of the time. If discrepancies are found in such blind checks, steps are taken to identify the potential causes of low repeatability and possible biases, and ameliorate them.

4.4. Uses of Large-Area Canopy Height Maps

We used a simple linear regression model to predict canopy height over the region of lidar coverage. Point density of lidar returns, the homogeneity of the lidar plot, and the species grouping (hardwoods vs. softwoods), were all found to be important factors in the efficacy predicting canopy heights. Canopy height maps are instrumental for wildfire risk management: Canopy height is one of the four structural parameters that are essential for estimating wildfire-related canopy fuel load, the others being canopy base height, canopy bulk density and available canopy fuel weight [27]. Most contemporary canopy fuel metric maps for such large areas have been from the Landscape Fire and Resource Management Planning Tools Prototype Project (LANDFIRE), which has a 30 m resolution [48]. Use of lidar data will enable improved understanding of the spatial heterogeneity of fuel loading, in turn potentially improving parameterization of semi-empirical wildfire behaviour modelling programs such as FARSITE and FlamMap. Another innovative use of canopy height maps is to improve the precision of forest parameters [49]. Currently, the FIA uses Landsat data for stratification but a canopy-height map derived from our methods or similar is likely to improve the precision of forest parameter estimates.

There has been emerging interest in the concept of lidar plots, which may be employed either as a proxy or as a means of extending field plots [50,51]. It was pointed out in a recent paper that surrogates for field plots are especially welcome in remote and relatively inaccessible regions, where the installation of a single field plot can cost several thousand dollars [50]. The authors go on to describe the laying out of 17 million lidar plots (each of 625 m² area) that sample almost the entire breadth of the Canadian boreal forest with reasonable accuracy estimates. As for extending field plots, the double-sampling approach implemented using ground plots and aerial photo inventories has been used to improve forest parameter estimates [52]. An innovative scheme on the same lines is detailed in a paper by Parker and Evans [51]. The paper describes a similar double-sampling scheme, where timber volume is estimated

from individual tree diameter and height estimates. Lidar plots are used (instead of the usual phase 1 photo-plots) to increase the sampling intensity by a factor of nine. The lidar data adjacent to the ground plots is only used as input to a statistical framework for estimating timber volumes. The advantage of augmenting ground plot information with lidar plots is manifest in the better confidence intervals of the timber volumes by age class. When there is better co-registration, we can envision that data similar to ours will increase the accuracy of USFS-FIA county level estimates.

4.5. Possible Improvements to Lidar Data

A disadvantage of using data collected for non-forestry work is that they tend to be taken acquired in leaf-off conditions. For example, about 95% of the projects used in this study were done in the January to March timeframe, which decreases the utility of the data for forestry applications [18]. In addition, there is clear need for more standardized collection and dissemination of project metadata for lidar surveys. For example, most projects do not currently measure vertical accuracy over forest land cover explicitly as part of their pre-acquisition quality check. Such measures would be very useful for forestry applications. In fact, the American Society for Photogrammetry and Remote Sensing and the Inter-Governmental Committee on Surveying and Mapping recommend collecting a minimum of 20 checkpoints (30 is preferred) in each of the major land cover categories representative of the area for which the vertical accuracy of lidar data is to be verified [53]. Various recommendations for project planning and data processing for forestry have been made [15]. For example, it was recommended that a sensor capable of multiple returns (at least three) be used. We had to discard data from several projects (including coverage of the whole state of Louisiana) that did not label multiple returns. Also, there were noticeable artifacts in many datasets when lidar data were merged from overlapping flight lines, which can change the outcome of tree recognition procedures [51] and may have affected our accuracies as well.

5. Conclusions

This study is the first known effort to integrate the national forest inventory of USA (*i.e.*, the FIA) with extensive public domain non-forestry lidar data (which covers over 30% of the US [19]). We showed the feasibility of our methodology by generating a canopy height map over a sample area. Hence, the model and metrics developed can be used to generate a coherent, wall-to-wall canopy height map over large areas of the southeastern US. This is a first step towards constructing a national wall-to-wall vertical vegetation structure map in the US, using the area-based approach which can then be used to ask germane questions regarding forest inventories, forest health, carbon sequestration, wildlife habitat suitability, the maintenance of biodiversity and fire risk mitigation. These questions are highly relevant, given the interest of the public (and hence scientists and policy makers) in sustainable management of natural resources. Also, future lidar acquisitions over the southeastern US might contribute to a multi-temporal lidar data stack which has been demonstrated to be helpful in answering questions about forest change detection [54,55]. It has also been suggested that height information would be a useful additional input layer for the classification algorithms of the next generation national land cover dataset and national ecoregions dataset [19]. Our results also give confidence about extending this method to the understory layer (the FIA collects understory metrics on a subset of their P2 plots), similar to other efforts in this direction.

In conclusion, we find that the combination of multiple-return ALS data collected for non-forestry purposes and national scale forest inventory data is promising for use in an area-based approach for generating large-area estimates, provided that: (1) Metadata related to ALS campaigns is standardized and readily available, leading to easier workflows and better statistical analysis; (2) Vertical accuracy of lidar heights are assessed at several check-points for forest land-cover category; (3) Co-registration errors are quantified well, and are small. Moreover, continued effort in this direction have the potential to inform work in estimating forest vertical structure (especially the understory) over large regions, leading to better biomass and wildfire-risk estimates.

Acknowledgments

This work was funded by the PINEMAP project (pinemap.org) sponsored by the USDA's National Institute of Food and Agriculture (NIFA). We would also like to acknowledge the United States Department of Agriculture (USDA) McIntire-Stennis Formula Grant.

Author Contributions

Ranjith Gopalakrishnan contributed to the research design, collected the data, designed and implemented the software for lidar-FIA data intersection, and led the interpretation of results and manuscript writing. Valerie Thomas and Randolph Wynne secured funding, proposed and developed the research design, helped in the interpretation of results and in manuscript writing and revisions. John Coulston helped with the research design, FIA data management and the intersection efforts, and in the interpretation of results.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Gopalakrishnan, R.; Thomas, V.; Coulston, J.; Wynne, R. Efficacy of using heterogeneous Lidar datasets in predicting canopy heights over a large region. In Proceedings of the SilviLaser conference, Beijing, China, 9–11 October 2013.
2. Jensen, J.L.; Humes, K.S.; Vierling, L.A.; Hudak, A.T. Discrete return Lidar-based prediction of leaf area index in two conifer forests. *Remote Sens. Environ.* **2008**, *112*, 3947–3957.
3. Goetz, S.; Steinberg, D.; Dubayah, R.; Blair, B. Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. *Remote Sens. Environ.* **2007**, *108*, 254–263.
4. Gorte, R. *The Rising Cost of Wildfire Protection*; Headwaters Economics: Bozeman, MT, USA, 2013.
5. Marshall, D.J.; Wimberly, M.; Pete, B.; Stanturf, J. *Synthesis of Knowledge of Hazardous Fuels Management in Loblolly Pine Forests*; US Department of Agriculture, Forest Service, Southern Research Station: Asheville, NC, USA, 2008.
6. Wear, D.; Greis, J. The Southern Forest Futures Project: Summary Report. Available online: <http://www.srs.fs.usda.gov/pubs/42526> (accessed on 20 March 2015).

7. Oswalt, S.N.; Thompson, M.T.; Smith, W.B. *US Forest Resource Facts and Historical Trends*; US Department of Agriculture, Forest Service: Knoxville, TN, USA, 2009.
8. Turner, D.P.; Koerper, G.J.; Harmon, M.E.; Lee, J.J. A carbon budget for forests of the conterminous United States. *Ecol. Appl.* **1995**, *5*, 421–436.
9. White, J.C.; Wulder, M.A.; Varhola, A.; Vastaranta, M.; Coops, N.C.; Cook, B.D.; Pitt, D.; Woods, M. A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. *For. Chronic* **2013**, *89*, 722–723.
10. Næsset, E. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sens. Environ.* **2002**, *80*, 88–99.
11. Simard, M.; Pinto, N.; Fisher, J.B.; Baccini, A. Mapping forest canopy height globally with spaceborne lidar. *J. Geophys. Res. Biogeosci.* **2011**, *116*, doi:10.1029/2011JG001708.
12. Lefsky, M.A.; Harding, D.J.; Keller, M.; Cohen, W.B.; Carabajal, C.C.; Del Bom Espirito-Santo, F.; Hunter, M.O.; de Oliveira, R. Estimates of forest canopy height and aboveground biomass using ICESat. *Geophys. Res. Lett.* **2005**, *32*, doi:10.1029/2005GL023971.
13. Wulder, M.A.; White, J.C.; Nelson, R.F.; Næsset, E.; Ørka, H.O.; Coops, N.C.; Hilker, T.; Bater, C.W.; Gobakken, T. Lidar sampling for large-area forest characterization: A review. *Remote Sens. Environ.* **2012**, *121*, 196–209.
14. Nord-Larsen, T.; Schumacher, J. Estimation of forest resources from a country wide laser scanning survey and national forest inventory data. *Remote Sens. Environ.* **2012**, *119*, 148–157.
15. Evans, J.S.; Hudak, A.T.; Faux, R.; Smith, A. Discrete return LiDAR in natural resources: Recommendations for project planning, data processing, and deliverables. *Remote Sens.* **2009**, *1*, 776–794.
16. Næsset, E. Effects of different sensors, flying altitudes, and pulse repetition frequencies on forest canopy metrics and biophysical stand properties derived from small-footprint airborne laser data. *Remote Sens. Environ.* **2009**, *113*, 148–159.
17. Yu, X.; Hyypä, J.; Hyypä, H.; Maltamo, M. Effects of flight altitude on tree height estimation using airborne laser scanning. *Laser-Scanners For. Landsc. Assess. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2004**, *36*, 96–101.
18. Næsset, E. Assessing sensor effects and effects of leaf-off and leaf-on canopy conditions on biophysical stand properties derived from small-footprint airborne laser data. *Remote Sens. Environ.* **2005**, *98*, 356–370.
19. Stoker, J.M.; Cochrane, M.A.; Roy, D.P. Integrating disparate Lidar data at the national scale to assess the relationships between height above ground, land cover and ecoregions. *Photogramm. Eng. Remote Sens.* **2014**, *80*, 59–70.
20. Davidson, M.A.; Miglarese, A.H. Digital coast and The National Map: A seamless cooperative. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 1127–1132.
21. NOAA Digital Coast: Frequently Asked Questions. Available online: <http://coast.noaa.gov/digitalcoast/sites/default/files/files/1394462725/DigCoast-FAQs.pdf> (accessed on 20 March 2015).
22. Bechtold, W.A.; Patterson, P.L. *The Enhanced Forest Inventory and Analysis Program: National Sampling Design and Estimation Procedures*; US Department of Agriculture Forest Service, Southern Research Station: Asheville, NC, USA, 2005.

23. Forest Inventory and Analysis: Sampling and Plot Design (FIA fact sheet series). Available online: <http://www.fia.fs.fed.us/library/fact-sheets/data-collections/Sampling%20and%20Plot%20Design.pdf> (accessed on 20 March 2015).
24. McRoberts, R.E. The effects of rectification and Global Positioning System errors on satellite image-based estimates of forest area. *Remote Sens. Environ.* **2010**, *114*, 1710–1717.
25. Hoppus, M.; Lister, A. The status of accurately locating forest inventory and analysis plots using the Global Positioning System. In Proceedings of the Seventh Annual Forest Inventory and Analysis Symposium, Portland, OR, USA, 3–6 October 2005; Volume 36, p. 179184.
26. Axelsson, P. DEM generation from laser scanner data using adaptive TIN models. *Int. Arch. Photogramm. Remote Sens.* **2000**, *33*, 111–118.
27. Erdody, T.L.; Moskal, L.M. Fusion of Lidar and imagery for estimating forest canopy fuels. *Remote Sens. Environ.* **2010**, *114*, 725–737.
28. Mutlu, M.; Popescu, S.C.; Zhao, K. Sensitivity analysis of fire behavior modeling with lidar-derived surface fuel maps. *For. Ecol. Manag.* **2008**, *256*, 289–294.
29. Gauch, H.G. Prediction, parsimony and noise. *Am. Sci.* **1993**, *81*, 468–478.
30. Kenyi, L.W.; Dubayah, R.; Hofton, M.; Scharadt, M. Comparative analysis of SRTM-NED vegetation canopy height to lidar-derived vegetation canopy metrics. *Int. J. Remote Sens.* **2009**, *30*, 2797–2811.
31. Andersen, H.E.; Reutebuch, S.E.; McGaughey, R.J. A rigorous assessment of tree height measurements obtained using airborne lidar and conventional field methods. *Can. J. Remote Sens.* **2006**, *32*, 355–366.
32. Brown, M.B.; Forsythe, A.B. Robust tests for the equality of variances. *J. Am. Stat. Assoc.* **1974**, *69*, 364–367.
33. Martinuzzi, S.; Vierling, L.A.; Gould, W.A.; Falkowski, M.J.; Evans, J.S.; Hudak, A.T.; Vierling, K.T. Mapping snags and understory shrubs for a lidar-based assessment of wildlife habitat suitability. *Remote Sens. Environ.* **2009**, *113*, 2533–2546.
34. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R. News* **2002**, *2*, 18–22.
35. White, J.C.; Wulder, M.A.; Vastaranta, M.; Coops, N.C.; Pitt, D.; Woods, M. The utility of image-based point clouds for forest inventory: A comparison with airborne laser scanning. *Forests* **2013**, *4*, 518–536.
36. Homer, C.; Dewitz, J.; Yang, L.; Jin, S.; Danielson, P.; Xian, G.; Coulston, J.; Herold, N.; Wickham, J.; Megown, K. Completion of the 2011 national land cover database for the conterminous United States—Representing a decade of land cover change information. *Photogramm. Eng. Remote Sens.* **2015**, *81*, 345–354.
37. Wasser, L.; Day, R.; Chasmer, L.; Taylor, A. Influence of vegetation structure on lidar-derived canopy height and fractional cover in forested riparian buffers during leaf-off and leaf-on conditions. *PLoS ONE* **2013**, *8*, doi:10.1371/journal.pone.0054776.
38. Heurich, M. Automatic recognition and measurement of single trees based on data from airborne laser scanning over the richly structured natural forests of the Bavarian Forest National Park. *For. Ecol. Manag.* **2008**, *255*, 2416–2433.
39. Takahashi, T.; Yamamoto, K.; Senda, Y.; Tsuzuku, M. Estimating individual tree heights of sugi (*Cryptomeria japonica* D. Don) plantations in mountainous areas using small-footprint airborne Lidar. *J. For. Res.* **2005**, *10*, 135–142.

40. Frazer, G.W.; Magnussen, S.; Wulder, M.A.; Niemann, K.O. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. *Remote Sens. Environ.* **2011**, *115*, 636–649.
41. Gobakken, T.; Næsset, E. Assessing effects of positioning errors and sample plot size on biophysical stand properties derived from airborne laser scanner data. *Can. J. For. Res.* **2009**, *39*, 1036–1052.
42. Johnson, C.E.; Barton, C.C. Where in the world are my field plots? Using GPS effectively in environmental field studies. *Front. Ecol. Environ.* **2004**, *2*, 475–482.
43. Bolstad, P.; Jenks, A.; Berkin, J.; Horne, K.; Reading, W.H. A comparison of autonomous, WAAS, real-time, and post-processed global positioning systems (GPS) accuracies in northern forests. *North. J. Appl. For.* **2005**, *22*, 5–11.
44. Wing, M.G.; Karsky, R. Standard and real-time accuracy and reliability of a mapping-grade GPS in a coniferous western Oregon forest. *West. J. Appl. For.* **2006**, *21*, 222–227.
45. Gatzliolis, D. Advancements in Lidar-Based Registration of FIA Field Plots. Available online: <http://www.nrs.fs.fed.us/pubs/gtr/gtr-nrs-p-105papers/72gatzliolis-p-105.pdf> (accessed on 20 March 2015).
46. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 821.
47. FIA Field Guide Volume 1: Field Data Collection Procedures for Phase2 Plots, Version 6.0. Available online: http://srsfia2.fs.fed.us/data_acquisition/FG_6_SRS_Final_legal_NOV.pdf (accessed on 20 March 2015).
48. Rollins, M.G.; Frame, C.K. *The LANDFIRE Prototype Project: Nationally Consistent and Locally Relevant Geospatial Data for Wildland Fire Management*; Gen. Tech. Rep.; RMRS-GTR-175. U.S. Department of Agriculture (Forest Service): Fort Collins, CO, USA, 2006.
49. Hawbaker, T.J.; Keuler, N.S.; Lesak, A.A.; Gobakken, T.; Contrucci, K.; Radeloff, V.C. Improved estimates of forest vegetation structure and biomass with a lidar-optimized sampling design. *J. Geophys. Res. Biogeosci.* **2009**, *114*, doi:10.1029/2008JG000870.
50. Wulder, M.A.; White, J.C.; Bater, C.W.; Coops, N.C.; Hopkinson, C.; Chen, G. Lidar plots—A new large-area data collection option: Context, concepts, and case study. *Can. J. Remote Sens.* **2012**, *38*, 600–618.
51. Parker, R.C.; Evans, D.L. Stratified light detection and ranging double-sample forest inventory. *South. J. Appl. For.* **2007**, *31*, 66–72.
52. Avery, T.E.; Burkhardt, H.E. *Forest Measurements*, 5th ed.; McGraw Hill: Boston, MA, USA, 2001.
53. Liu, X. Accuracy assessment of lidar elevation data using survey marks. *Surv. Rev.* **2011**, *43*, 80–93.
54. Solberg, S.; Næsset, E.; Hanssen, K.H.; Christiansen, E. Mapping defoliation during a severe insect attack on Scots pine using airborne laser scanning. *Remote Sens. Environ.* **2006**, *102*, 364–376.
55. Vepakomma, U.; St-Onge, B.; Kneeshaw, D. Spatially explicit characterization of boreal forest gap dynamics using multi-temporal lidar data. *Remote Sens. Environ.* **2008**, *112*, 2326–2340.

Chapter 3

Shrub detection using disparate airborne laser scanning acquisitions over varied forest cover types

This chapter is being prepared for submission to the *International Journal of Remote Sensing*
(<http://www.tandfonline.com/loi/tres20>).

Most formatting used is according to the journal specifications.

Shrub detection using disparate airborne laser scanning acquisitions over varied forest cover types

Abstract: We explore the possibility of extending the national forest inventory based point data of understory presence using region-wide, disparate lidar data for the Southeastern United States. For this, we developed a simple inferential model that helps to understand the basic underlying relationships and associations between lidar predictor metrics and forest understory shrub presence over a wide range of forest types and topographic conditions. The model (a LASSO-based logistic regression model) had fair predictive performance (accuracy = 62%, kappa = 0.23). Hence, we were able to propose a set of biophysically meaningful predictor variables that represent understory (4), canopy (3), topographic conditions (1) and sensor characteristics (1). The single most important predictor variable was the understory layer canopy density, the ratio of lidar returns in the understory to those near the ground. Hence we demonstrate that the interplay of several factors affect understory vegetation condition. Overall, our work highlights the potential value of using lidar to characterize understory conditions.

1. Introduction:

The vertical structure of forest plant communities plays an important role in determining ecosystem health and function. For instance, the susceptibility of forest stands to deleterious external factors such as windthrow (tree uprooted or broken by wind), insect damage (defoliation or bark damage by insects) and wildfires is influenced by vegetation structure, such as the relative bulk of canopy stratification layers and the gap between them (Morsdorf et al. 2010). Moreover, forest vertical structure (especially the shrub layer) critically influences habitat suitability for wildlife (Camprodon and Brotons 2006) and thus exploring remote sensing techniques for shrub detection is quite significant from a biodiversity management standpoint. Specifically, the presence of an understory shrub layer can have important trophic significance as it increases the availability of food sources (arthropods and fruits) for birds and other animals. Shrubs are a well-defined component of the forest vertical structure, and they are defined as plants with medium to small, woody, perennial aboveground stems, mostly less than 3 meters tall (Kimmins 2004). Along with trees (defined as plants with large, woody, perennial aboveground stems, generally taller than 3 meters) they form two of the five major layers of forest vertical structure.

Possible changes in wildland fire regimes have become a societal concern in the United States in the recent years, with increasing media reports of wildfires, especially very large ones (> 100,000 ha). There are many pockets of high wildfire potential along the Southeastern coast and the Florida panhandle (Oswalt, Thompson, and Smith 2009). In fact, the Southeast is the leading region in the United States regarding the number of wildfires per year: approximately 30,000 to

50,000 fires per year were reported from 2002 to 2006 (Andreu and Hermansen-Baez 2008). Some examples of large and dangerous wildfire conflagrations in the recent past are the Volusia Fire (44,900 ha), the Flagler/St. John Fire that occurred in Florida in 1998 (38,300 ha), the Okefenokee Fires of 2007 in Georgia and Florida (over 243,000 ha) and the 2008 Evans Road Fire in North Carolina (over 16,500 ha) (Wear and Greis 2011). Also, there has been an increase in residential developments along the Wildland Urban Interface across the South over the past 30 years, which makes such wildfires even more perilous (Andreu and Hermansen-Baez 2008; Stewart et al. 2005). Factors such as extensive fire suppression since the 19th century, combined with grazing, and the increasing effects of climate change are thought to play major roles in these estimated change (Westerling et al. 2006). In this context, it is important to recognize the role of forest understory in determining fire behaviour: the understory has the potential to act as both surface fuels (by which fires can propagate along the ground, in the shrub layer) or as ladder fuels (by which fires can propagate to the canopy). Unplanned wildfires are especially problematic in managed forest stands (common in the Southeast) where biomass consumed by fires result in lost harvest profits. Additionally, many of the effects of fires on soils and hydrology are deleterious to future stand regeneration and productivity. Despite the importance of understory for wildfires, little is known about the spatio-temporal variability of understory presence across large swaths of the United States. Hence, exploring the possibility of understory detection using passive and active remote sensing methods (the focal point of this work) is therefore critical to the task of assessing fuel loads and predicting fire behaviour in the Southeast US.

Predicting understory vegetation is challenging (Martinuzzi et al. 2009; Wing et al. 2015; Jakubowski, Guo, and Kelly 2013). The ability to predict the presence (versus absence) of understory shrubs from lidar data for managed, mixed-temperate coniferous forest stands on topographically complex terrain has been examined in Martinuzzi et al., (2009) (Martinuzzi et al. 2009). In that study, regression trees (i.e., the randomForest algorithm) using three lidar metrics (two associated with height bins and a third topographic one) was able to predict presence/absence with 83% accuracy. In a different study by Wing et al., the importance of high-quality lidar data and field data for understory cover prediction has been illustrated on an interior ponderosa pine forest site in Western USA (Wing et al. 2015). There, they stem-mapped every suitable shrub and also measured their crown dimensions, thus giving very robust field estimates of understory cover for the whole plot. Their modeling effort yielded low RMSEs of ~6% for shrub cover. Understory density is well-predicted with high point density lidar (~6 returns per square meter) for managed loblolly pine (*Pinus taeda* L.) forests; the leaf area index at 1.0 and 2.5 meters above the ground (thus bracketing the understory layer) was predicted with model R^2 levels of as much as 0.75 (Sumnall et al. 2016). Morsdorf et al., (2010) showed that small footprint ALS data could be used to identify and distinguish between three different vegetation strata in a multi-layered Mediterranean ecosystem (Morsdorf et al. 2010). There, the importance of intensity information was demonstrated: they noted that including ALS echo intensity as an additional feature was crucial in separating layers that could not be separated by pure return height information alone.

Leaf-off lidar has also shown to be useful in detecting understory invasive species in urban environments (Singh, Davis, and Meentemeyer 2015). In that study, the importance of topography variables and the spectral attributes of lidar returns (i.e., the intensity values) to

detect and map such species were noted. Several studies have used discrete-return airborne lidar data to aid in the characterization of wildland species habitats and predation risk (Zellweger et al. 2014; Lone et al. 2014; Nijland et al. 2014; Simonson, Allen, and Coomes 2012). Terrestrial lidar has also been used to estimate biomass and leaf area index of shrubs in the arctic region (Greaves et al. 2015). There has also been attempts to quantify understory presence using optical remote sensing data such as those from landsat. Pduzzi et al. assessed whether understory presence could be sensed from landsat images by examining the correlation between understory leaf area index (LAI) and landsat derived vegetation indices (Pduzzi, Allen, and Wynne 2010). Moderate, but significant, correlation between understory summer LAI and a landsat-based vegetation index (0.43) was reported. The difference between spring and winter landsat-based LAI (attributed to deciduous understory presence) has also been shown to be a promising correlate for understory presence in some forest types (Blinn et al. 2012).

A significant drawback of previous efforts to map understory abundance is that they do not explore strategies to produce understory maps over much larger areas. Another drawback is that the prior works noted above were performed over small study areas and therefore the associated results (such as the best metrics for understory detection) cannot be extrapolated to other dissimilar areas and forest types. Addressing these two drawbacks is the main motivation of our study. Given that there is no single national-level lidar data acquisition project in the US (or plans of one) as of 2016 to aid in the generation of such large area maps (such as the entire Southeast USA), we explore the use of disparate lidar projects instead. This is used in conjunction with national-inventory field data from the Forest Inventory and Analysis (FIA) wing of the United States Forest Service (USFS) (Bechtold and Patterson 2005).

In this work, we examine how well understory shrub cover (i.e., below a forest canopy) can be estimated from a disparate lidar dataset. Our work focuses on generating an inferential, interpretable model that helps to discern the underlying relationships between such heterogeneous remotely-sensed and field-measured data over a large spatial extent. The utility of an inferential model versus a purely predictive one for understanding underlying data relationships has been well reviewed and established (Hastie et al. 2013). Our specific research objectives for this work are: (a) to explore the prediction of understory presence/absence over a large area spread over several forest types by combining data from disparate lidar acquisitions, (b) to interpret key drivers of the predictions, from a physical standpoint. It has been shown that such a parsimonious model (with just two lidar metrics) can reasonably estimate plot-level canopy height over 76 different lidar datasets (Gopalakrishnan et al. 2015). This work extends the effort to the understory, exploring the possibility of a similar area-based approach (White et al. 2013) to understory mapping. That is, we ask the question of whether FIA understory point-measurements can be similarly extrapolated to larger spatial and ecological scales using airborne lidar data.

2. Methods:

Study Area and Lidar Data Used:

The study area was potentially all forested regions with adequate lidar coverage in the southern states of USA (thirteen states; bold black outline in Figure 1). Public-domain lidar datasets

managed by three federal agencies were acquired as part of this study. These agencies were: (1) The U.S. Geological Survey (USGS); (2) National Oceanic and Atmospheric Administration (NOAA); and (3) Natural Resources Conservation Service (NRCS). The first two are coordinating agencies of national-level lidar data sharing partnerships. In this capacity, they maintain web-based interfaces with which all lidar data collected by numerous partner agencies are disseminated. From these, we could get lidar data only for nine states, namely Alabama, Arkansas, Florida, Georgia, Louisiana, North Carolina, South Carolina, Tennessee and Texas. The spatial extents of these acquisitions can be seen in Figure 1. The forests found in these regions are quite varied, ranging from broadleaf (e.g., oak-hickory and sweetgum/yellow-poplar forests), to conifer forests (e.g., both natural and plantation pine stands) or even mixed forests (e.g., oak-pine, oak-gum-cypress) (oak is genus *Quercus*, hickory is the genus *Carya*, pine is the genus *Pinus*, gum and sweetgum are the genus *Liquidambar*, cypress is the family *Cupressaceae*, yellow-poplar is *Liriodendron Tulipifera*).

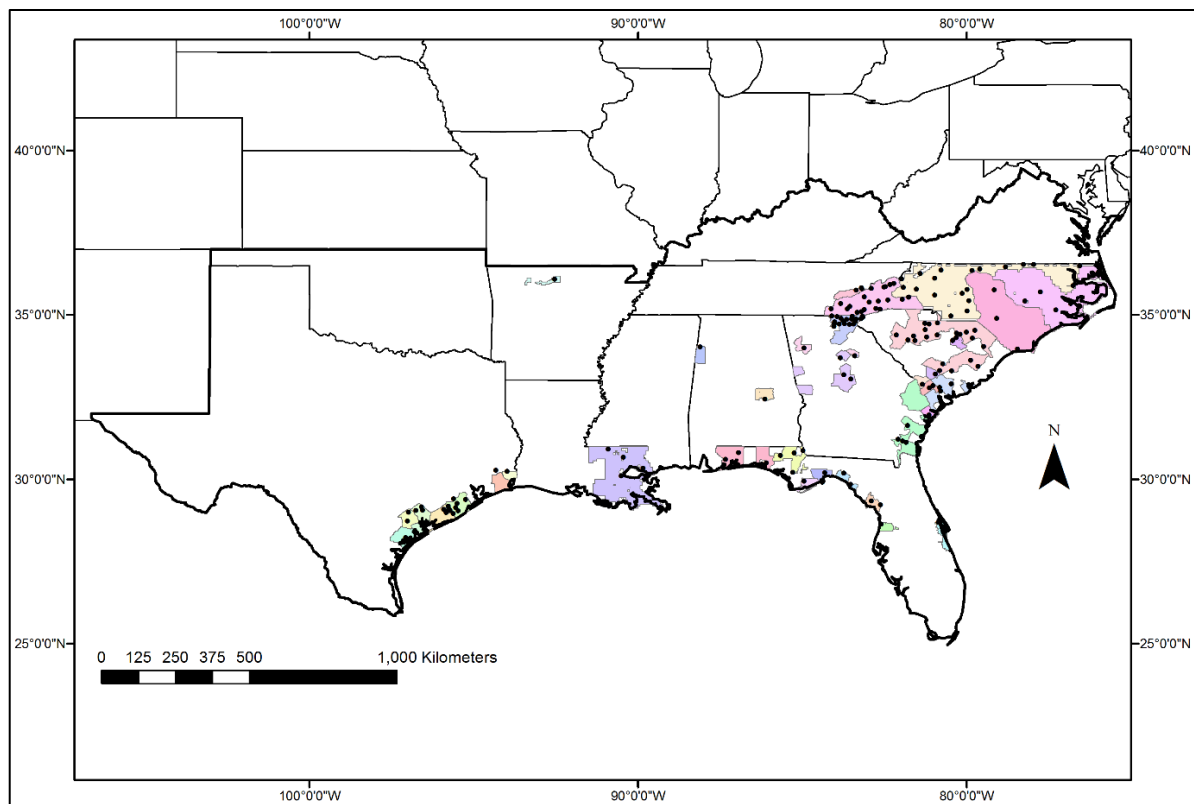


Figure 1: Our study region, which potentially includes all forested areas with lidar coverage in 13 states of the Southeast USA. Each colour represents a different lidar acquisition (total of 37). The black dots represent the location of the FIA P3 plots (explained later). The lidar acquisition extent span several major ecological and physiographic zones of the Southeast USA.

Region-wide field data:

We used shrub cover measurements from phase 3 plots of the FIA (Westfall and Woodall 2007) (henceforth FIA plots). Figure 2 shows the layout of one such plot. Understory is measured only at a limited area; only the four microplots shown (53.8 m² in total). Shrub cover was calculated as the sum of live shrub cover and dead shrub cover (reported separately by the FIA). These

shrub cover values are ocularly estimated by the FIA crew and this contributes to a lack repeatability of measurements. But Westfall et al. reported acceptable repeatability metrics for shrub cover: 64% for live shrub cover and 89% for dead shrub cover (Westfall and Woodall 2007).

Categorical dependent variable for understory presence:

We derived our dependent variable from the four microplots measurements in the following fashion:

1. Take the average of the shrub cover measured for the four microplots;
2. Categorical classification: Assign the plot into a categorical class, based on the computed average. The Multi-Resolution Land Characteristics Consortium had established a 25% shrub cover threshold to define the class “shrubland” for the 1992 Land Cover map (Vogelmann et al. 2001). Moreover, a 25% threshold has been proposed for assessing habitat suitability for woodpeckers (and for all bird species, overall) (Martinuzzi et al. 2009). Hence, we classify a forest plot as “understory present” or “understory absent” based on the 25% average shrub cover threshold, along the lines of previous literature.

Lidar plot size and metrics computed:

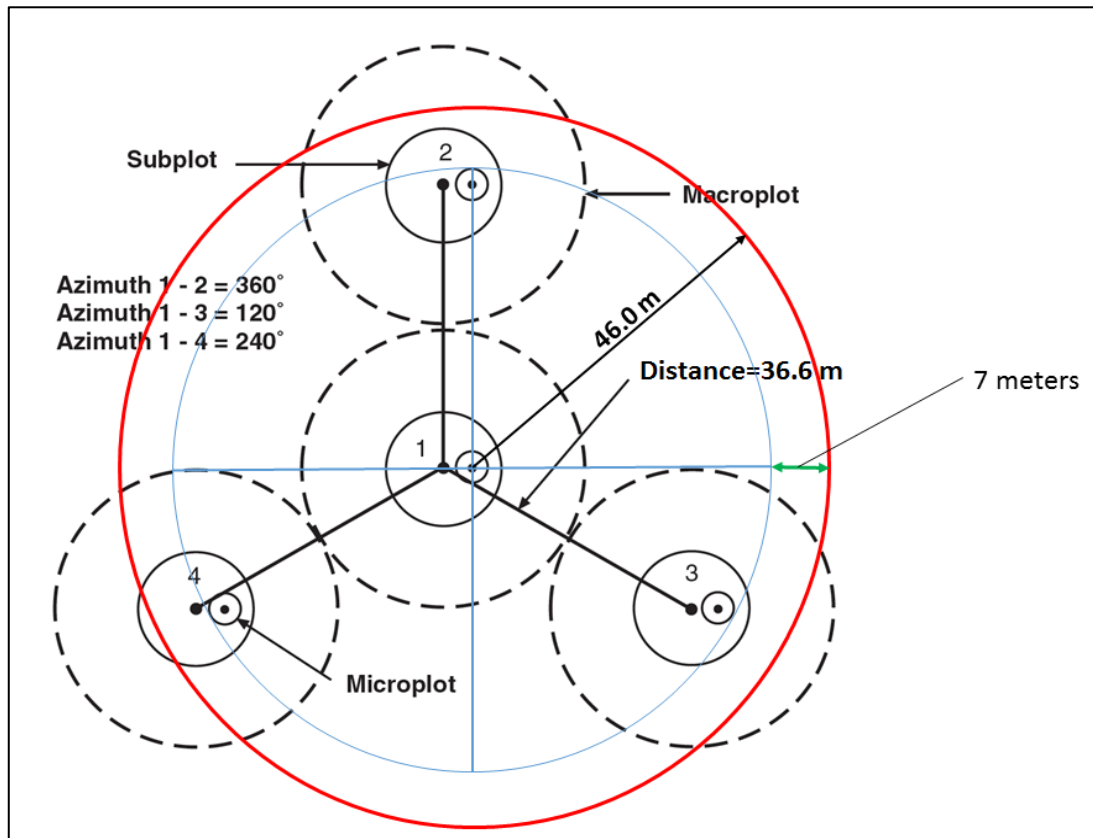


Figure 2: Size and position of the lidar plot, with respect to the FIA plot. Figure adapted from (Bechtold and Patterson 2005). The black lines and circles mark the FIA plot, the blue line and

circle denotes a circular plot that exactly subtends all four FIA microplots (radius = 38.6 meters) and the red circle denotes our lidar plot (radius = 46 meters) which is obtained after applying an additional buffer of 7.0 meters.

The *lidar plot* is the plot used for extracting the point cloud (from the region-wide lidar data) corresponding to the FIA plot. There are two main considerations for determining the lidar plot size and shape:

1. The plot should encompass all four FIA microplots. This implies that a circular plot of radius 38.6 meters, centred at the centre of the FIA plot.
2. The plot should be buffered so as to account for possible co-registration errors in FIA plot location data. There have been a few field studies that estimated the typical Global Positioning System (GPS) location errors associated with the FIA plots. Researchers at the USFS FIA program have come up with a probability density function of GPS errors, based on extensive field trials and review of other factors (McRoberts 2010). From that function, one can conclude that most GPS errors were within 15 meters. Recently, the FIA had done a more rigorous field study, where they used survey grade GPS instruments, and the re-measurements indicated that a majority of GPS measurement errors of plot location are within 7.0 meters (M. Russell, personal communication, October 2015).

Hence, we decided to use a circular plot buffered by 7.0 meters. The position of the circular lidar plot with respect to the FIA plot elements can be seen in figure 2.

The steps involved in the extraction of lidar plot-level data from the region-wide lidar data were similar to that described in (Gopalakrishnan et al. 2015). They are:

Step 1: For any given FIA P3 plot location, we checked for the presence of a spatially corresponding lidar acquisition within ± 2 years of the plot measurement. If one was available, a circular lidar plot centred at the *recorded* centre-point location of the FIA plot (as recorded by the FIA; which may be several meters off, see discussion above) was cut out from the lidar dataset. We extracted a total of 277 such lidar plots from the lidar data.

Step2: We used the *lasground* tool of *lastools* (version 141117; <http://rapidlasso.com/lastools/>) for classification of the returns on the lidar plot (ground versus non-ground), the default step size of 5.0 meters was used. Lastools uses the established method of progressive TIN (triangulated irregular network) densification for this.

This resulted in a set of plots for which both lidar data and FIA P3 field measurement data (including shrub cover measurements) were available. All further analysis and modeling was done on this set of paired elements.

We calculated several metrics over these lidar plots for characterization of plot understory, overstory and topographic conditions (see table 1). Many of these metrics have been shown to be

useful in predicting various forest structure attributes, by previous studies. Additionally, we calculated several *horizontal lidar metrics*, which quantify the horizontal structure of forest vegetation on the plot. That is, they try to capture the spatial distribution of plot vegetation with respect to the planimetric (i.e., x, y) coordinates. Such metrics have been shown to capture spatial characteristics that help separate natural forests from plantations (Sverdrup-Thygeson et al. 2016). The horizontal metrics we calculated are:

1. $pGridCells_{LPI>5\%}$: This metric is an evaluation of the proportion of the plot for which adequate amount of sunlight penetrates through the canopy and reaches the understory. It is based on the laser penetration index (LPI), which has shown to be well-correlated to LAI (Peduzzi et al. 2012) and is hence a good proxy for canopy cover. We calculated LPI_{3m} at the plot level in the following way: it is calculated as the percent of lidar returns of height less than 3.0 meters (ie, penetrates the canopy), out of the total returns from the plot. That is:

$$LPI_{3m} = \frac{n_{understoryReturns(h<3.0m)}}{n_{allReturns}} * 100.0 \quad (1)$$

The metric $pGridCells_{LPI>5\%}$ is an extension of this; we first defined a *lidar square-plot* as a square plot of size 100x100m, centred at the centre of the lidar plot, and oriented in a north-south direction. Then, this lidar square-plot is gridded into 400 smaller 5x5m gridcells. For each such gridcell, the LPI is calculated in the same fashion as LPI_{3m} above. The light penetration for such a gridcell is assumed to be “adequate” if its LPI_{3m} is more than 5%. Then, $pGridCells_{LPI>5\%}$ is calculated as the proportion of such “adequate” gridcells, for the whole plot.

2. $stdev_{h90_3m}$, CV_{h90_3m} : These two metrics essentially capture the “patchiness” of the understory: whether its height is uniform or uneven over the lidar plot. For these horizontal metrics, the lidar square-plot was divided into 100 gridcells, each of size 10x10 meters. For each such gridcell, understory returns are defined as returns below a height of 3.0 meters. The 90th percentile height of these returns is calculated for each gridcell. Then, the variation in these individual heights were calculated: standard deviation for $stdev_{h90_3m}$ and coefficient of variation for CV_{h90_3m} .
3. $range_{DEMhts}$, mad_{DEMhts} , sd_{DEMhts} , $skew_{DEMhts}$, $kurt_{DEMhts}$, IQR_{DEMhts} , var_{DEMhts} : Horizontal metrics related to topography. Statistic capturing the variation of lidar-generated DEM heights over the FIA plot. For this, the lidar square-plot was divided into 100 smaller 10x10m gridcells, and DEM heights are calculated on each of these gridcells. The aggregation statistics used for these metrics were range, mean absolute deviation, standard deviation, skewness, kurtosis, interquartile range and variance, respectively.
4. EPC : Effective plot coverage. This metric is an evaluation of the proportion of the plot that has been adequately sampled by lidar returns. For calculating this, the lidar square-plot is gridded into 1000 smaller 1x1m gridcells. A gridcell is assumed to be “adequately covered” if there is at least one understory return from it. Then, EPC is calculated as the proportion of such adequately covered gridcells in the plot. That is:

$$EPC = \frac{n_{adequatelyCoveredGridcells}}{n_{allGridcells}} * 100.0 \quad (2)$$

Information about both overstory features (such as canopy cover, canopy structure, leaf area index and species characteristics) and acquisition methodologies (such as point densities, scan angles and scanning pattern) are intrinsic in this metric (Wing et al. 2015).

From the descriptions above, one can note that different gridcell sizes were used for each horizontal metric (e.g., 5x5 meter gridcells for $pGridCells_{LPI>5\%}$, 10x10 meter gridcells for $stdev_{h90_3m}$ and cv_{h90_3m}). This was done after examining several plots and estimating the resolution that best captured the vegetation or topography parameter in question, for each horizontal metric.

Multiple land use FIA plots: The current FIA plot design involves fully randomized plot locations: plots were laid out based purely on randomized co-ordinates. Hence they may not be fully forested, and can potentially straddle multiple land use types, such as two different forest types, or a forest and a meadow. Such different land-uses on the same plot are referred to as “conditions” in the FIA literature. Usually, randomized plots for sampling forest areas have been moved to ensure that they are homogeneously forested. For example, a major study that attempted to map forest fuels (including understory) using optical remote sensing methods used over 300 forested plots (Rollins, Keane, and Parsons 2004). These circular plots “were established ≥ 50 m from any edge that represented a distinct boundary between cover types or structural stages”. Similarly, to ensure a degree of homogeneity of forested lidar plots, we also screened out lidar plots that had straddled more than one major land use type. In this context, the following metrics were defined:

1. Percent forested, estimated from FIA data: This is the proportion of the FIA plot denoted as being in a forested land use.
2. Percent forested, estimated from lidar data ($pFor_{lidar}$): The proportion of the 100x100 meter lidar plot that is forested, is estimated. This is done as follows: A uniform grid is used to divide the plot into 100 square gridcells, each of size 10 meters. For each such gridcell, it is assumed to be forested if the 95th percentile height of returns from it are over a height of 3.0 meters. That is, these returns are assumed to be canopy returns, and their presence indicates that there is vegetation from trees in that gridcell. Then, the proportion of gridcells which are thus labelled as forested is taken as the lidar-based estimate of percent lidar plot being forested.

Then, the following screening criteria are used, so that partially forested FIA plots are screened out:

1. Over 80% of the FIA plot is forested, as estimated from FIA data.
2. Over 80% of the lidar plot is forested, as estimated from lidar data.

We were left with lidar 153 plots after the above screening criteria have been applied. These set of plots can be assumed to be forested plots, with the effect of other land uses being minimal. All later analysis and modeling was done on these set of 153 plots. The approximate location of these plots can be seen in figure 1.

Heterogeneity of lidar acquisition parameters, forest types sampled:

The final set of 153 lidar plots represents a variety of lidar projects with various acquisition parameters (figure 3). This results in a variety of forest types and physiographic zones being sampled (figure 4) covering numerous ecological zones (figure 1).

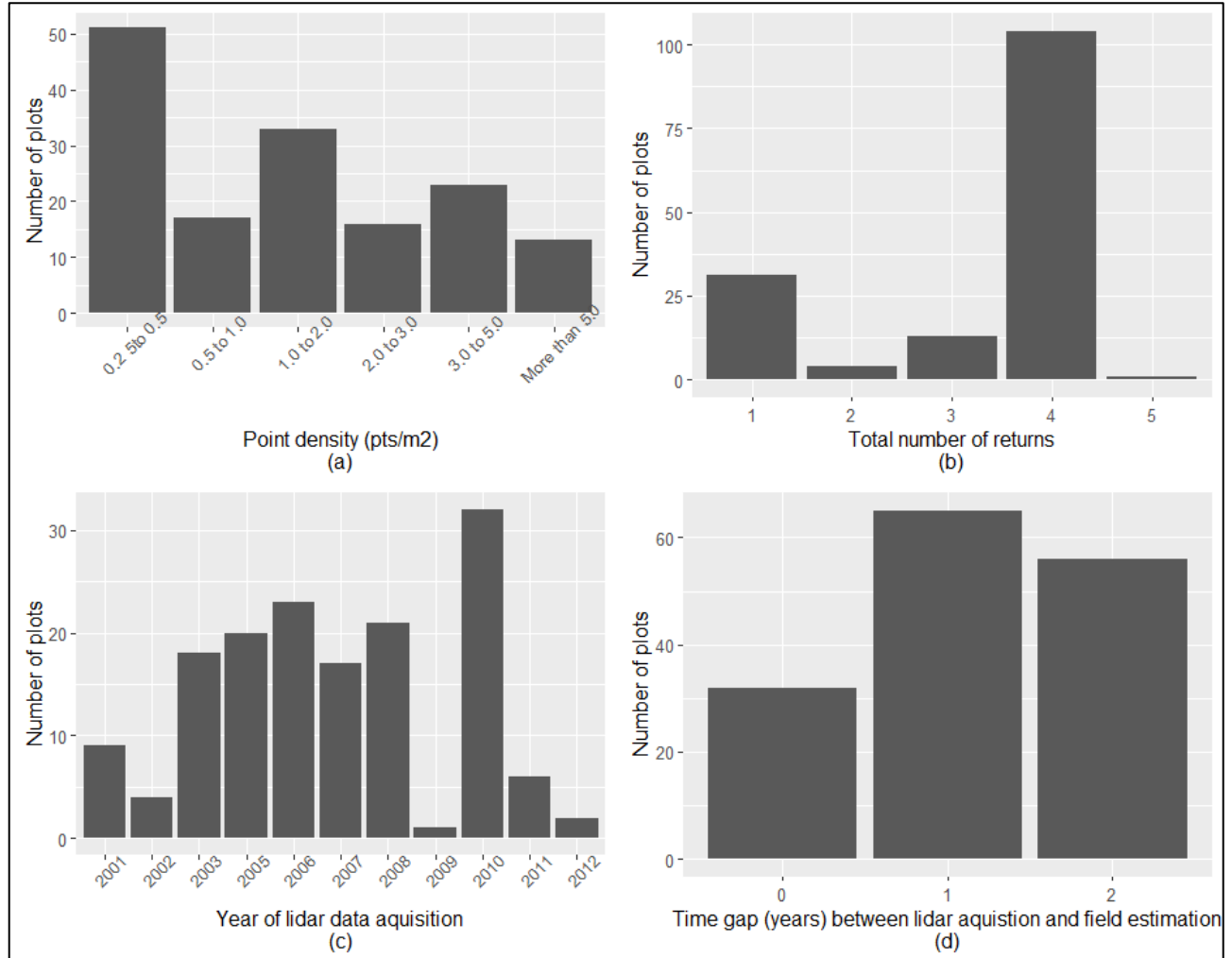


Figure 3: The distribution of various lidar acquisition parameters, for the 153 lidar plots used. (a) Point density of the various projects. The median point density is 1.22 pts/m²; (b) Total number of returns labelled. Lidar data on about 20% of plots ($n=31$) are from single-return systems; (c) Year of lidar data acquisition, which represents a good sampling of various technologies involved; (a) Time gap between lidar data and field measurement. Note that around 36% of plots ($n=56$) have a time-gap of 2 years; this increases the chance of understory conditions being substantially different between field measurement and lidar acquisition time.

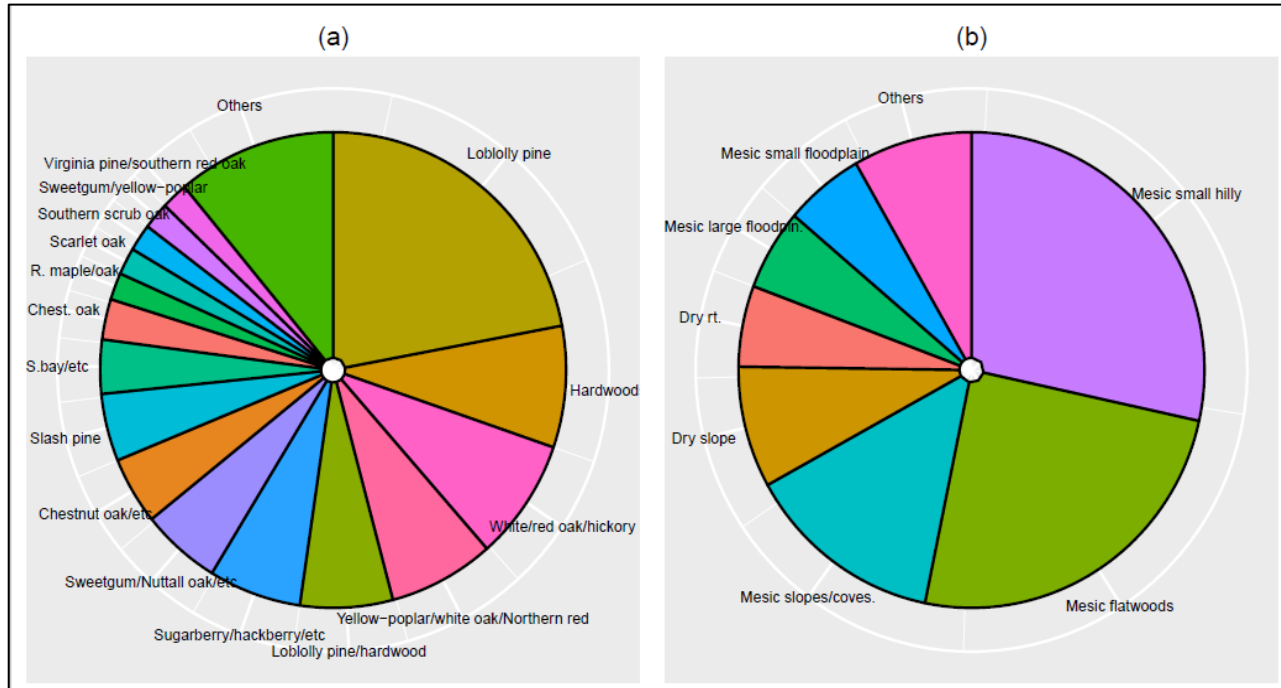


Figure 4: (a) Proportion of various forest types represented (as reported by FIA) in the 153 plots used. Some of the names have been abbreviated from their longer versions, for chart readability and clarity. That is, “S. bay/etc” denotes the “sweet bay/swamp tulepo/red maple” mixed-forest complex (where sweet bay is *magnolia virginiana*, swamp tulepo is *nyssa biflora*, red maple is *acer rubrum*), “Chest. oak” is “chestnut oak” (*quercus prinus*), “R. maple” is “red maple” and “Hardwood” represents “mixed upland hardwoods” class of forests. (b) The physiographic class of the landform, as denoted by FIA. This captures the general landform shape, topographical position, and soil moisture availability class of the plot. Some of the labels have been abbreviated: “Dry rt.” is “dry ridgetop”. These charts demonstrate that a wide spectrum of forest types and landform types have been covered.

Main model specification: We used a multiple logistic regression model to link the binomial categorical dependent variable and the continuous independent variables (i.e., the lidar metrics). We used the ‘least absolute shrinkage and selection operator’ technique (henceforth shortened as ‘LASSO’), a technique known for good and stable feature selection from correlated predictors (Tibshirani 1996; Kuhn and Johnson 2013). We used the shrinkage parameter (λ , in the LASSO literature) that resulted in the minimum cross-validated classification error rate. As a limited number of plots ($n=153$) were only available, independent training and test sets were not possible. Hence, we used all 153 plots to train the main model. Model accuracy was assessed by cross validation: we used repeated 10-fold cross-validation (100 randomized repeats) to assess the prediction accuracy. We chose to use $K=10$ folds, considering the bias-variance trade-off associated with that choice. Using either 5 or 10 folds have been empirically shown to generate test error rate estimates that suffer neither from excessively high bias nor from very high variance (Hastie et al. 2013). Another advantage of using parametric models such as logistic regression is that the absolute value of the coefficients of the regression equation is a straightforward indication of how large the effect of the variable is. Hence, we also calculated the relative effect size of each variable, in the final model.

Table 1: Metrics used as predictor variables for shrub presence/absence, in our modeling effort. All returns of each pulse (ie, single, first, last, etc) have been used for computing these metrics. This was done because a good proportion of the plots used had data acquired with single return systems. Hence, for example, calculation of canopy height metrics using first returns (as is commonly done in literature) was not possible.

Metric Name	Description
<u>Vegetation metrics (from lidar)</u>	
$h5_{canPts}$, $h50_{canPts}$, $h85_{canPts}$, $h95_{canPts}$	Representative height percentiles of canopy returns (canopy returns defined as returns of height > 3.0 meters from the ground) (Erdody and Moskal 2010).
$stddev_{canPts}$, CV_{canPts} , $skew_{canPts}$, $kurt_{canPts}$	The standard deviation, coefficient of variation, skewness and kurtosis of canopy return heights (García et al. 2010).
$p_{groundRets}$	Percent of returns classified as ground returns
p_{hb1} , p_{hb2}	Percentage of returns (both ground and non-ground) in two understory-related height bins. Height bin 1 is between 0.5 and 2.0 meters, height bin 2 is between 0.5 and 3.0 meters.
$ULCD_1$, $ULCD_2$	Two flavours of the <i>ULCD</i> metric, adapted from (Wing et al. 2015). $ULCD_1$ is defined in the following way: $ULCD_1 = \frac{UP}{RGP}$; where <i>UP</i> (understory points) is the number of returns in the understory (between 0.5 and 2.0 meters) and <i>RGP</i> (relative ground point) is the number of returns between 0.0 and 0.5 meters (effectively considered to be ground points). For $ULCD_2$, <i>UP</i> is the number of points in a slightly larger understory bin (0.5 to 3.0 meters).
$h10_{usPts}$, $h50_{usPts}$, $h70_{usPts}$, $h90_{usPts}$	Height percentiles of understory returns (defined as non-ground returns of height less than 3.0 m)
LPI_{3m}	A version of the Laser Penetration Index (LPI). For more details, see text.
$pGridCells_{LPI>5\%}$	A modified version of LPI, see text for more details.
$stdev_{h90_3m}$, CV_{h90_3m}	Metrics for quantifying the patchiness of the understory; see text for more details.
<u>Topographic metrics (calculated from the Lidar-generated DEM)</u>	
$range_{DEMhts}$, mad_{DEMhts} , sd_{DEMhts} , $skew_{DEMhts}$, $kurt_{DEMhts}$, IQR_{DEMhts}	Metrics for quantifying the variation in the terrain; see text for more details.

*var*_{DEMhts}

Slope Slope on the 100x100m lidar square-plot (Martinuzzi et al. 2009)

SCOSA Percent slope * cos(aspect) (Stage 1976)

SSINA Percent slope * sin(aspect) (Stage 1976)

FIA plot-level metrics:

p_{sw} Percentage (by basal area) of softwood trees, in the four FIA subplots. For more details on this estimate, see (Gopalakrishnan et al. 2015)

p_{pine} Percentage (by basal area) of pine trees, in the four FIA subplots. The calculations are the same as that for *p_{sw}* (see above), but basal area is partitioned into pine and non-pine classes, instead.

p_{natFor} Percentage of the FIA plot estimated to be natural forests (versus plantations). This is calculated from the condition-level data of the FIA plot (Bechtold and Patterson 2005)

Auxiliary variables:

YrAq_{lidar} Year of lidar data acquisition; to capture progression in sensor technology.

gap_{years} Time gap (number of years) between lidar data collection and lidar acquisition. This variable can take values of either 0, 1 or 2.

p_{For}_{lidar} Percent of 100x100m plot that is forested, as estimated from lidar (see text for more details)

numRets Total number of returns per pulse. For example, for a single-return system, *numRets* = 1.

EPC Quantifies the proportion of the plot where the understory has been adequately sampled by lidar returns. For computing this, the lidar square-plot is gridded into smaller 1x1m gridcells; EPC is the proportion of such gridcells that have least one understory return. For more details, see text.

3. Results:

Model tuning and feature selection was done using the LASSO method, as described before. The performance of the final logistic regression model selected is fair (accuracy = 62%, kappa = 0.23). The model is better than a null-model (*p*-value for McNemar's test on kappa is < 0.0001), demonstrating the utility of lidar data.

The confusion matrix between the classes can be seen in Figure 5(a). The variables (predictor lidar metrics) selected in the final model are shown in Table 2, along with relative effect size. The evolution of model classification accuracies for increasing average percent shrub cover (estimated from the FIA plot data) can be seen in figure 5(b). The correlation matrix for these nine selected variables can be seen in figure 6.

We see that several lidar metrics related to ratios of plot-level height bins ($ULCD_1$, $pGridCells_{LPI>5\%}$, p_{hb2}) have been chosen. For understanding the efficacy of these metrics, consider the lidar metric p_{hb2} . This metric is a straightforward quantization of the number of returns in an understory bin (between 0.5 and 3.0 meters) compared to the total number of returns from the plot. This is illustrated in figure 6.

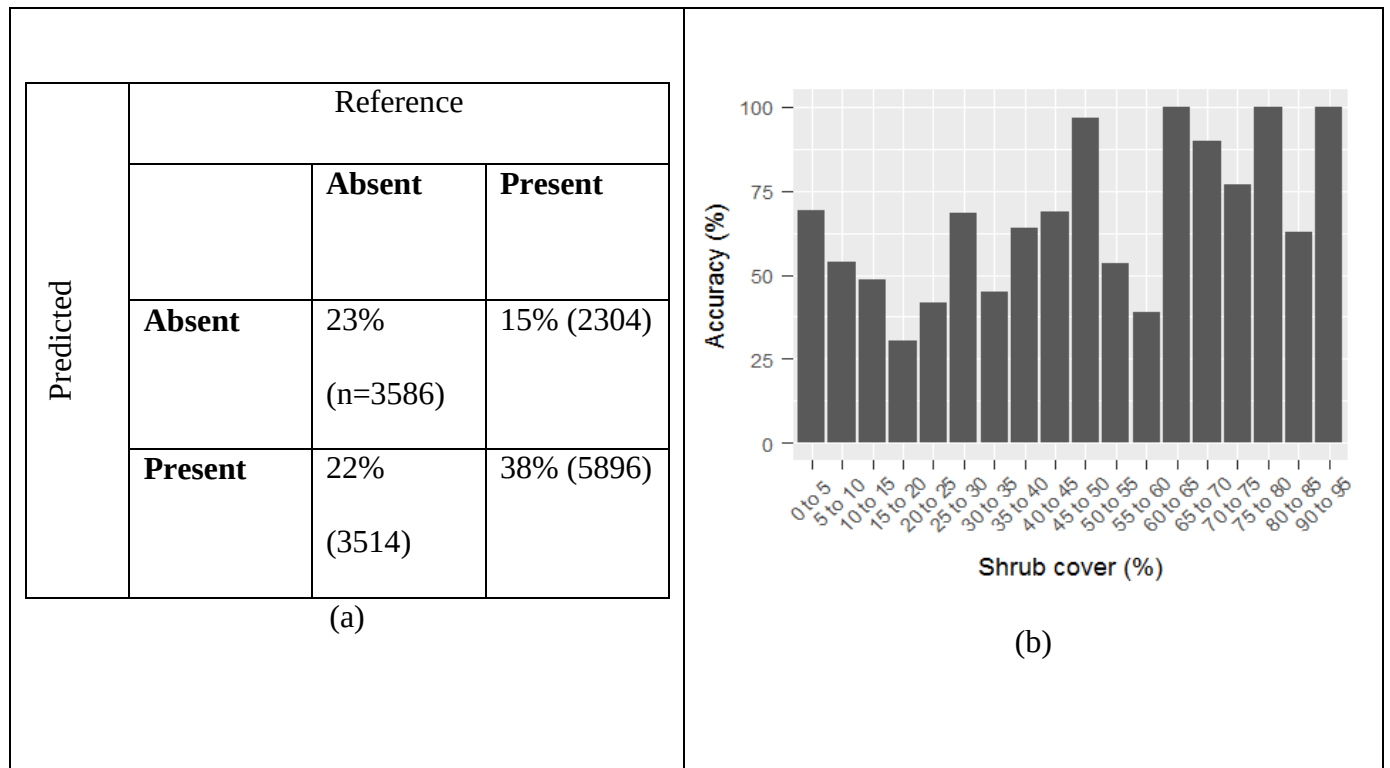


Figure 5 (a): Confusion matrix of the classification, expressed as percent of total test plots used during the repeated 10-fold cross validation ($n=15300$); (b) Trends in model classification accuracy for varying percent shrub cover. Classification results of 15300 test plots generated during the 10-fold cross validation is depicted. Classification accuracies are generally better at the ends of the distribution (that is, when shrub cover is less than 5% or more than 60%).

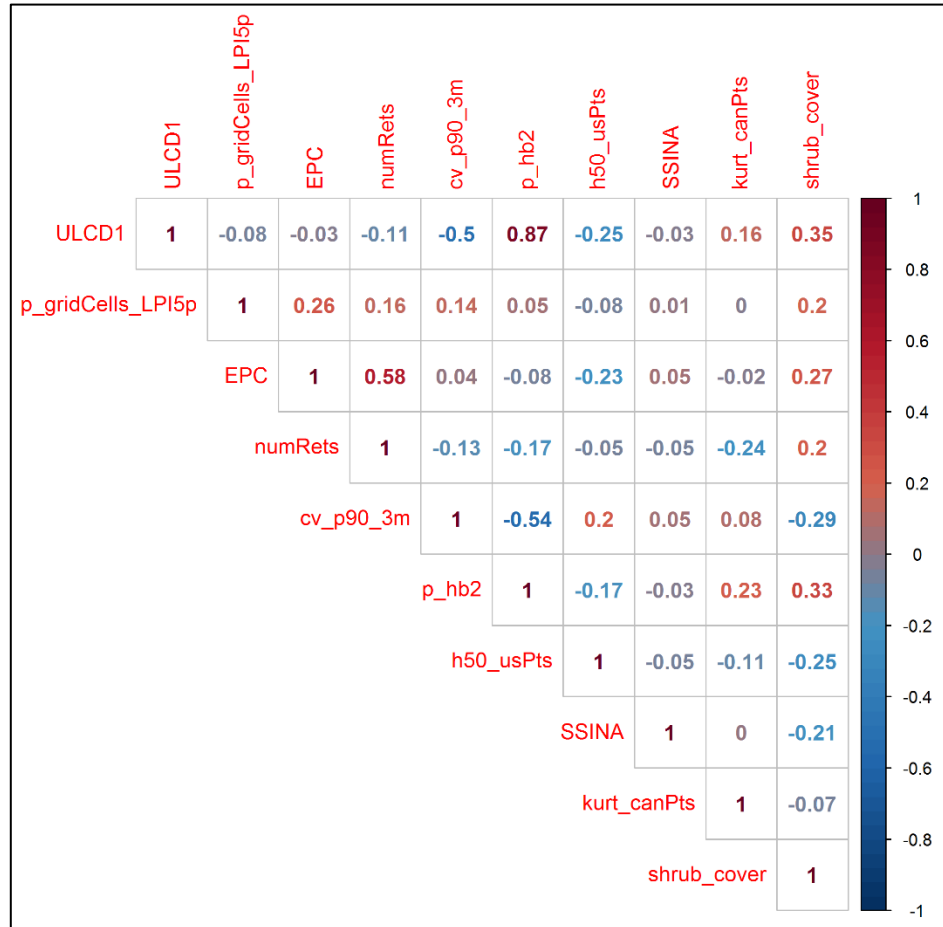


Figure 6: Correlation matrix of the selected variables, along with the dependent variable (continuous variant). Some variables have been shortened for better readability. For example, $pGridCells_{LPI>5\%}$ is shown as $p_gridCells_LPI5p$. Here, $shrub_cover$ is the percent shrub cover (continuous variable). There is high correlation between $ULCD_1$ and p_{hb2} . This is expected, as they quantify the percent of returns from the understory bin, but in different ways. The LASSO technique may have selected both of them as each has unique power in explaining the dependent variable in certain situations.

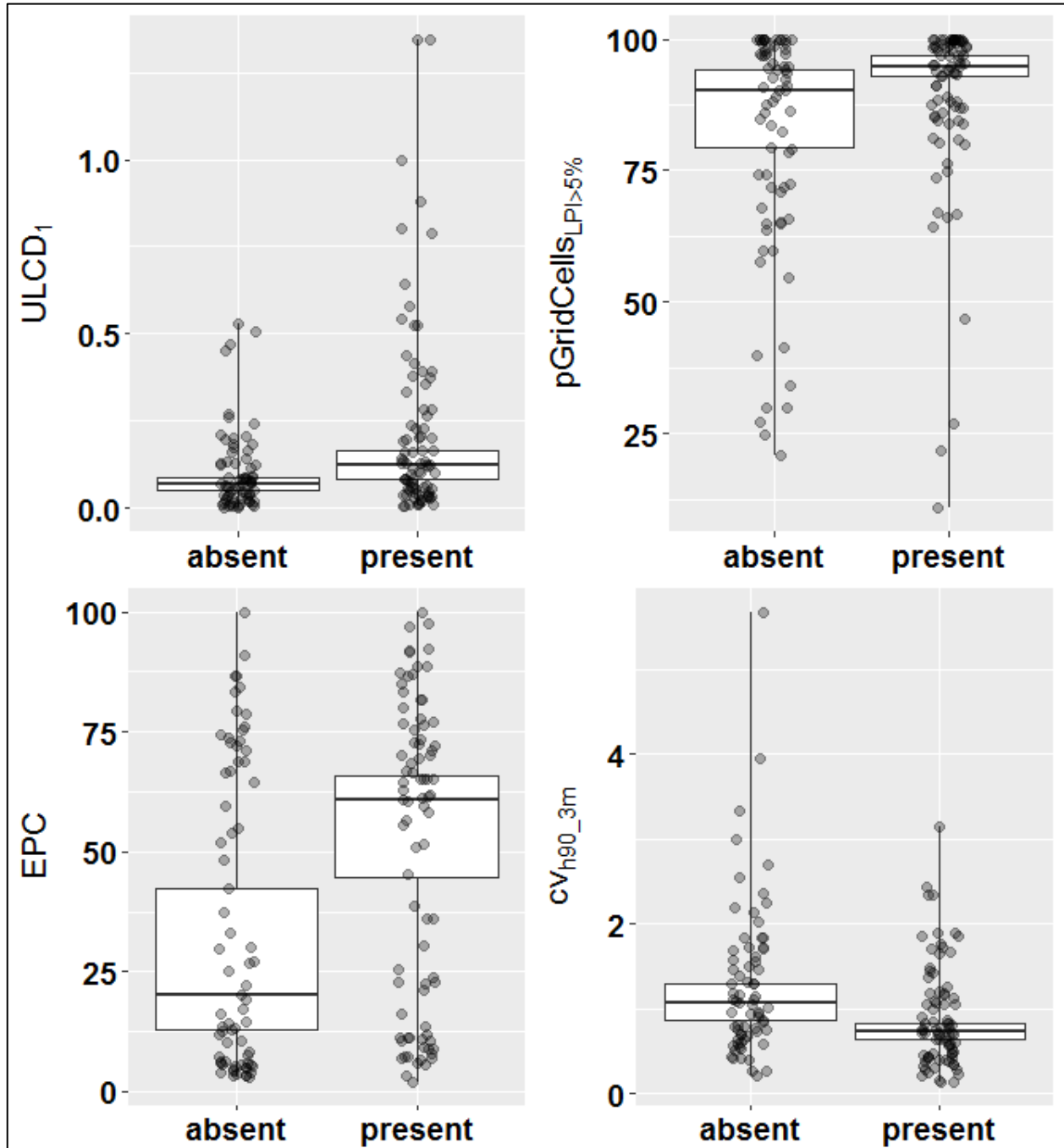


Figure 7: Box-plots depicting the distribution of the four most important variables selected for understory condition discrimination (absent versus present). The median of the population is non-parametrically estimated, along with the uncertainty, using 2000 bootstrap samples generated from the original sample. The top and bottom vertical bounds of the box here represent the upper and lower 95th percentile confidence intervals for the median. It is useful to keep the physical significance of these variables in mind while interpreting these graphs: $ULCD_1$ is a proxy for the amount of plant material in the understory layer, relative to the ground. The metric $pGridCells_{LPI>5\%}$ is a light penetration metric, of how open the canopy is. The metric EPC is an indicator of how well the understory part of the forest plot is sampled by lidar returns, which is another indicator of canopy openness. Lastly, CV_{h90_3m} indicates how patchy the understory layer

is (higher values indicate more patchiness). Only relatively small differences in variable distributions are seen between understory conditions, explaining the poor model performance.

Our main model confirms the utility and efficacy of several lidar metrics, some of which have been used in similar studies. To understand this further, the distributions of the top four important variables are shown in figure 7 (the lidar-acquisition related auxiliary variable *numRets* was excluded). The metric $ULCD_1$ has high importance in understory detection: an intuition of how this metric works can be got from figure 8. A height-bin based metric very similar to $ULCD_1$ was found to explain the greatest amount of variability in understory vegetation cover for Ponderosa Pine forests (Wing et al. 2015). From figure 7, it can be seen that values of $ULCD_1$ are higher when shrubs are present, as there would be more returns from the understory layer in this case. Also, it is noteworthy that the metric $pGridCells_{LPI>5\%}$ was found to be quite important. Stands with higher light penetration tend to have more understory (as seen in figure 7) because light is usually one of the most limiting factors affecting understory plant mortality and growth (Kimmins 2004). Overstory characteristics such as LAI and canopy gaps affect light penetration and the amount of light that ultimately reaches the understory. The fact that the horizontal metric $pGridCells_{LPI>5\%}$ was chosen over LPI_{3m} shows the efficacy of such horizontal metrics in better characterizing the uniformity of light penetration over the entire plot. EPC is another horizontal metric that is also correlated with the prevalence of canopy gaps, hence we see that higher EPC values mostly implies more understory. Another interesting finding is the importance of topography, as SSINA was selected. The negative value of the associated coefficient suggests that western aspects are slightly more suitable for the development of understory shrubs in our region of study. This may be due to the fact that western aspects in this region are known to be drier, thus leading to sparser crowns and overstory vegetation (see Beers et al. (1966) (Beers, Dress, and Wensel 1966), where it was reported that southwest aspects are most limiting for forest growth). Hence, more sunlight may reach the understory which, in turn, promotes more shrub growth. In a similar study, a topography-based slope-aspect transformation very similar to SSINA was one of the three variables selected for the main model (Martinuzzi et al. 2009).

Table 2: Important predictor variables identified for modeling understory, along with short descriptions (see table 1 and text for more details). As standardized predictor variable were used for model development, the absolute value of the coefficient associated with a predictor indicates the relative importance of the predictor, in terms of effect size. Hence, the % effect size denotes the relative contribution of the absolute value of that predictor.

Variable	Description	Type	Coefficient value (% effect size)
$ULCD_1$	Ratio of number of returns from understory height bin to those from ground bin	Understory	0.247 (21.9)
$pGridCells_{LPI>5\%}$	Quantification of light penetration through the canopy	Canopy	0.171 (15.2)
EPC	Effective plot coverage. Proportion of the plot where the understory has been adequately sampled by lidar returns.	Canopy+ auxiliary	0.167 (14.9)

<i>numRets</i>	Maximum number of returns registered from the plot	Auxiliary	0.158 (14.0)
<i>CV_{h90_3m}</i>	Coefficient of variation of understory heights over the plot	Understory	-0.124 (11.0)
<i>p_{hb2}</i>	Percent of points in height bin (0.5 to 3.0 meters)	Understory	0.093 (8.3)
<i>h50_{usPts}</i>	Plot-level height of understory	Understory	-0.091 (8.1)
SSINA	Percent slope * sin(aspect). Quantifies plot aspect	Topography	-0.069 (6.1)
<i>kurt_{canPts}</i>	Kurtosis of canopy return heights	Canopy	-0.005 (0.4)

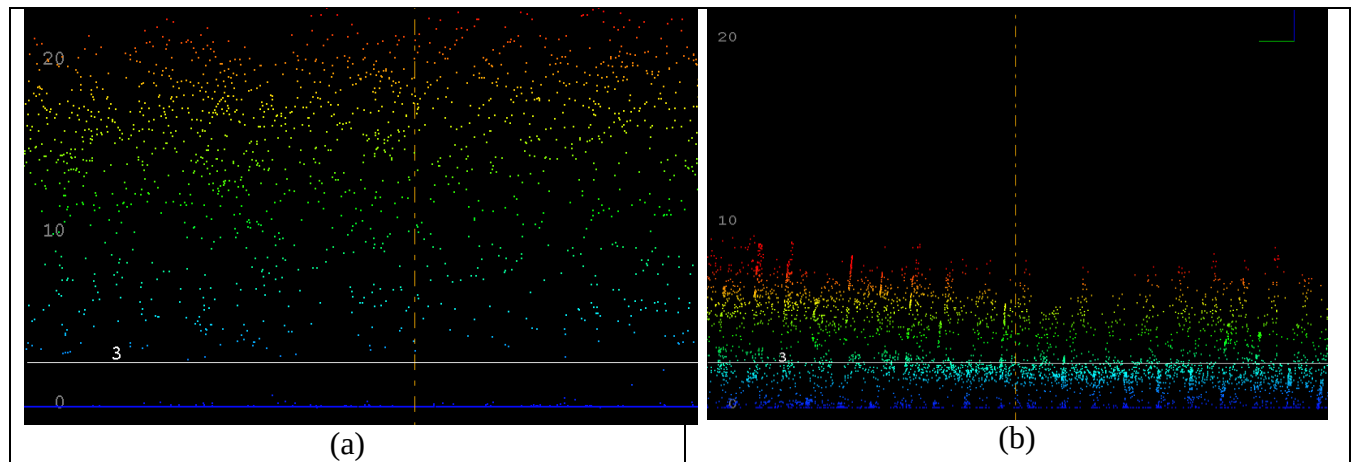


Figure 8: Example lidar point cloud profiles of forest plots which illustrates the efficacy of the the p_{hb2} lidar metric in distinguishing between understory absent and present conditions. (a) An unmanaged hardwood stand with understory practically absent. Here (and in (b) too), grey line with the annotation “3” marks the 3.0m height level. The light grey numbers on the left side mark the height, in meters. The p_{hb2} value computed is low (0.2%) (b) A planted pine stand with high understory presence. In this case, p_{hb2} is on the higher side (44%). Hence, one can observe the greater number of returns in the understory height bin, compared to (a).

4. Discussion:

Understory vegetation cover has been difficult to estimate and predict, especially over large spatial extents (Wing et al. 2015). Operational efforts often use products that may be substandard in many places and regions: for example, for LANDFIRE products, shrub cover is inferred from the remote-sensed dominant vegetation cover (i.e., tree canopy cover type in forests) using a rule-based method (Ottmar et al. 2007). Forest shrub layer maps are important to conservation managers too: for example, in Mediterranean forests, more than 40% of the total floristic richness, as quantified by forest plants, is located in the shrub layer. In this context, efforts such as ours to better understand the utility of lidar data in extending FIA field data are important. We have developed a simple and robust inferential model that demonstrates the important underlying relationships and helps explore the associations between lidar data and understory presence. The

model performance, as such, was fair. Some reasons for the reduced model performance are illustrated in figure 9. The importance of point density can be noted from figure 9(a). The median point density for our dataset is 1.2 returns/m², and plots that have point density significantly lower than this would not be represented well by the model. Another consequence of such low point density is that the understory sampling by the lidar is quite limited. Problems of heterogeneity are illustrated in 9(c) and 9(d). In 9(c), it can be seen that the forest is quite patchy, with the ground clearly visible (blue colour) in several large patches. This greatly increases the chance of the FIA microplots being non-representative for the entire FIA plot. A similar problem of heterogeneity can be seen in 9(d), where possible different management regimes are implicated. Also, work by Korpela et al. suggests that return echo-triggering probabilities may be dependent on the understory species involved, too (Korpela, Hovi, and Morsdorf 2012). A few suggestions for improvement, in the context of the overall methodology, are also given below:

1. Make lidar metadata uniformly available across all projects: Several sensor-level attributes are important to effectively detect understory, and differentiate it from nearby elements (such as ground or canopy). First of all, the accuracy of the instrument (as estimated from ground control points and over vegetated areas) should be high. Then, the scan angle is important: at high off-nadir scan angles, there is a much greater chance of both understory occlusion and of pulses hitting tree boles at the understory level. Another factor of importance in multi-return system is the *instrument dead-time*, which is the minimum vertical distance needed between two forest elements to be recorded as two separate returns. Such metadata could not be factored in into our analysis, as these were not readily available for most of the lidar acquisitions. Using such data is expected to give better explanatory power to the models.
2. Intensity values: The intensity values of lidar returns is a function of surface reflectance of the reflecting surface and hence is useful for differentiating between understory components (shrubs versus herbaceous vegetation, saplings, stumps, tree boles and coarse woody debris). Such values recorded by the more recent sensors have been shown to be very useful for understory studies. For example, return intensity values have been used to filter and remove points associated with unwanted understory components (Wing et al. 2015), for cluster analysis based vegetation strata separation (Morsdorf et al. 2010) and for separating snags from live vegetation (Martinuzzi et al. 2009). But, in spite of their demonstrated potential in the understory context, we could not use intensity values in our studies as they were not available or well-calibrated for most of our acquisitions.
3. Improvements in the FIA sampling scheme: Although the FIA P3 data is the best available database for large-scale regional understory measurement, it suffers from some major shortcomings. The most important was that the number of plots available for our analysis was rather limited (n=153), given the heterogeneity in landform physiographic conditions, forest types and forest succession stages involved. A major reason for this is the need to screen out plots with heterogeneous land-use, from the original set: in our case, as much as 44% of plots were thus dropped. As mentioned before, this is because the FIA does not move their plots so that they do not straddle multiple forest conditions. Having an augmented sampling intensity of P3 plots in “generally well-forested regions” (which should be identified from historic land-use data) would be a possible solution to this. Another reason is the inability to better co-register FIA and lidar data; this has been extensively discussed in the literature before (Gopalakrishnan et al. 2015).

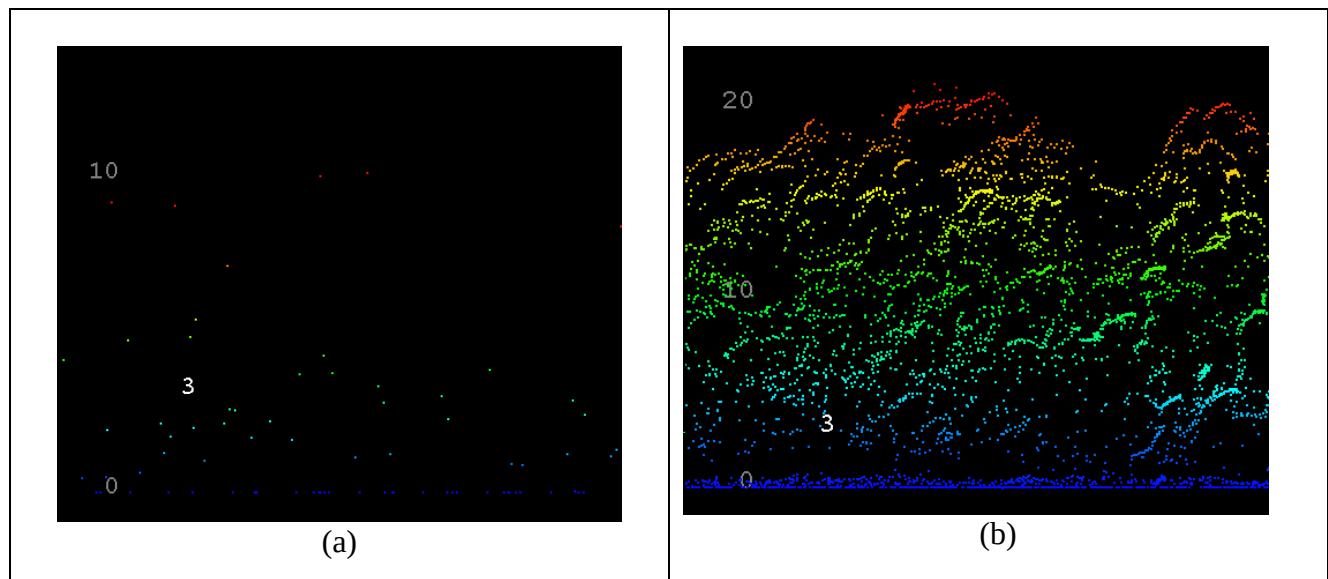
4. Improvements in FIA's shrub measurement protocols: A comprehensive internal study done by the FIA looked at the repeatability of understory measurements for 139 P3 plots, these measurements being taken over a span of 3 years (Westfall and Woodall 2007). Across the suite of measurement variables that comprise the forest fire fuel load inventory, it was found that as much as 15 of the 27 field variables did not have the required repeatability levels. Also, the microplots are rather small, given the heterogeneity of understory condition. A good empirical estimate of plot size for proper sampling of understory vegetation in temperate zones is 50 to 200 m² (this estimate varies between such wide limits because it depends on the type of vegetation community being sampled) (Ellenberg and Mueller-Dombois 1974). Meanwhile, the total area sampled by all four FIA microplots (53.8 m²) is almost at the lower limit, and is much smaller than the upper limit.

It should be noted that the fairly low overall accuracy that we report (62%) is similar to that of several studies that which used lidar to detect the occluded understory layer. For example, a similar effort at understory detection (shrub present/absent) reported a low accuracy of 48%, even after being aided by lidar intensity information and high-quality field data (Morsdorf et al. 2010). Another effort performed on terrain covered by dense, mixed-conifer forests also reported poor RMSEs for shrub cover (20 to 30%), even when high-quality lidar data was used (10 pulses/m²) (Jakubowski, Guo, and Kelly 2013). The authors go on to suggest that relatively higher density data would be ideal for all cover metrics (like tree cover & shrub cover). This is supported by results reported by Korpela et al. (2012) that the R^2 of lidar metric based models involved in prediction of understory stem density (stems ha⁻¹) increased with increased pulse density (Korpela, Hovi, and Morsdorf 2012).

To quantify and understand this further, we calculated the understory point density (UPD), or the density of returns (returns/m²) in the height bin of 0.0 to 3.0 meters. This quantifies the ability of the lidar instrument to penetrate the overstory and sample the vegetation beneath it. UPD is relatively low, with a mean of 0.82 and a median of 0.53 pts/m². Also, as much as 70% of the plots have UPD below 1.0 pts/m². It is known that some vegetation metrics are sensitive to low point density. For example, the estimation of forest variables such as volume and canopy height from lidar metrics is still possible with low point density (Gobakken and Næsset 2008) but for individual tree detection and delineation, much higher pulse densities (7 to 15 pts/m²) are recommended (Vauhkonen et al. 2011). As shrub cover estimation is similar to tree detection in some sense, we think that understory point density is important, and might explain our low accuracies. If so, more comparative studies like (Gobakken and Næsset 2008) and (Vauhkonen et al. 2011) are required to ascertain optimal understory point densities for shrub mapping in various forest types.

In a recent work on tropical forest biomass estimation using ALS lidar, Asner et al. (2012) suggested that there is a consistency in the way trees naturally fill tropical forest canopy space and are thus sensed by lidar (Asner et al. 2012). They then demonstrated this consistency, and exploited it to develop a simple and universal model for biomass estimation over a wide variety of tropical forest types. Our work is on similar lines: we derived a set of a lidar metrics as important for understory presence estimation over a wide range of temperate forest types and topographic conditions in the Southeastern US (refer to table 2). An examination of the effect

sizes in the main model highlights the relative importance of overstory and topographic factors. These effect sizes could be compared with similar parameters of broad-based forest ecology models of the region, and discrepancies should be further investigated. The choice of $pGridCells_{LPI>5\%}$ by the modeling framework as the second-most important determinant for understory presence points to two inferences: (1) Light penetration at the plot level through canopy gaps is important for understory growth; (2) Such canopy gaps need to be distributed across the plot (and not clustered) for better understory growth. Most forest floor regions tend to be highly sunlight-limited: the average light intensity found beneath a well-stocked forest stand is quite below the optimal level for most plant species (Kimmins 2004). For example, the sunlight intensity beneath an eastern white pine forest is approximately 27% of the full intensity, and it could be as low as 2-5% for other common deciduous forest types (Kimmins 2004). The importance of slope and aspect in the determination and regulation of sunlight intensities during the day is also well known. Aspects help decide how wet or dry a forested site is, and the effect of the aspect is further amplified by increasing slopes. For example, it had been observed that southwest aspects are usually the most severe sites for forest growth and regeneration (Beers, Dress, and Wensel 1966).



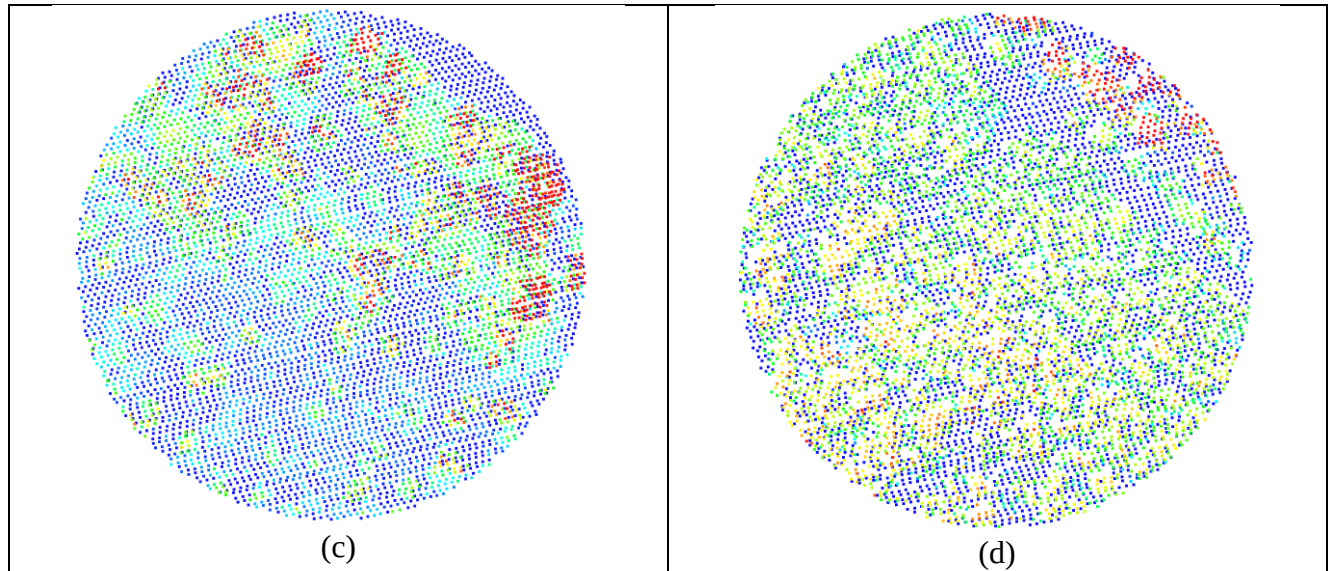


Figure 9: Prominent reasons for incorrect model class predictions are illustrated here. All figures show lidar point-clouds, with the colour gradation (blue to red) representing the height above ground (blue represents points near ground, red represents points near the top of the canopy). (a) Vertical profile of the plot point cloud. Here (and in (b) too), the annotation “3” marks the 3.0m height level from the ground. The misclassification as “shrub absent” is due to low point density (0.3 pts/m^2); (b) Vertical profile of point cloud. FIA data labels this plot as “shrub absent”; the most probable reason for misclassification is vegetation regrowth. This is a non-plantation forested plot where the lidar acquisition was done two after FIA measurements. It seems from the lidar profile that there is quite some understory returns which points to the possibility of understory vegetation regrowth since FIA measurement; (c) Planimetric top-view of plot (same for (d)). A very heterogeneous forest condition is seen; (d) A managed pine plantation. The straight lines (blue) correspond to likely forest roads, they divide the forest into stands of unequal height. Understory conditions can vary widely across such differing management regimes. In (c) and (d) the heterogeneity most likely causes the sampling microplots to be non-representative.

5. Conclusions:

We propose a set of a lidar metrics as correlated for shrub presence over several Southeastern forest types and topographic conditions. We also give a ranking of the importance of these predictor variables. Not surprisingly, the most important one is a straight-forward measure of the percentage of returns from the understory bin. This work also demonstrates the utility of the relatively new and untested ULCD and other metrics based on height bin ratios to better help predict understory. We were successful in demonstrating the efficacy of horizontal metrics in characterizing overstory and understory vegetation and terrain features. We also demonstrated the major importance of overstory characteristics in dictating understory conditions by noting the prevalence of overstory-related metrics in our models. The importance of topography and that it can be well-sensed by lidar is demonstrated by the importance of the aspect-based topographic metric (SSINA) to the model. Overall, this work highlights the promise in combining disparate lidar datasets with field data to develop understory presence maps.

Acknowledgements: The authors would like to acknowledge LISA, the statistical consulting group of the statistics department, Virginia Tech for their help during this work.

References:

- Andreu, A., and L. A. Hermansen-Baez. 2008. "Fire in the South 2: The Southern Wildfire Risk Assessment." <http://www.treeseearch.fs.fed.us/pubs/34087>.
- Asner, Gregory P., Joseph Mascaro, Helene C. Muller-Landau, Ghislain Vieilledent, Romuald Vaudry, Maminaiaina Rasamoelina, Jefferson S. Hall, and Michiel van Breugel. 2012. "A Universal Airborne LiDAR Approach for Tropical Forest Carbon Mapping." *Oecologia* 168 (4): 1147–1160.
- Bechtold, William A., and Paul L. Patterson. 2005. *The Enhanced Forest Inventory and Analysis Program: National Sampling Design and Estimation Procedures*. US Department of Agriculture Forest Service, Southern Research Station. http://www.srs.fs.usda.gov/pubs/gtr/gtr_srs080/gtr_srs080.
- Beers, Thomas W., Peter E. Dress, and Lee C. Wensel. 1966. "Notes and Observations: Aspect Transformation in Site Productivity Research." *Journal of Forestry* 64 (10): 691–692.
- Blinn, Christine E., Timothy J. Albaugh, Thomas R. Fox, Randolph H. Wynne, Jose L. Stape, Rafael A. Rubilar, and H. Lee Allen. 2012. "A Method for Estimating Deciduous Competition in Pine Stands Using Landsat." *Southern Journal of Applied Forestry* 36 (2): 71–78.
- Camprodon, Jordi, and Lluís Brotons. 2006. "Effects of Undergrowth Clearing on the Bird Communities of the Northwestern Mediterranean Coppice Holm Oak Forests." *Forest Ecology and Management* 221 (1): 72–82.
- Ellenberg, Dieter, and Dieter Mueller-Dombois. 1974. *Aims and Methods of Vegetation Ecology*. Wiley New York, NY. https://www.researchgate.net/profile/Dieter_Mueller-Dombois/publication/275652823_Aims_Methods_Veg_Ecology_ch5/links/5542f1770cf24107d3948c8d.pdf.
- Erdody, T. L., and L. M. Moskal. 2010. "Fusion of LiDAR and Imagery for Estimating Forest Canopy Fuels." *Remote Sensing of Environment* 114 (4): 725–737.
- García, Mariano, David Riaño, Emilio Chuvieco, and F. Mark Danson. 2010. "Estimating Biomass Carbon Stocks for a Mediterranean Forest in Central Spain Using LiDAR Height and Intensity Data." *Remote Sensing of Environment* 114 (4): 816–830.
- Gobakken, Terje, and Erik Næsset. 2008. "Assessing Effects of Laser Point Density, Ground Sampling Intensity, and Field Sample Plot Size on Biophysical Stand Properties Derived from Airborne Laser Scanner Data." *Canadian Journal of Forest Research* 38 (5): 1095–1109. doi:10.1139/X07-219.
- Gopalakrishnan, Ranjith, Valerie A. Thomas, John W. Coulston, and Randolph H. Wynne. 2015. "Prediction of Canopy Heights over a Large Region Using Heterogeneous Lidar Datasets: Efficacy and Challenges." *Remote Sensing* 7 (9): 11036–11060.
- Greaves, Heather E., Lee A. Vierling, Jan UH Eitel, Natalie T. Boelman, Troy S. Magney, Case M. Prager, and Kevin L. Griffin. 2015. "Estimating Aboveground Biomass and Leaf Area of Low-Stature Arctic Shrubs with Terrestrial LiDAR." *Remote Sensing of Environment* 164: 26–35.
- Hastie, Trevor, Gareth James, Robert Tibshirani, and Daniela Witten. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer, New York,.

- Jakubowski, Marek K., Qinghua Guo, and Maggi Kelly. 2013. "Tradeoffs between Lidar Pulse Density and Forest Measurement Accuracy." *Remote Sensing of Environment* 130: 245–253.
- Kimmins, J. P. 2004. *Forest Ecology; A Foundation for Sustainable Forest Management and Environmental Ethics in Forestry. Third Edition*. Prentice Hall, New Jersey.
- Korpela, Ilkka, Aarne Hovi, and Felix Morsdorf. 2012. "Understory Trees in Airborne LiDAR data—Selective Mapping due to Transmission Losses and Echo-Triggering Mechanisms." *Remote Sensing of Environment* 119: 92–104.
- Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer.
<http://link.springer.com/content/pdf/10.1007/978-1-4614-6849-3.pdf>.
- Lone, Karen, Leif Egil Loe, Terje Gobakken, John DC Linnell, John Odden, Jørgen Remmen, and Atle Mysterud. 2014. "Living and Dying in a Multi-Predator Landscape of Fear: Roe Deer Are Squeezed by Contrasting Pattern of Predation Risk Imposed by Lynx and Humans." *Oikos* 123 (6): 641–651.
- Martinuzzi, Sebastián, Lee A. Vierling, William A. Gould, Michael J. Falkowski, Jeffrey S. Evans, Andrew T. Hudak, and Kerri T. Vierling. 2009. "Mapping Snags and Understory Shrubs for a LiDAR-Based Assessment of Wildlife Habitat Suitability." *Remote Sensing of Environment* 113 (12): 2533–2546.
- McRoberts, Ronald E. 2010. "The Effects of Rectification and Global Positioning System Errors on Satellite Image-Based Estimates of Forest Area." *Remote Sensing of Environment* 114 (8): 1710–1717.
- Morsdorf, Felix, Anders Måre, Benjamin Koetz, Nathalie Cassagne, Francois Pimont, Eric Rigolot, and Britta Allgöwer. 2010. "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* 114 (7): 1403–1415.
- Nijland, Wiebe, Scott E. Nielsen, Nicholas C. Coops, Michael A. Wulder, and Gordon B. Stenhouse. 2014. "Fine-Spatial Scale Predictions of Understory Species Using Climate- and LiDAR-Derived Terrain and Canopy Metrics." *Journal of Applied Remote Sensing* 8 (1): 083572–083572.
- Oswalt, Sonja N., Michael T. Thompson, and W. Brad Smith. 2009. *US Forest Resource Facts and Historical Trends*. US Department of Agriculture, Forest Service.
- Ottmar, Roger D., David V. Sandberg, Cynthia L. Riccardi, and Susan J. Prichard. 2007. "An Overview of the Fuel Characteristic Classification System—Quantifying, Classifying, and Creating Fuelbeds for Resource Planning This Article Is One of a Selection of Papers Published in the Special Forum on the Fuel Characteristic Classification System." *Canadian Journal of Forest Research* 37 (12): 2383–2393.
- Peduzzi, Alicia, H. Lee Allen, and Randolph H. Wynne. 2010. "Leaf Area of Overstory and Understory in Pine Plantations in the Flatwoods." *Southern Journal of Applied Forestry* 34 (4): 154–160.
- Peduzzi, Alicia, Randolph H. Wynne, Valerie A. Thomas, Ross F. Nelson, James J. Reis, and Mark Sanford. 2012. "Combined Use of Airborne Lidar and DBInSAR Data to Estimate LAI in Temperate Mixed Forests." *Remote Sensing* 4 (6): 1758–1780.
- Rollins, Matthew G., Robert E. Keane, and Russell A. Parsons. 2004. "Mapping Fuels and Fire Regimes Using Remote Sensing, Ecosystem Simulation, and Gradient Modeling." *Ecological Applications* 14 (1): 75–95.

- Simonson, William D., Harriet D. Allen, and David A. Coomes. 2012. "Use of an Airborne Lidar System to Model Plant Species Composition and Diversity of Mediterranean Oak Forests." *Conservation Biology* 26 (5): 840–850.
- Singh, Kunwar K., Amy J. Davis, and Ross K. Meentemeyer. 2015. "Detecting Understory Plant Invasion in Urban Forests Using LiDAR." *International Journal of Applied Earth Observation and Geoinformation* 38: 267–279.
- Stage, Albert R. 1976. "Notes: An Expression for the Effect of Aspect, Slope, and Habitat Type on Tree Growth." *Forest Science* 22 (4): 457–460.
- Stewart, Susan, Volker Radeloff, Roger Hammer, Jeremy Fried, Sherry Holcomb, and Jason McKeefry. 2005. "Mapping the Wildland Urban Interface and Projecting Its Growth to 2030: Summary Statistics." *USDA Forest Service North Central Research Station: Evanston, IL, USA*.
- Sumnall, Matthew J., Thomas R. Fox, Randolph H. Wynne, Christine Blinn, and Valerie A. Thomas. 2016. "Estimating Leaf Area Index at Multiple Heights within the Understorey Component of Loblolly Pine Forests from Airborne Discrete-Return Lidar." *International Journal of Remote Sensing* 37 (1): 78–99.
- Sverdrup-Thygeson, Anne, Hans Ole Ørka, Terje Gobakken, and Erik Næsset. 2016. "Can Airborne Laser Scanning Assist in Mapping and Monitoring Natural Forests?" *Forest Ecology and Management* 369: 116–125.
- Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Vauhkonen, Jari, Liviu Ene, Sandeep Gupta, Johannes Heinzl, Johan Holmgren, Juho Pitkänen, Svein Solberg, et al. 2011. "Comparative Testing of Single-Tree Detection Algorithms under Different Types of Forest." *Forestry*, cpr051.
- Vogelmann, James E., Stephen M. Howard, Limin Yang, Charles R. Larson, Bruce K. Wylie, and Nick Van Driel. 2001. "Completion of the 1990s National Land Cover Data Set for the Conterminous United States from Landsat Thematic Mapper Data and Ancillary Data Sources." *Photogrammetric Engineering and Remote Sensing* 67 (6).
<https://trid.trb.org/view.aspx?id=686441>.
- Wear, DN, and JG Greis. 2011. *The Southern Forest Futures Project: Summary Report*.
<http://www.srs.fs.usda.gov/futures/>.
- Westerling, Anthony L., Hugo G. Hidalgo, Daniel R. Cayan, and Thomas W. Swetnam. 2006. "Warming and Earlier Spring Increase Western US Forest Wildfire Activity." *Science* 313 (5789): 940–943.
- Westfall, J. A., and C. W. Woodall. 2007. "Measurement Repeatability of a Large-Scale Inventory of Forest Fuels." *Forest Ecology and Management* 253 (1): 171–176.
- White, Joanne C., Michael A. Wulder, Andrés Varhola, Mikko Vastaranta, Nicholas C. Coops, Bruce D. Cook, Doug Pitt, and Murray Woods. 2013. "A Best Practices Guide for Generating Forest Inventory Attributes from Airborne Laser Scanning Data Using an Area-Based Approach." *The Forestry Chronicle* 89 (6): 722–723.
- Wing, Brian M., Martin W. Ritchie, Kevin Boston, Warren B. Cohen, and Michael J. Olsen. 2015. "Individual Snag Detection Using Neighborhood Attribute Filtered Airborne Lidar Data." *Remote Sensing of Environment* 163: 165–179.
- Zellweger, Florian, Felix Morsdorf, Ross S. Purves, Veronika Braunisch, and Kurt Bollmann. 2014. "Improved Methods for Measuring Forest Landscape Structure: LiDAR

Complements Field-Based Habitat Assessment.” *Biodiversity and Conservation* 23 (2): 289–307.

Chapter 4

Predicting forest site index for managed pine in Southeastern USA using disparate lidar datasets and Landsat time-series maps

Abstract

We present a method for combining discrete return airborne laser scanning data from disparate acquisitions with Landsat disturbance mapping based proxy-age products to generate site index maps for plantation pine forests over large areas in the southeast United states. For predicting canopy heights, a linear regression model was developed using field data from the Forest Inventory and Analysis (FIA) program of the US forest service ($n=217$ plots). The model was parsimonious (the 70th percentile height of canopy returns was the single predictor variable) and robust ($R^2=0.84$, RMSE=1.85 meters). For forest stand ages, a modified version of the vegetation change tracker (VCT) algorithm was used. Planted pine areas were predicted by using a Landsat multitemporal-data based classification algorithm. The estimated bias of the method was low (-0.75 meters) and the RMSE was consistent with other similar approaches (3.8 meters). The relative RMSE estimate was 19%. We generate a 30 meter gridcell resolution current site index map (as of approximately 2009) for planted pine stands for parts of two distinct and geographically separated regions: Central South Carolina and Eastern Virginia (generated map area: 832 sq. km). The species was assumed to be loblolly pine (*Pinus taeda* L.) as it was dominant in this area. Using this map, we estimate that approximately 83% of the forest stand area surveyed had low levels of productivity (which we define as site index <22.0 meters for base age of 25 years). Further, using a hydrological unit (watershed) level analysis, we rank regions in the Southeast by their predicted levels of planted pine productivity. These results have important implications for policy makers and forest managers and would help guide better regional-level forest management interventions.

4.1 Introduction

Forests in the southeastern part of the United States store a substantial amount of carbon (around 12 Gtons), which is estimated to be as much as 36% of all forest carbon sequestered by forests in the conterminous United States [1]. They also absorb a large amount (around 13%) of regional greenhouse gas emissions [2]. However, some prominent model outputs suggest that the magnitude these carbon sinks could be highly diminished in the future in the face of climate change [3]. Hence, there is a need to better understand both current and projected forest site productivity at local and regional scales, which strongly motivates the need for site productivity maps that are robust and can be periodically updated. Here, site productivity is defined as the potential of a site to produce biomass (for a good review of this and related concepts, refer to [4]). Such maps based on remote sensing data are also useful for calibrating models related to land-use and land surface processes [5], [6]. Finally, robust forest productivity maps can provide salient inputs for *precision forestry* managers, in which spatially specific management interventions (such as fertilization, herbaceous weed control and density management) are carried out on forest stand locations of relatively low productivity [7].

Site index of a forest site is defined as the average height that dominant and co-dominant trees are projected to reach for a given base age (usually 25 or 50 years) at that site. Site classification by such stand height (that is, the use of site index curves) has become a universal practice in forestry and is recognized as one of the most suitable indicators of site productivity in even-aged forest stands [4]. A sizable proportion (around 30%) of southeastern US forests are dominated by the genus *pinus* ("pine"), with around 10 million hectares in naturally regenerated stands and another 14 million hectares of pine plantations [8], [9]. The spatial distribution of forest productivity and site index of this region has been studied over the past three decades using two distinct approaches: the theoretical model approach and the sampling plot based approach. The theoretical model approach consists of developing spatially explicit models of forest site index for pine forests of this region. Parametric and nonparametric empirical models that relate site index of plantation loblolly pine to biophysical variables (climate, edaphic, and physiographic factors) on a large scale have been developed [10]. A version of the dynamic, process-based 3PG model customized for loblolly pine has been developed and has been validated over a range of sites in Georgia, USA [11]. But in these efforts, the predicted quantity is *potential productivity* or the productivity under the assumed plantation coverage area and management regimes. A different value of productivity might actually manifest on the landscape, known as *expressed site productivity*. This is pine productivity which results after the manifestation of the effects of silvicultural treatments [12].

In the sampling plot based approach, one can analyze the variation of expressed loblolly pine productivity over a wide area in the Southeast, using plot-level field-measured data. The analysis of seven individual studies that were established across the natural range of loblolly pine to examine factors that control and limit the productivity of managed stands has been reported [13]. Another region-wide study of thirteen loblolly pine plantations spanning seven southern states and four physiographic provinces (Lower Coastal Plain, Middle Coastal Plain, Hilly Coastal Plain and Piedmont) primarily looked at the effects of controlling competing vegetation [14]. A factorial combination experimental design of woody and herbaceous control treatments were replicated at all of the sites. Changes in expressed site productivity over the geographic range of loblolly pine are reported in Zhao et al. (2016), an extensive analysis involving six long-term studies and over 250 plots [8]. While there have been quite a few studies that leveraged plot-level data, the potential of remote sensing techniques (with its concomitant advantages of large-area, standardized and synoptic coverage) for site index prediction for this region has not been studied. The generation of a site index map for a relatively large area by such techniques will help answer questions such as the proportion of forest stands with low productivity in that area, and whether there is (generally) significant variation of productivity at the intra-stand level.

There has been prior work regarding prediction and verification of site index from lidar in other regions. The first article that described explicit lidar-based site index mapping was by Gatziolis (2007) for a 9500-ha

temperate rainforest site in Oregon, USA [15]. Wulder et al. (2010) combined lidar-measured tree heights with inventory age information to predict site index for a 2500-ha forested area in British Columbia, Canada [16]. Tompalski et al. (2015) detailed a more involved study in the same region, where both lidar and Landsat time-series data were used together to predict site index over 100 square km of western hemlock dominated forests (RMSE = 5.55m) [17]. However, none of these previous studies address the unique challenges a similar effort faces in Southeast USA, namely: a) The lack of a single, homogeneous large-area lidar acquisition effort, thus necessitating the need to combine disparate lidar projects; b) The lack of standardized planting records for deriving stand age (such records were used by some efforts mentioned above [15], [16]; 3) A mixed landscape of both softwood and hardwood species types.

In this study, we focus on even-aged, planted pine forests, which could be managed at various levels of intensity. Recent work relevant to the unique challenges of the southeast was described by Gopalakrishnan et al. (2015) [18], where it was shown that the area-based approach is effective at producing robust models of dominant canopy height over as much as 76 disparate lidar projects ($R^2 = 0.74$). Further, it was shown there that canopy heights are better predicted for softwoods and over more homogeneous forested areas. We leverage on those two key insights in the design of this present study. Moreover, it has been shown that forest ages can be predicted reasonably for a large area in the Southeast USA using Landsat time series data [19] and that such time series data can be combined with spectral data to identify planted pines [20], [21]. Repeated lidar data collection is expected to happen for this area in the future; the United States Geological Survey (USGS) is coordinating an inter-agency effort aiming to periodically collect lidar for the conterminous United States [22], thus adding to the relevance of our work.

The objective of the present study is to explore the possibility of combining the datasets mentioned above towards generating reasonably robust expressed site index maps over the Southeast USA region. We try to answer the following specific research questions:

1. How effective is lidar collected for non-forestry purposes coupled with Landsat-based disturbance products (and using time since disturbance as proxy for age) in predicting site index over pine plantations? And, if reasonably effective:
2. What percentage of the study area has poor productivity (site index < 22.0 meters, or approximately 72 feet, at base age 25) and medium to high productivity (site index $\geq$ 22.0 meters)?
3. Are there detectable differences between potential and expressed pine productivity, over the landscape? If so, which regions register the greatest differences?

Hence, we aim to evaluate lidar data collected over large areas (mainly to aid in the generation of digital elevation models) as another tool that can be used to understand and analyze local and regional trends in southeastern forest productivity.

4.2 Methods

4.2.1 Study area

Our study area focused on all even-aged pine plantation forests in the area of lidar coverage (see figure 4.1). The selection criteria for the lidar projects used is explained below.

4.2.2 Remote sensing datasets

The following three datasets were used in our study:

1. Lidar data: We considered many lidar projects that were (a) available in the public domain as of 2011, (b) from the southeastern states of USA, (b) had recorded multiple returns (i.e., two or more) for each pulse, (c) there was at least one planted pine FIA plot in their area of coverage. This resulted in 38 projects being selected, spread across eight southeastern states (Virginia, North Carolina, South Carolina, Georgia, Florida, Alabama, Mississippi and Texas), as shown in figure 4.1. We used georeferenced point-cloud data from these projects. Their acquisition dates ranged from 2005 to 2012, and 31 of these projects used instruments capable of recording up to four returns.
2. Landsat-based prediction of plantation age: The vegetation change tracker (VCT) algorithm is a fairly well-established and extensively tested technique that effectively leverages multitemporal Landsat data to detect disturbances to forest stands (such as clearcutting) [23]. Recently, additional steps that partitioned such disturbances into stand-clearing ones (i.e., clearcuts) and non-stand-clearing ones (such as thinning, insect damage and ice damage) were proposed as part of an enhanced vegetation change tracker algorithm (henceforth eVCT) [19]. Identifying and demarcating the stand-clearing disturbance helps delineate the stand, and the number of years since disturbance is a proxy for stand age.
3. Landsat-based identification of planted pine stands: We used a map of planted pine forest stands in the Southeast USA generated by a combination of spectral and multitemporal Landsat data [20], [21]. The following were the prominent datasets used in this classification: 1) National Land Cover Database 2011, Landsat-based land cover map; 2) the difference between summer Normalized Difference Vegetation Index (NDVI) and winter NDVI from Landsat; 4) the global forest change dataset by Hansen et al. (2000-2013) [24] ; 5) the Vegetation Change Tracker dataset (VCT; 1985-2011 [23]); From the generated map, it was estimated that about 37% of the total forest area of Southeast USA is covered by pine plantations.

4.2.3 Canopy height model

We first developed a lidar-based dominant canopy height model for pine plantations that was applicable over the lidar coverage area shown in figure 4.1.

The field data for this effort came from the Forest Inventory and Analysis (FIA) section of the United States Forest Service [25]. We used data from FIA's *phase 2* (P2) plots. Figure 4.2 shows the nationally standardized sampling design for these plots. As part of the inventory, all standing trees of diameter at breast height greater than or equal to 12.7 cm (5.0 inches) are measured for crown height. The stated accuracy of such field height measurements is $\pm 10\%$. This is expected, as field height was measured using clinometers and there are limitations to these instruments in closed-canopy forest conditions.

We used a square north-south oriented lidar plot of size 30 meters as the other two Landsat-based data sources had map resolutions of 30 meters. The position of this square lidar plot with respect to the FIA plot elements is shown in figure 4.2. The corresponding field plot for this effort is *subplot 1* from the FIA, a circular plot of 7.32 meters (24.0 feet) radius, centered at the center of the FIA plot.

We applied the following criteria to select FIA plots in the southeast region, as per our domain of interest (planted pine forests):

1. The plots had to be *single-conditioned*. "Conditions" is a term used by the FIA to refer to major land-use types on the plots [25]. A consequence of FIA adopting fully randomized plot locations is that some plots can straddle multiple land-use types (eg: a plot may be 70% pine stand, 20% hardwoods stand, 10% water). Hence, we filtered out such non-homogeneous plots.
2. The plot was marked to have clear evidence of artificial regeneration (i.e., is a plantation).

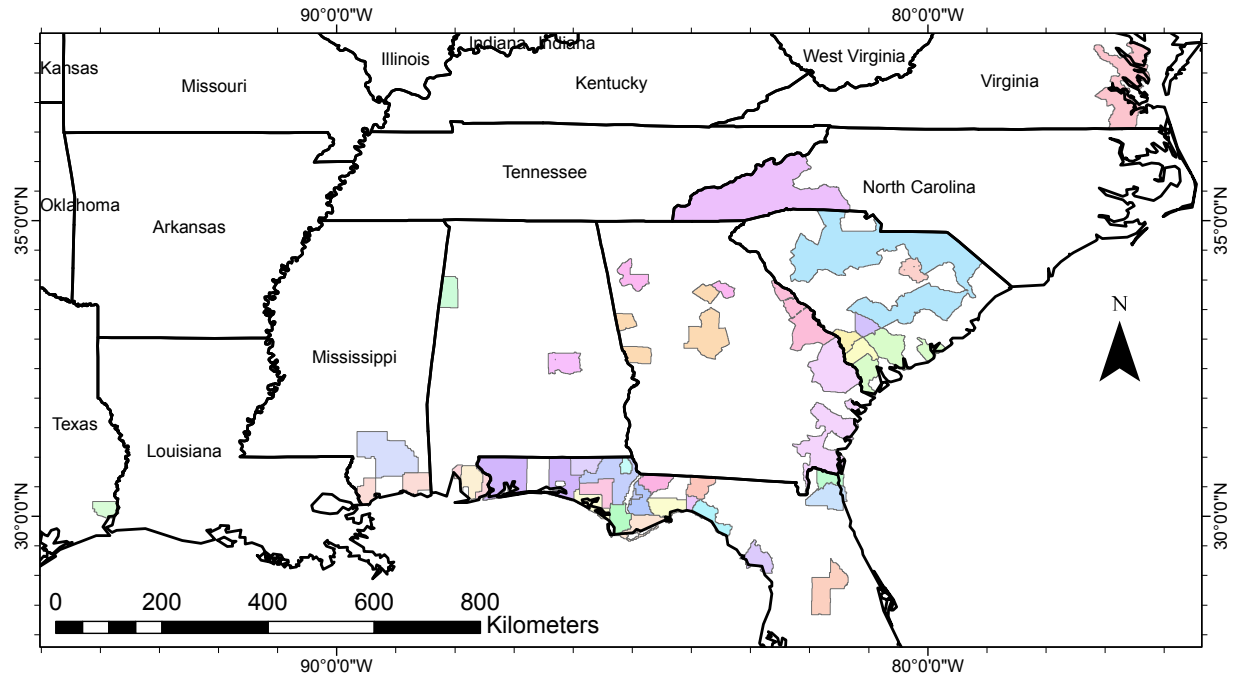


Figure 4.1: Lidar data from 38 projects spread across eight states were used in this study. Each project is shown in a different color. This region is dominated by loblolly pine, with some slash pine (*Pinus elliottii*) in the extreme south

3. The percent by basal area of the plot being of pine species was $\geq 95\%$.
4. Lidar data within 2 years was available from the year of visit and measurement of the FIA plots. That is, there is a maximum of ± 2 years difference between the years of lidar acquisition and of FIA field measurement.

After screening out plots that were either clear-cut between field measurements and lidar acquisition and plots with few uneven-aged trees, we had 217 FIA plots left, spanning 38 different lidar projects.

We extracted lidar plots for each of these 217 FIA plots from our lidar data. At this point, one should note that co-registration errors are an issue given that GPS instruments used by FIA are not survey-grade. But McRoberts et al. reported that most positional errors are within 15.0 meters [26], and had later estimated that most are within 7.0 meters (M. Russell, personal communication, October 2015). Hence, the relatively large size of the lidar plot (compared to the FIA center subplot) and the homogeneous nature of the forest stands selected (i.e., single-conditioned) addresses the co-registration issue adequately.

For each point-cloud of returns corresponding to the lidar plots, the following steps were done:

1. Ground classification using the lasground tool of lastools (version 141117; <http://rapidlasso.com/lastools/>). The default step size of 5.0 meters was used.
2. The understory points were removed using a threshold of 3.0 meters. This threshold was determined after inspecting 12 FIA plots (of the 217) for which understory height information was available.
3. Lidar distributional metrics: The standard percentiles of height above ground (h_5, h_{10}, h_{20} , etc) were extracted from the point cloud.

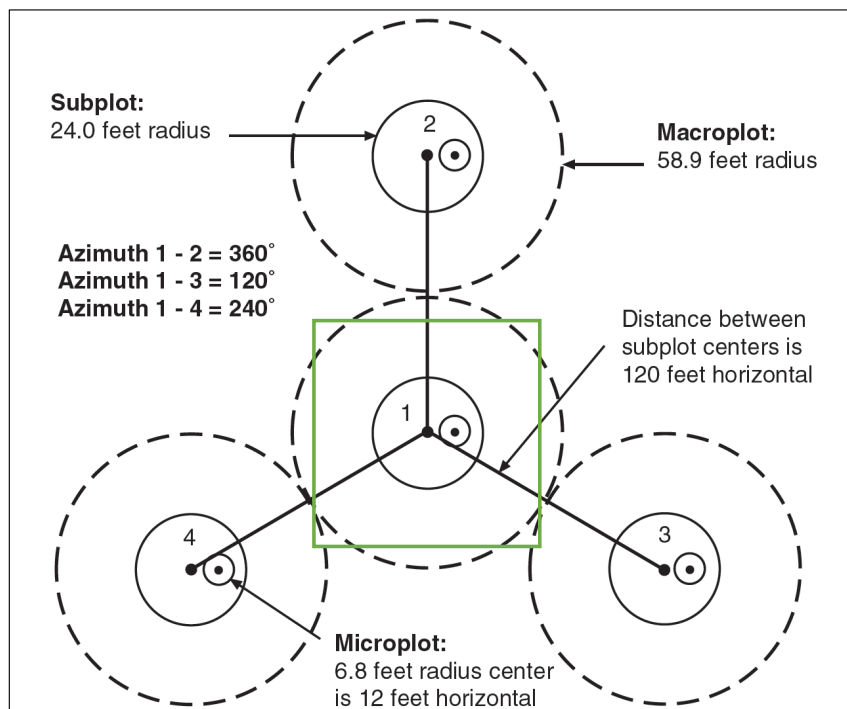


Figure 4.2: The lidar plot (green outline) in reference to the FIA plot (black markings). Figure adapted from [25].

4. The dominant height of all trees on our field plot (*subplot 1* of the FIA) was calculated as the average height of the five tallest trees on that subplot.

At this point, we had a set of square lidar plots of 30 meter size with associated field-measured dominant heights. First, pairwise correlations between the percentiles of height (h_5, h_{10}, h_{20} , etc) and the plot level dominant height were examined to select the most correlated percentile metric. Then, an OLS (ordinary least squares) based bivariate linear regression model was developed between this metric and the field-measured dominant height. We then ensured that the assumptions of linear regression (near-linear trend, normality of distribution of variables, homogeneity of variances) were adequately met.

4.2.4 Generation of a sample site index map from lidar and Landsat data

The generation of canopy height maps for the entire lidar coverage area shown in figure 4.1 is beyond the scope of this study as it involves processing a relatively huge amount of data (few terabytes). Hence, we decided to concentrate on a smaller region shown in figure 4.3. We generated site index maps for all pixels labelled as planted pine in this region. Here, plantations are almost always of the loblolly pine species (for a detailed study on the ecology and culture of this species, refer to [27]). To confirm this, we examined 211 FIA plantation pine plots (as of year 2009) in the region of lidar coverage. Approximately 94% of these plots were loblolly pine and 5% were longleaf pine. We used a dynamic site index equation based on a long-term loblolly pine plantation study that used a large number ($n=186$) of permanent sample plots [28]. These plots adequately sampled the loblolly pine range stretching from Virginia (in the east) to Texas (in the west). We only considered stands that were at least 15 years of age, as site index measurements are more reliable for such older stands. The site index was calculated for each 0.09 ha (30x30m) Landsat pixel inside these forest

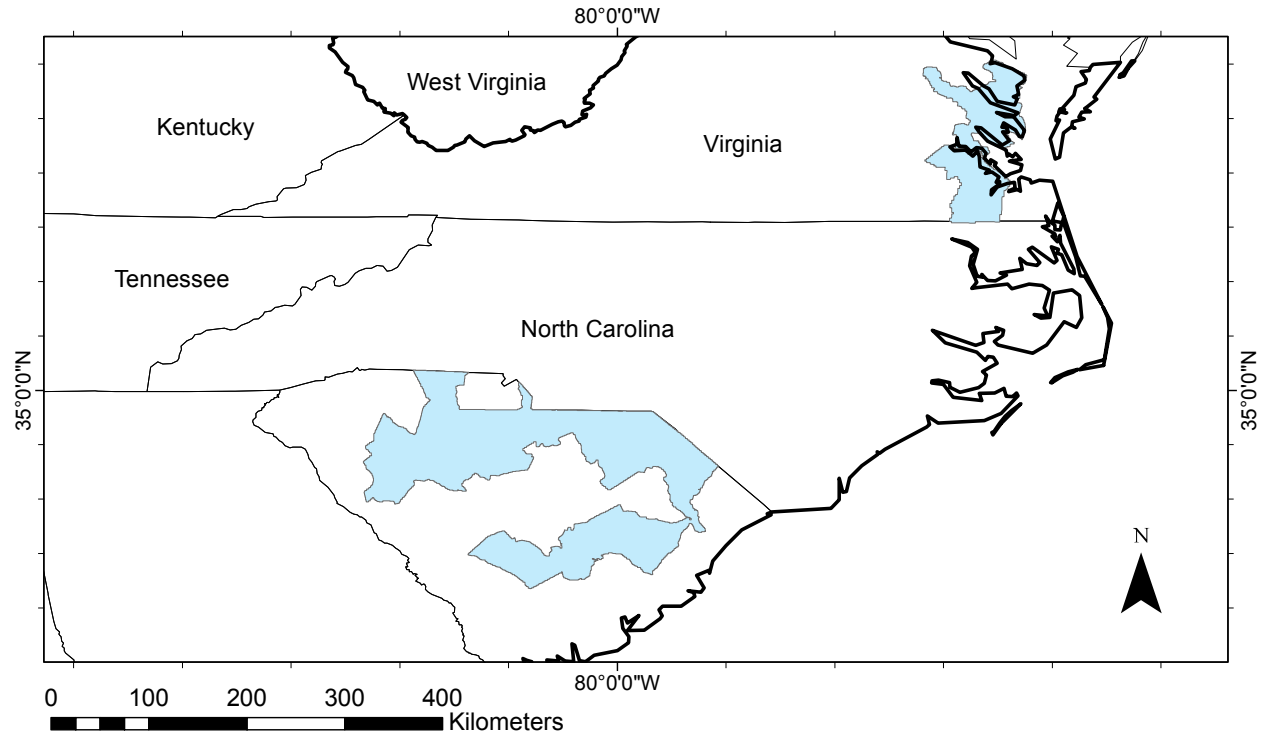


Figure 4.3: The extend of the two lidar projects over which the site index mapping was done are shown in blue shading

stands; this will help us bring out the intra-stand variation in site index. The workflow for the generation of this map is shown in figure 4.4. We then calculated the proportion of area in a low site index class versus that in a medium to high site index class using a threshold of 22.0 meters.

4.2.5 Accuracy assessment of site index predictions

We used a monte-carlo approach to estimate the accuracy of the site index map. This method has been used in other remote-sensing based studies to propagate uncertainties in model input parameters [29]. This approach is popular when the output (site index) is a complex function of the inputs and when the distribution of the errors in the inputs are independently known or can be estimated.

In our case, the dynamic site index equation [28] states that:

$$\text{Site Index} = fn(\text{Dominant height}, \text{age}) \quad (4.1)$$

The error distribution of the two right-hand-side terms of equation 4.1 are estimated in the following way:

1. Dominant height: This would be a normal distribution of mean 0.0 and standard deviation (σ) equal the root mean square error of the OLS-based linear model.
2. Age, predicted by eVCT: For this, we used the following independent ground-truth sources:
 - (a) Appomattox-Buckingham State Forest plot set: This is a set of 23 loblolly pine forest plots located in stands of known age in central Virginia, USA [30]. We have good estimates of the age of these

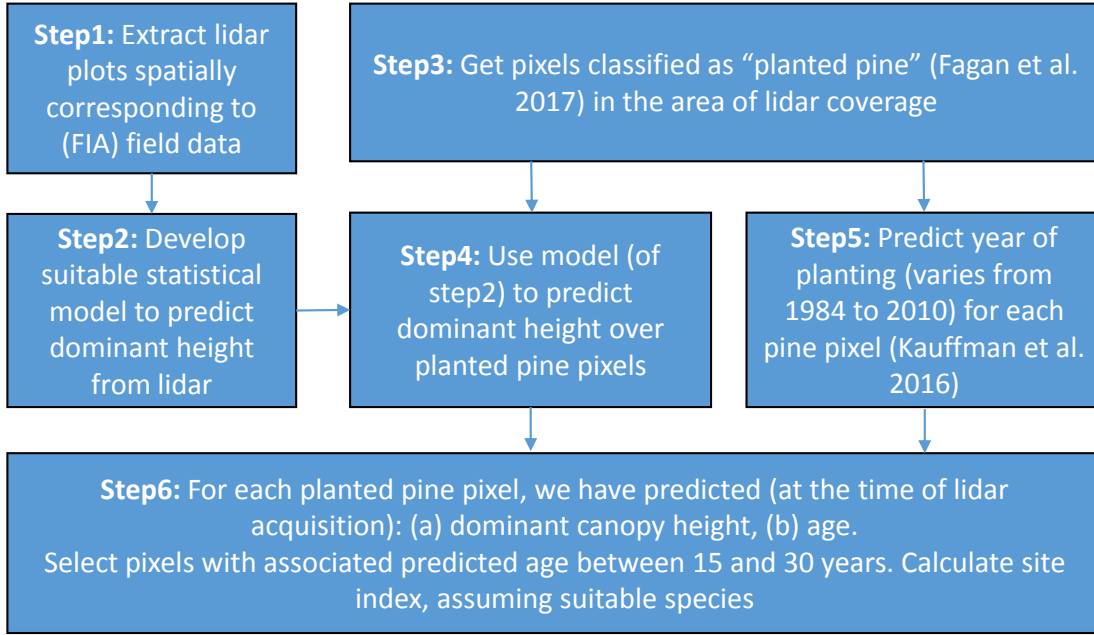


Figure 4.4: Major steps involved in the generation of the site index map

stands, based on planting records. But these plots are relatively few in number and their coverage is geographically limited.

- (b) FIA plot set: This is a set of 75 FIA plantation pine plots for which ages (year of planting) were available. The accuracy of these is variable, as the availability of planting records and reliable tree core data is limited. The lack of agreement between the FIA field estimates of age and eVCT is *assumed* to be due to erroneous eVCT ages. Note that this *may not always be true*, especially given the uncertainty in FIA age estimates.

The eVCT-based age error distribution was then estimated by randomly sampling among these plots, and calculating the difference between the ages.

A paired set of $(\hat{y}_i, y_i)$ values were generated where $\hat{y}_i$ is the lidar-based site index predictions, y_i is the *estimated* reference values (obtained using the error estimates above) and $\bar{y}$ is the average of these reference values. The bias, the RMSE and the relative RMSE (%RMSE) were then calculated using standard formulas:

$$bias = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i) \quad (4.2)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (4.3)$$

$$\%RMSE = \frac{RMSE * 100.0}{\bar{y}} \quad (4.4)$$

We then spatially aggregated the pixel-level site index values to distinct hydrological units (corresponding to watersheds, henceforth HUs), a biophysically meaningful unit of aggregation for forest vegetation studies

[31]. The 12-digit hydrologic unit (HUC-12) was the unit of aggregation used. We also overlaid level two (L2) ecological regions from the US Environmental Policy Agency (EPA) over the maps generated [32] as they have been used previously to partially attribute measured plot-level differences in forest productivity [33], [34]. There are two such ecological regions within our lidar coverage area; the EPA labels them as 'Mississippi alluvial and Southeast USA coastal plains' (along the coast, henceforth 'coastal plains') and the 'Southeastern USA plains' (interior, henceforth 'Southeastern plains'). The aggregation operation on the raster was done using 50,000 sample points, for better computational tractability. The Kruskal-Wallis rank-based test was used to ascertain if the differences in means (using the samples) noticed at the HU level were statistically significant. Then, the Steel-Dwass all-pairs test (a non-parametric version of Tukey's HSD test) was used to identify groups of HUs with such statistically significant differences in site index values.

4.2.6 Change in site index

Historical loblolly pine site index values for the region were obtained from the USDA Natural Resource Conservation Service's Soil Survey Geographic Database (henceforth SSURGO) [35]. As this data was mostly collected in the 1960 and 1970s, this is a representation of the site index that had prevailed during those times. Loblolly pine site index was converted from base age 50 to 25 by using the base age invariant site index model developed by Diegues-Aranda et al. (2006) [28]. We then estimated the change in site index over the region over the past four decades by subtracting the SSURGO site index from the lidar-predicted site index (again, using a representative sample of 50,000 points). This was also aggregated to the HU level. The Kruskal-Wallis test and the Steel-Dwass all-pairs test were again used to analyze the statistical significance of the changed site index values at the HU level.

4.3 Results

4.3.1 Canopy height model

We selected the 70th percentile of lidar heights (h_{70}) as the independent variable for the regression model after examination of the correlation between plot-level lidar distributional metrics and dominant canopy heights. The correlation of h_{70} with dominant heights over all plots was 82%.

The linear regression model has the form:

$$\text{Dominant height} = 0.90138 * h_{70} + 4.2867 \quad (4.5)$$

The R^2 of the fit was 0.84 and the RMSE was 1.85 meters. Also, h_{70} was highly significant ($p < 0.0001$). The scatterplot illustrating the fit can be seen in figure 4.5.

4.3.2 Accuracy assessment of site index map

The accuracy of site index values generated was assessed using the methods outlined in subsection 4.2.5). We used:

1. The Appomattox-Buckingham State Forest plot set: The bias is estimated to be -0.75 meters, the RMSE estimate is 3.0 meters, the relative RMSE is 15%.
2. FIA plot set: In this case, the estimated bias is -0.29 meters, the RMSE is 3.8 meters, the relative RMSE is 19%.

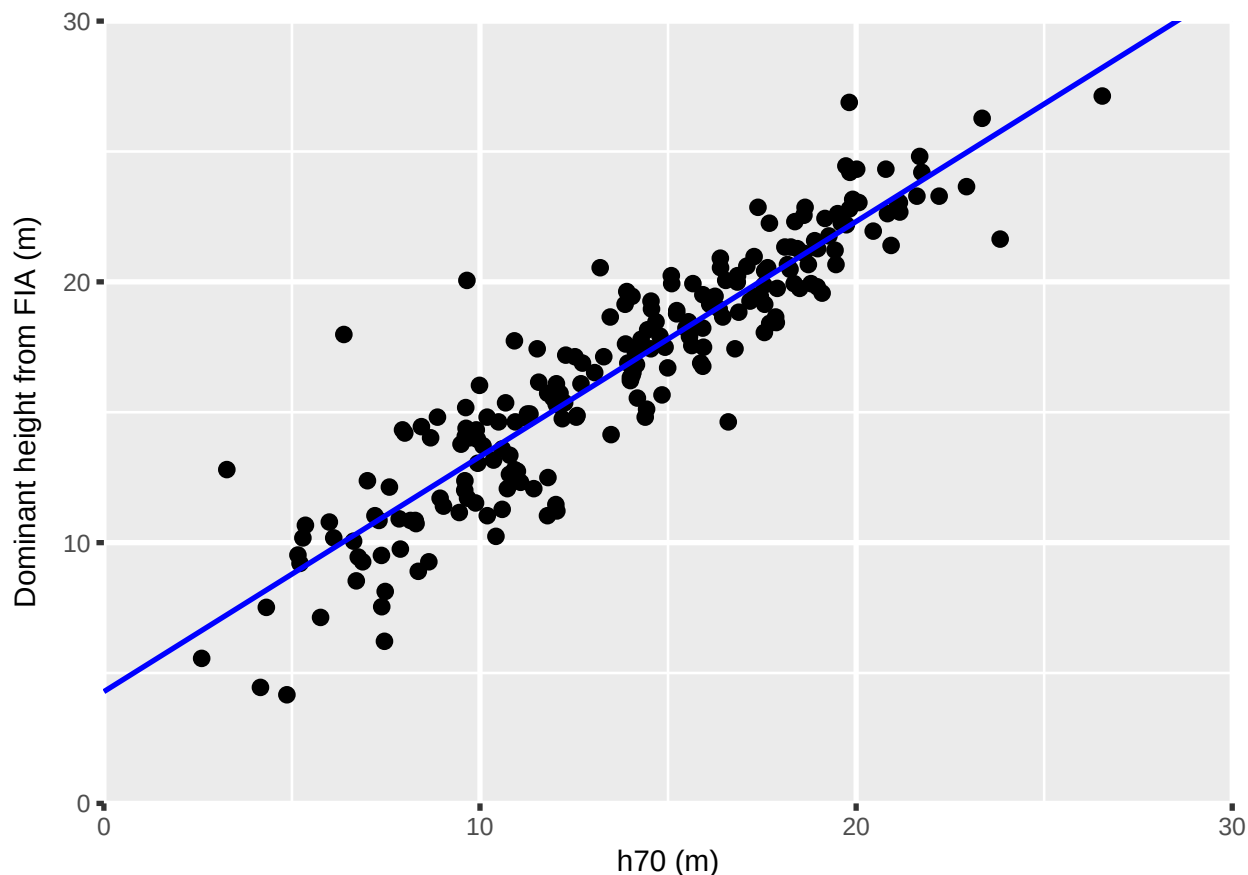


Figure 4.5: Scatterplot showing the plot-level agreement between the 70th percentile of lidar first return heights and the FIA-measured average dominant tree height ($n = 217$ plots). The fitted linear model is indicated by the blue line

Overall, the bias is quite low, while the RMSE is consistent with other similar approaches, thus affirming that one can use non-forestry-quality lidar data from disparate projects combined with Landsat-based age maps to reasonably predict site index over large areas (refer to our first research objective). The accuracy values compare favorably with levels reported by Tompaski et al. (bias=0.7 m, RMSE=5.5 m) [17] which used similar remote sensing data sources for site index prediction.

We generated a site index map as per the steps outlined in section 4.2. A total of 22027 forest stands of ages between 15 and 27 years (at the time of lidar acquisition) were identified by the combination of eVCT [19] and Landsat-based pine stand identification algorithms [20]. The site index was predicted at the individual Landsat pixel level; that is, for a total loblolly pine forest area of 832 sq. km. As the generated site index map is too large and too sparse for presentation in its entirety here, we show a sample section in Newberry County, South Carolina (see figure 4.6).

The distribution of site index can be seen in figure 4.7, with a median value of 19.8 meters, a 5th percentile value of 12.6 and a 95th percentile value of 24.0 meters. The proportion of area calculated under the two site index categories is shown in table 4.1.

Most of the HUs have moderate to high site index values (greater than 18.5 meters). A distinct divide can be noticed with respect to the ecoregions. Almost all the high site index HUs are in the coastal plains while

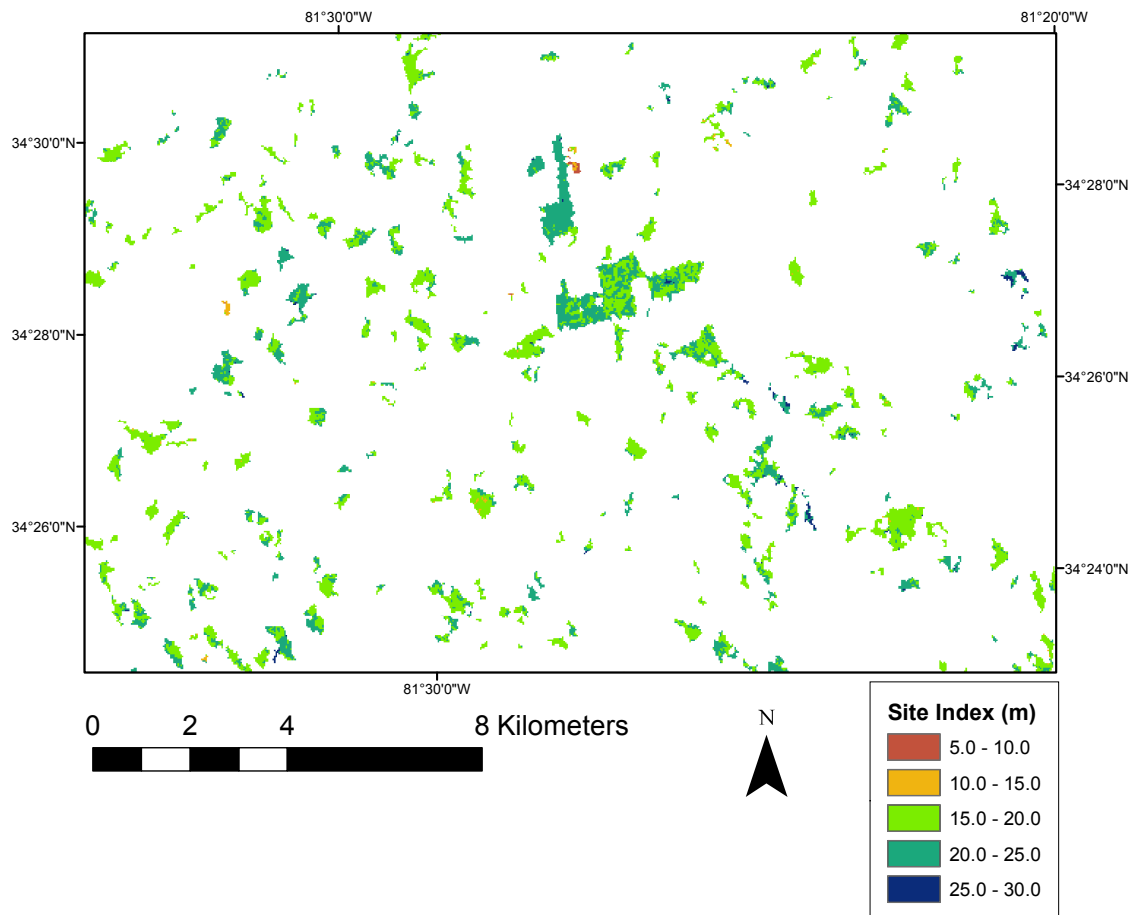


Figure 4.6: A part of the site index map generated is shown here at a scale where forest stands are visible (Newberry County, South Carolina). A very small percent of pixels (less than 0.1%) where site index values more than 30.0 meters were predicted are not shown (as they were mostly model over-predictions). Most stands register reasonable to good site index values (greater than 20.0 meters) while some under-perform (less than 10.0). A fairly large variation of site index can be seen over the landscape, too.

almost all the low site index HUs are on the right side of the figure, in the Southeastern plains ecoregion. This may indicate that coastal plain soils and climatic conditions are better suited for forest vegetation growth. HUs that perform poorly seem to cluster together, which seems to indicate that their performance is more strongly linked to broad geographic reasons, rather than site-specific management ones.

Table 4.1: The proportion of area surveyed falling in two site index classes. Only 17.4% of the area is in a medium-to-high productivity class

Site index range (meters)	% area	Total area (sq. km)
Less than 22.0	82.6	687.2
Greater than or equal to 22.0	17.4	144.8

A map of the difference between lidar predicted loblolly pine site index and SSURGO estimated site index was generated. A histogram of the differences over the entire lidar coverage region can be seen in figure 4.9. A general increase in site index is noticed for the whole region, mostly attributed to management interventions [8], [37], although the increase in carbon dioxide concentrations ([38], [39]) and nitrogen deposition may be

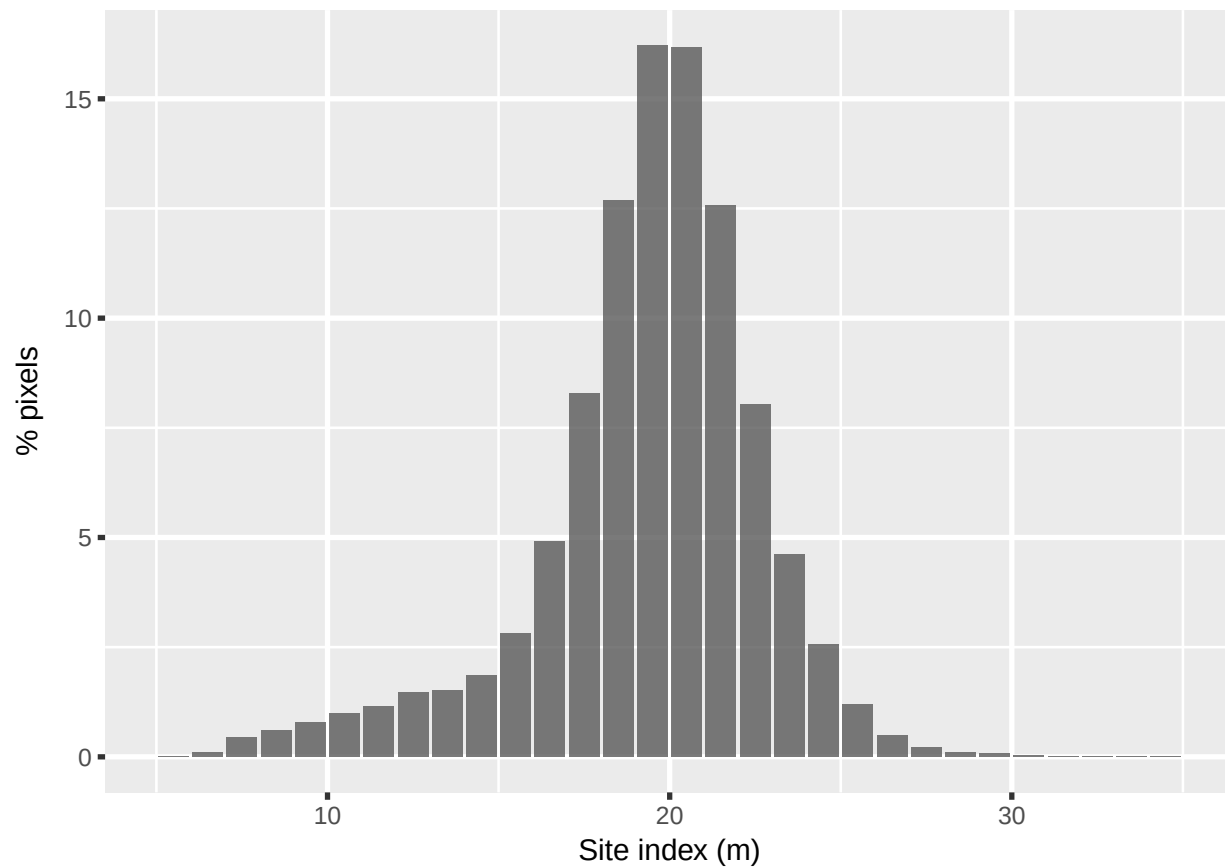


Figure 4.7: The distribution of site index values over the surveyed landscape. More than 50% of the area has a site index value less than 20.0 meters (65.6 feet)

factors. Notably, around 16% of the area surveyed registered a decrease in site index over the years. A similar phenomenon has been reported by Zhao et al. (2016) [8] where they report that as much as 26% of the 850 plots surveyed had a near-zero or negative increase in site index. We aggregated the site index difference map to distinct HUs (see figure 4.10). Here, in South Carolina, a strong trend of large productivity increases in the western interior parts can be seen. This indicates that management interventions may have been most effective on the poorer soils of those interior regions.

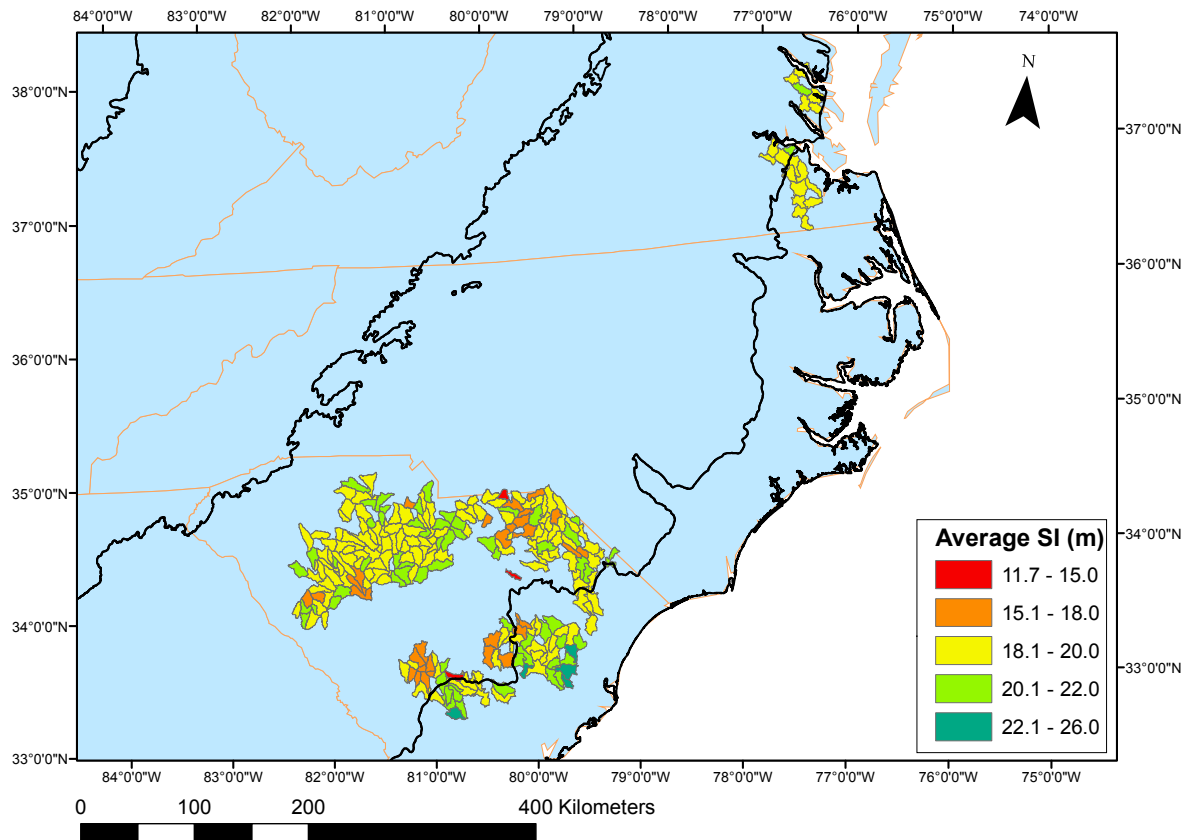


Figure 4.8: The average site index for all stands that fall in the hydrological unit is shown. Only hydrological units with at least 4.5 ha of surveyed forest stand area (50 pixels) are shown. The thick black lines demarcate EPA level 2 ecoregions (Omernik 1987), see text for more details. The areas of decreased productivity (less than 18.0 m) near the transition zone between these two ecoregions roughly correspond to the South Carolina sandhills region, where poor soils are documented [36]

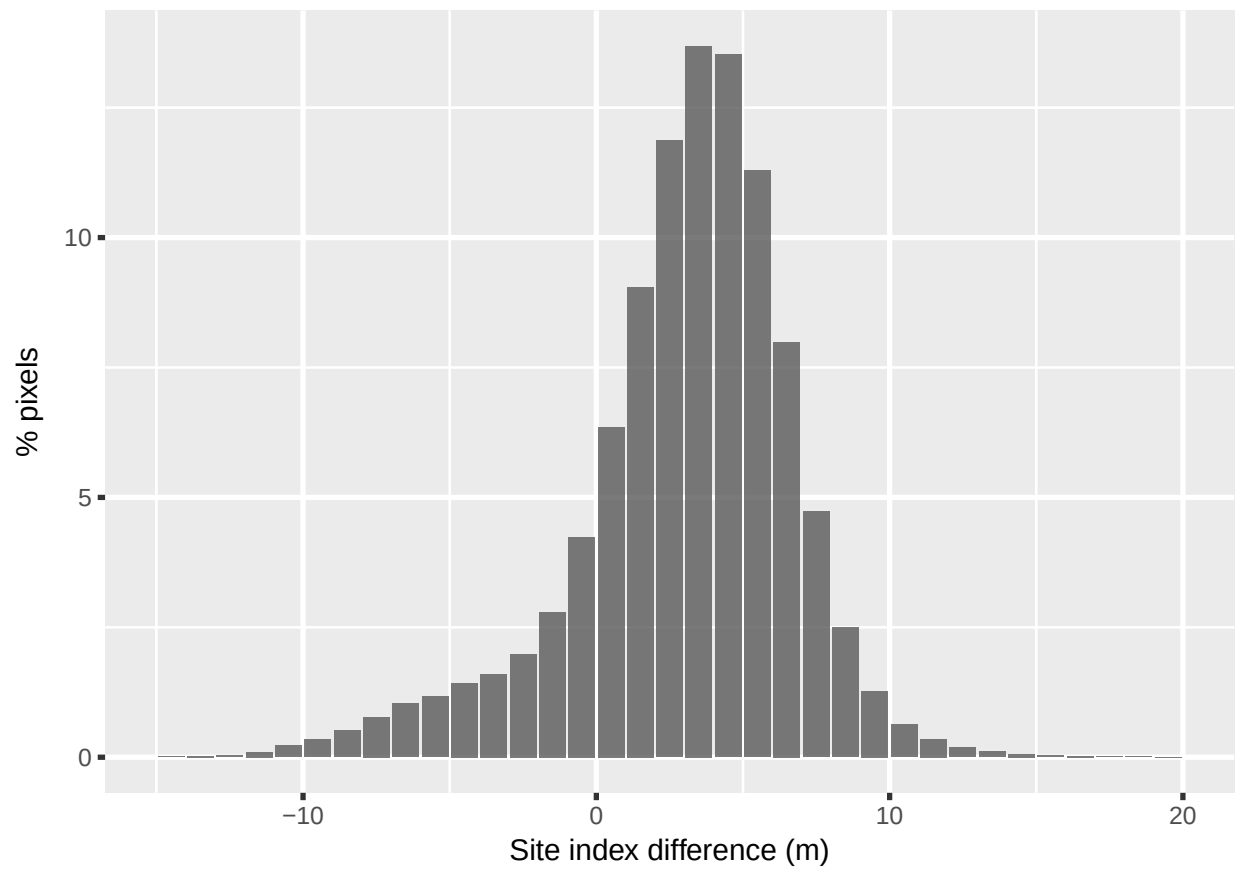


Figure 4.9: The difference between lidar-predicted site index (as of 2008 and 2010) and SSURGO-estimated site index (around the 1970s).

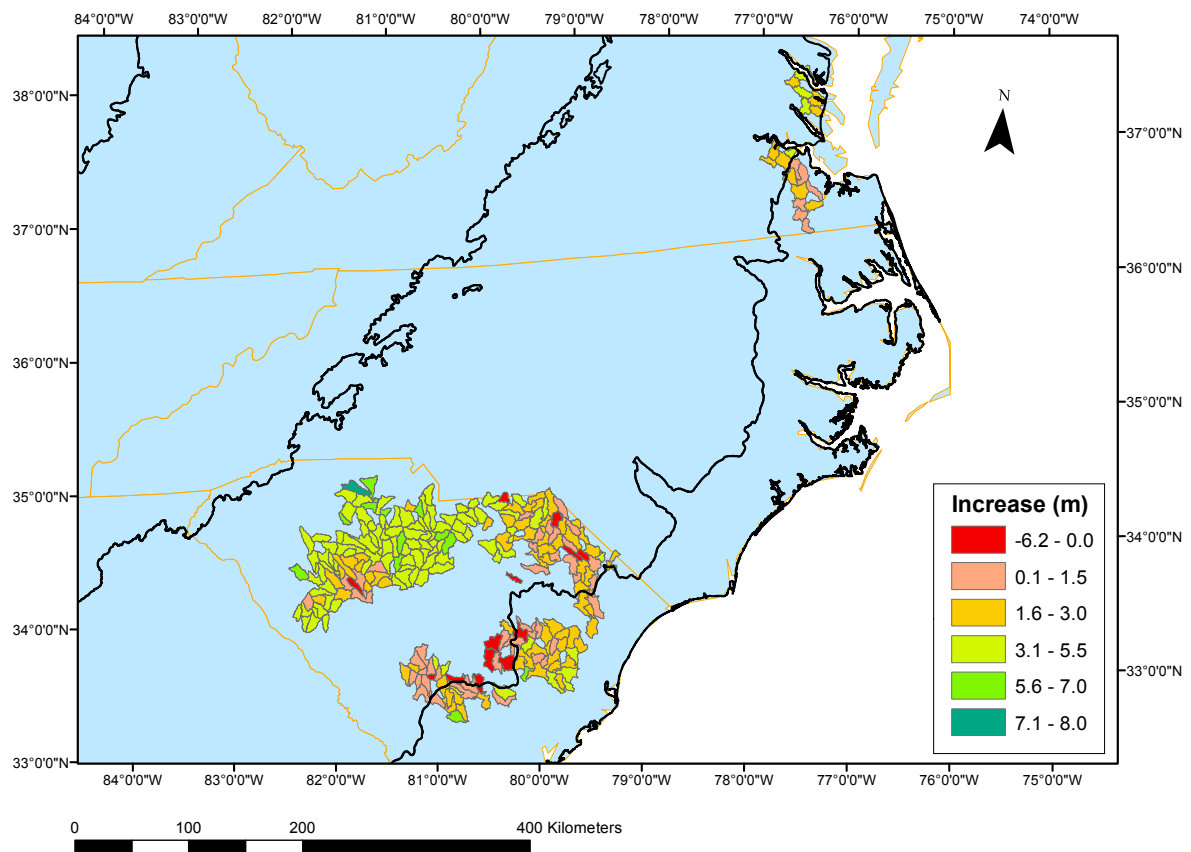


Figure 4.10: The average site index difference (between current and historical SSURGO estimates) for stands that fall in the hydrological unit is shown. Only hydrological units with at least 4.5 ha of surveyed forest stand area (50 pixels) are considered. The connotation of the thick black line is same as that in figure 4.8.

4.4 Discussion

In this study, we have combined various remote-sensing products to generate a site index map for planted pine in large areas of Southeast USA. The accuracy of the map is contingent on several factors: (1) the efficacy of the planted pine stand detection algorithm [20], [21], (2) that all such detected stands are of loblolly pine (as we assumed), (3) the applicability of the site index equation-form to all loblolly pine stands in the region [28], and (4) robustness of the dominant height and age predictions (discussed in detail below). The species assumption we made is that all the plantation pine in figure 4.3 is loblolly pine. This is not the case; a small number of longleaf pine (*pinus palustris*) plantations are also present in that area (estimated to be around 5%). Nevertheless, these two species have similar growth patterns [40] and hence the errors associated with these assumptions are not expected to be large. Alternatively, refinements to the algorithm used for selecting planted pine may allow for selection of pure loblolly pine pixels. Although the site index equation used here was from a study covering much of the native range of loblolly pine and age classes [28], it may not be representative for some forest stands. An alternative way to make site index predictions is the *guide curve method* [17], where a one-time measurement of a large sample of trees is used to construct a site index model. Our methods are more similar to Wulder et al. (2010) where prior site index curves (for specified species) were used along with lidar predictions of stand height and inventory-based stand ages [16].

Site index predictions are sensitive to errors in associated prediction errors of the age model. To further understand and quantify this in our case, we first analyzed the agreement of forest stand ages predicted by eVCT (remote sensing) and those based on planting records or field visits (figure 4.12). Age predictions of eVCT are relatively unbiased, which explains the unbiased nature of our site index predictions. The absolute age bias of approximately 1.1 years in both cases shown in the figure is similar to the bias of 2.2 years reported by Tompalksi et al. [17], where fixed *disturbance index* thresholds were used to detect stand-clearing disturbances from Landsat time-series maps. In figure 4.12 (b), one can notice that a large number of plots were assigned an age of either 24 or 26 years. This is because a corresponding large number of plots were assigned high disturbance magnitudes by the underlying VCT algorithm for the year 1984. This was the year Landsat 5 was launched, and as such, the beginning of the VCT time series. VCT is known to ascribe less reliable disturbance magnitudes at the beginning of the analyzed time series, in some instances [23].

We further looked at a small set of planted pine plots ($n=19$) where we had both site index predictions from lidar and field based site index estimates from FIA (figure 4.13). The R^2 of the fit is low: 1.8%. Similar regressions of lidar-derived site index with values from forest inventories have produced either low (0.03 in [16]) or medium (0.42 in [15]) R^2 values. Nevertheless, in our case, if one considers plots where both age estimates are within ± 5 years ($n=12$), the R^2 of the fit improves to 46%; similarly for ± 3 years ($n=9$), the R^2 is 77.5. This can be seen in figure 4.13, where plots with similar age estimates (from eVCT and FIA) tend to be near the 1:1 line, which shows that age discrepancies are the major cause of the overall lack of fit. Over-detection of disturbances at the year 1984 (see discussion above) is clearly responsible for many age discrepancies. Other likely causes have to be further investigated. Figure 4.13 also implies that dominant heights are predicted well from lidar data. This is in agreement with previous literature that lidar estimates of plot-level height metrics are quite robust [41]–[43]. Our estimates of the region-wide increase in site index since the 1970s (see figure 4.9) compares well to a similar estimate by Subedi et al. (2015) [44]. In that article, the authors compared the current site index value with that of SSURGO-based values for various loblolly pine stands across 21 study sites, spanning the geographic breadth of the Southeast. They found that site index had increased by around 5 to 7 meters (on an average) on these stands. This compares well with the 4 meter increase that we observe over the whole region. It should also be noted that we estimated a negative bias of approximately 0.3 to 0.7 for our site index prediction method (see section 4.3.2), which means that the actual increase could be closer to 5.0 meters.

An interesting method of simultaneous computation of both age and site index maps from multi-temporal canopy height rasters (4 images per stack, spread over 58 years) is presented in Vega et al. (2009) [45].

Height maps were generated from historic aerial photographs by stereoscopic methods. Age and site index pairs that best fit (i.e., minimum residual) the four canopy height values at each gridcell are then selected. A similar approach could be designed for our study area where high-resolution digital photos (from which heights can be extracted using stereo matching) are available.

It is forecast that the area of planted pine land use in Southeast US could more than double in the next 40 years. It is expected to grow from 19% of the total forest area (thus, from 39 million acres of planted pine in 2010) to between 24 and 36% by 2060 [9], which highlights the relevance of this research. HUs registering low site index in figure 4.8 should be further analyzed and considered for possible management and policy interventions. For example, macronutrient addition has been found to be widely effective; a large and consistent growth increase (around 25%) after midrotation fertilization with nitrogen and phosphorous has been reported on the majority of soil types [37]. Deficiencies of other nutrients such as potassium, calcium, copper and boron have also been implicated [46]. Other soil-related issues pertinent to pine plantations such as texture, compaction, soil aeration and moisture and pH value [47] may be harder to address. This may be the reason Zhao et al. (2016) reported that as much as 26% of plots was found to have poor response to treatments (such as nutrient additions) on a regional level [8]. Nevertheless, there seems to be ample scope for additional management interventions; for example, it was reported by Han et al. (2007) that percent of regional total greenhouse gas emission sequestered by forests in 11 Southeast states could be increased from 13% to 23% through proper policies and using best land management practices [2]. Lastly, the modeling community could also benefit from such site index maps: Two classes of models that could be better calibrated and validated with such maps are forest productivity models and regional wood supply models. For example, Subedi et al. (2015) points to a mechanism by which stand-level forest productivity could be predicted (and updated for changing climate conditions and management regimes) for loblolly pine stands on a regional scale [44]. They proposed a method by which SSURGO-based soil fertility rating could be used as input to $3PG_{lob}$, a loblolly pine productivity model. Our site index predictions would be useful as validating reference data for such efforts. Alternatively, our forest productivity estimates could be used to deduce the intensity of management for specific sites.

The spatial variability of site index within a forest stand is an important but poorly-studied topic, with interesting associated questions [48]. Variations in microsite quality have been implicated to explain the diversity of tree heights and crown shapes seen in stands [49]; genotype variation may also be a cause. Soil differences and related moisture and nutrient differences have caused mature tree heights to range from 15 to 34 meters within a short distance [27]. Large variations in site index have also been reported in the Southeast, from 10 to 30 meters [27]. Figure 4.11 illustrates such a variation in site index. We have selected stands of uniform age (predicted by eVCT) so that site index prediction variations are solely linked to canopy height. A wide range of site index values is predicted within both the stands. Some contiguous patches within the stands seem to be better managed, on the whole. For example, the patch on the left in figure 4.11(b) has an overall high site index. Forest patches with low productivity can be associated with poor leaf area index (LAI) development, as a strong relationship between the two has been demonstrated [50]. It has been reported that stands often fall short of potential LAI (about 4.0) primarily due to limitations of macronutrients such as nitrogen and phosphorous observed at mid-rotation [47].

The results of this study have the following implications: (a) Lidar acquired for mostly non-forestry purposes (such as topographic mapping) combined with Landsat-based land-cover tracking algorithms, is a good tool that enables periodical assessment of site index trends for large areas, especially given that site index is a dynamic quantity and may need periodic updates [16], (b) A large proportion of loblolly pine forest stand area has low site index, and may have a sizable potential for additional carbon sequestration through better management interventions, and (c) There is a substantial variation of site index when calculated at an intra-stand level.

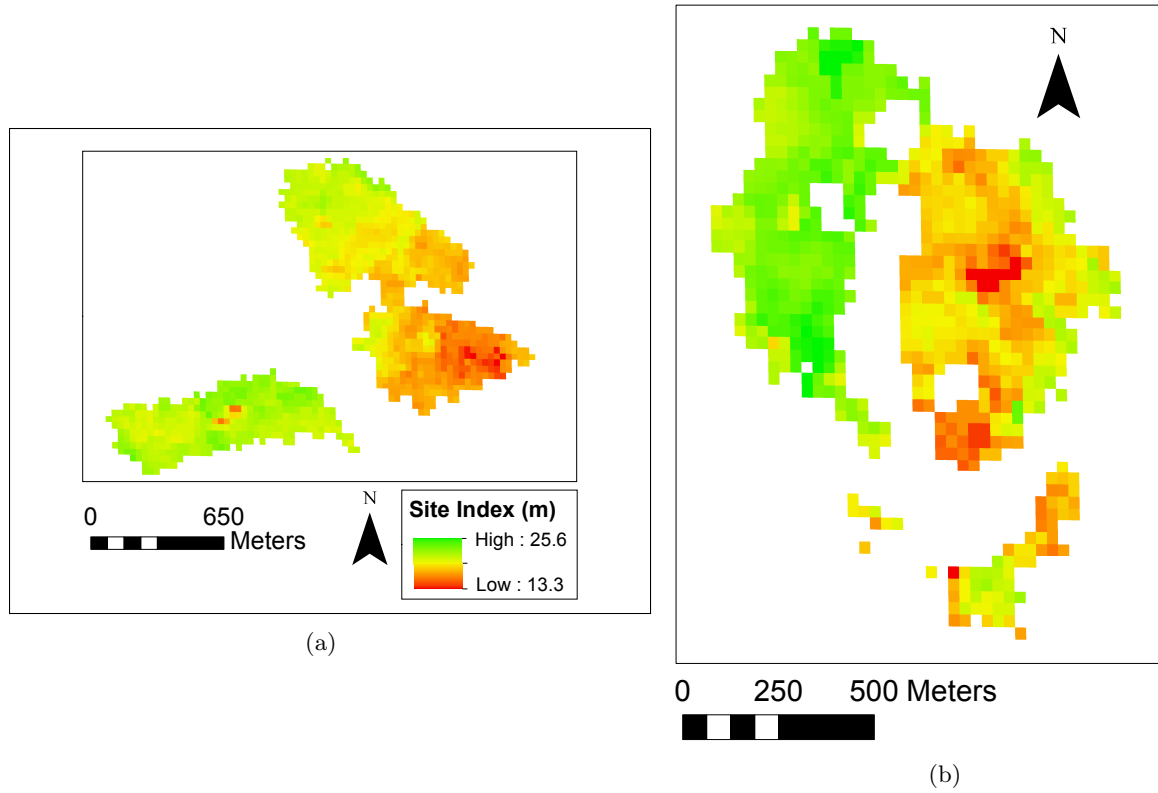


Figure 4.11: (a) Variation of site index for a uniformly-aged stand (age = 26 years) in the Isle of Wight county, Virginia, centered at (latitude, longitude) = (36.862, -76.711) decimal degrees, (b) Variation of site index inside a uniformly-aged stand (age = 21 years) in Williamsburg county, South Carolina, centered at (latitude, longitude) = (33.406, -79.781) decimal degrees. The scale here is the same as (a). Site index varies by as much as 12 meters (40 feet) within the stand.

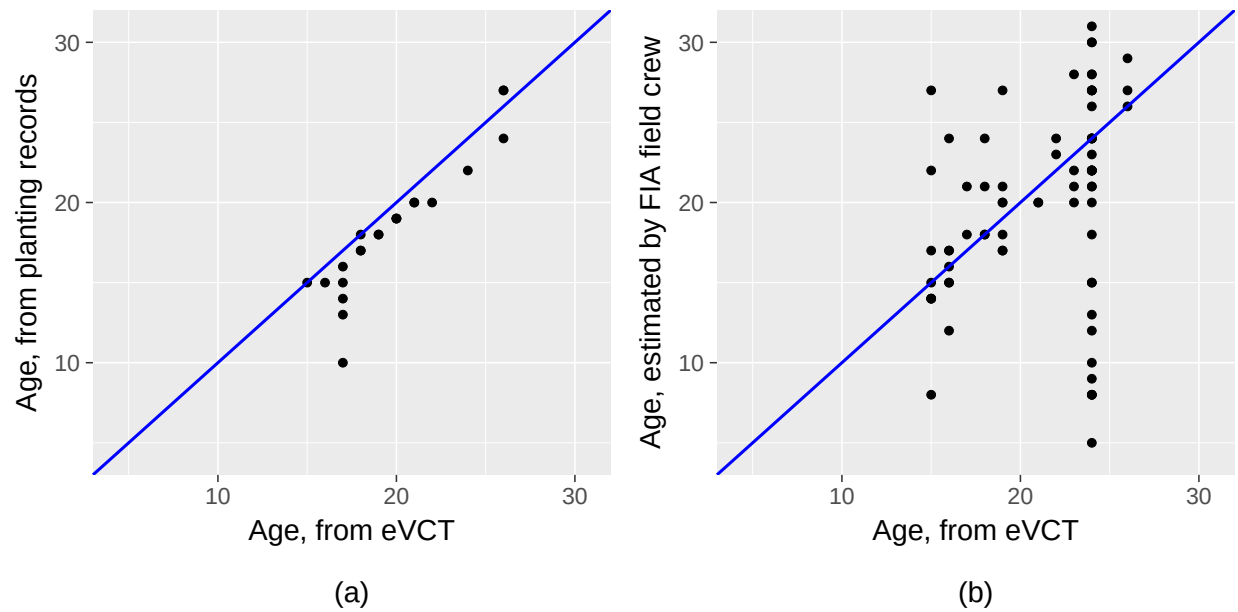


Figure 4.12: Correspondence between forest stand ages predicted by eVCT and estimated from other sources (all ages are in years). The 1:1 line is shown in blue. Only plots that were indicated by eVCT to be either 15 years of age or above were examined. (a) Age agreement, using planting record data from the Appomattox-Buckingham State Forest plot set (23 plots, see section 4.4.3). One extreme outlier point at $(x, y) = (10, 72)$ is excluded. It can be seen that most points are near the 1:1 line. The bias is -1.1 years. Around 87% of plots have ± 3 years age agreement while 91.3% of plots have ± 5 years, (b) Age agreement, using the set of 75 FIA plots. The pronounced "vertical line" of points corresponding to eVCT age = 24 is most likely due to VCT being less reliable at the beginning of the time series, see text for more details. The overall agreement is much weaker in this case. The bias is rather small (1.1 years). Around 65.3% of plots have ± 3 years age agreement, 73.3% of plots have ± 5 years and 85.3 have ± 8 years.

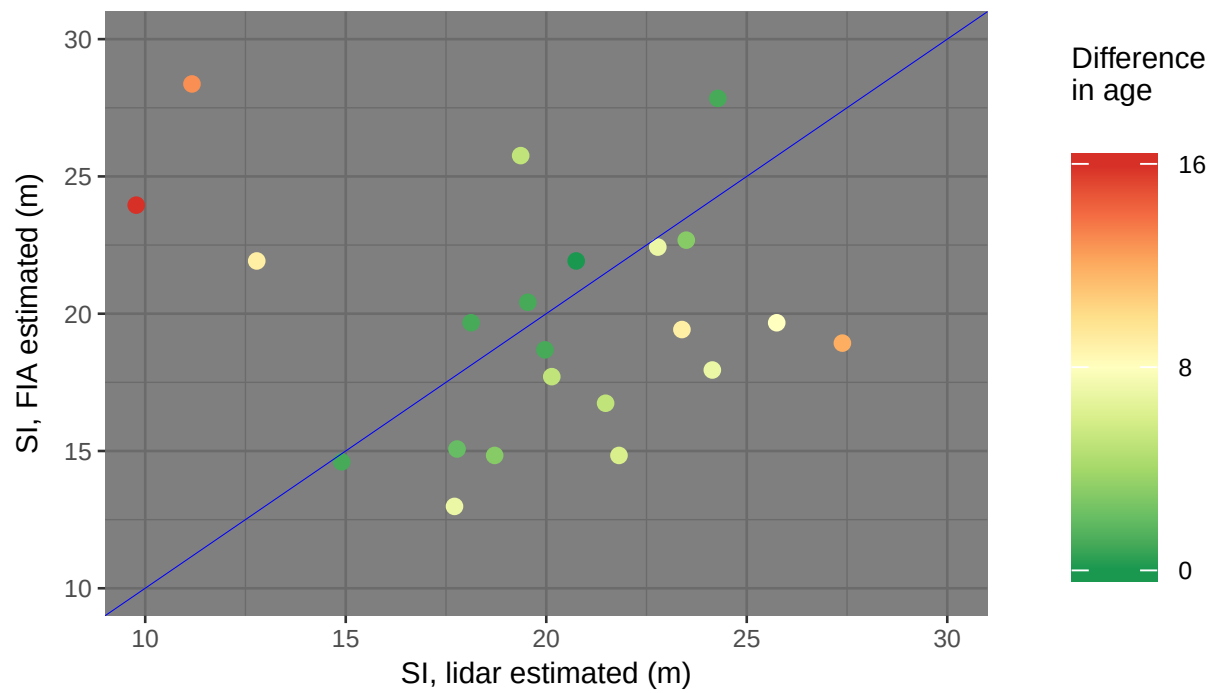


Figure 4.13: Correspondence between site index predictions from lidar and estimates from FIA for 19 planted pine plots. The absolute difference in age (years) predicted by eVCT and estimated by FIA is also indicated. The 1:1 line is shown in blue

4.5 Conclusions

In this study, we presented a method to generate site index estimates over large areas in the Southeast US, an important region with respect to carbon sequestration and forest plantation activity. Several areas were identified as having low productivity in Virginia and South Carolina, which might have important ramifications for land-use policy makers and forest managers. Further research aimed at explaining the cause of such low productivity could lead to insights pertaining to forest ecology and socio-economics of land use. Periodic generation of such site index maps when updated lidar data becomes available could support a plethora of scientific needs such as better understanding of the impacts of climate change, facilitation of better carbon budget modeling and better calibration and development of anthropogenic land-use models.

Bibliography

- [1] D. P. Turner, G. J. Koerper, M. E. Harmon, and J. J. Lee, "A carbon budget for forests of the conterminous United States," *Ecological Applications*, vol. 5, no. 2, pp. 421–436, 1995. [Online]. Available: <http://onlinelibrary.wiley.com/doi/10.2307/1942033/full>.
- [2] F. X. Han, M. J. Plodinec, Y. Su, D. L. Monts, and Z. Li, "Terrestrial carbon pools in southeast and south-central United States," *Climatic Change*, vol. 84, no. 2, pp. 191–202, 2007.
- [3] U. D.o.A.U. F. Service, *Future of America's forest and rangelands: Forest Service 2010 Resources Planning Act Assessment. General Technical Report GTR-WO-87*. USDA Forest Service Washington, DC, 2012.
- [4] J. P. Skovsgaard and J. K. Vanclay, "Forest site productivity: A review of the evolution of dendrometric concepts for even-aged stands," *Forestry*, vol. 81, no. 1, pp. 13–31, 2008. [Online]. Available: <http://forestry.oxfordjournals.org/content/81/1/13.short>.
- [5] P. J. Sellers, W. J. Shuttleworth, J. L. Dorman, A. Dalcher, and J. M. Roberts, "Calibrating the simple biosphere model for Amazonian tropical forest using field and remote sensing data. Part I: Average calibration with field data," *Journal of Applied Meteorology*, vol. 28, no. 8, pp. 727–759, 1989. [Online]. Available: [http://journals.ametsoc.org/doi/abs/10.1175/1520-0450\(1989\)028%3C0727:CTSBMF%3E2.0.CO%3B2](http://journals.ametsoc.org/doi/abs/10.1175/1520-0450(1989)028%3C0727:CTSBMF%3E2.0.CO%3B2).
- [6] M. Launay and M. Guerif, "Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications," *Agriculture, ecosystems & environment*, vol. 111, no. 1, pp. 321–339, 2005. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167880905002902>.
- [7] B. Dyck, "Precision forestry—The path to increased profitability," in *Proceedings of the 2nd International Precision Forestry Symposium, Seattle, Washington; University of Washington, Seattle*, 2003, pp. 3–8. [Online]. Available: http://www.academia.edu/download/4565218/pfc_final_symp_03.pdf#page=12.
- [8] D. Zhao, M. Kane, R. Teskey, T. R. Fox, T. J. Albaugh, H. L. Allen, and R. Rubilar, "Maximum response of loblolly pine plantations to silvicultural management in the southern United States," *Forest Ecology and Management*, vol. 375, pp. 105–111, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378112716302742>.
- [9] D. Wear and J. Greis, "The Southern Forest Futures Project: Summary Report," Tech. Rep., 2011. [Online]. Available: <http://www.srs.fs.usda.gov/futures/>.
- [10] C. O. Sabatia and H. E. Burkhart, "Predicting site index of plantation loblolly pine from biophysical variables," *Forest Ecology and Management*, vol. 326, pp. 142–156, 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378112714002588>.

- [11] C. Bryars, C. Maier, D. Zhao, M. Kane, B. Borders, R. Will, and R. Teskey, "Fixed physiological parameters in the 3-PG model produced accurate estimates of loblolly pine growth on sites in different geographic regions," *Forest Ecology and Management*, vol. 289, pp. 501–514, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378112712005737>.
- [12] C. A. Carlson, T. R. Fox, J. Creighton, P. M. Dougherty, and J. R. Johnson, "Nine-year growth responses to planting density manipulation and repeated early fertilization in a loblolly pine stand in the Virginia Piedmont," *Southern Journal of Applied Forestry*, vol. 33, no. 3, pp. 109–114, 2009. [Online]. Available: <http://www.ingentaconnect.com/content/saf/sjaf/2009/00000033/00000003/art00001>.
- [13] E. J. Jokela, P. M. Dougherty, and T. A. Martin, "Production dynamics of intensively managed loblolly pine stands in the southern United States: A synthesis of seven long-term experiments," *Forest Ecology and Management*, vol. 192, no. 1, pp. 117–130, 2004. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378112704000180>.
- [14] J. H. Miller, B. R. Zutter, S. M. Zedaker, M. B. Edwards, and R. A. Newbold, "Growth and yield relative to competition for loblolly pine plantations to midrotation—a southeastern United States regional study," *Southern Journal of Applied Forestry*, vol. 27, no. 4, pp. 237–252, 2003. [Online]. Available: <http://www.ingentaconnect.com/content/saf/sjaf/2003/00000027/00000004/art00002>.
- [15] D. Gatzliolis, "LiDAR-derived site index in the US Pacific Northwest—challenges and opportunities," 2007. [Online]. Available: <http://www.treesearch.fs.fed.us/pubs/33211>.
- [16] M. A. Wulder, J. C. White, G. Stinson, T. Hilker, W. A. Kurz, N. C. Coops, B. St-Onge, and J. T. Trofymow, "Implications of differing input data sources and approaches upon forest carbon stock estimation," *Environmental monitoring and assessment*, vol. 166, no. 1-4, pp. 543–561, 2010. [Online]. Available: <http://link.springer.com/article/10.1007/s10661-009-1022-6>.
- [17] P. Tompalski, N. C. Coops, J. C. White, M. A. Wulder, and P. D. Pickell, "Estimating Forest Site Productivity Using Airborne Laser Scanning Data and Landsat Time Series," *Canadian Journal of Remote Sensing*, vol. 41, no. 3, pp. 232–245, 2015. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/07038992.2015.1068686>.
- [18] R. Gopalakrishnan, V. A. Thomas, J. W. Coulston, and R. H. Wynne, "Prediction of Canopy Heights over a Large Region Using Heterogeneous Lidar Datasets: Efficacy and Challenges," *Remote Sensing*, vol. 7, no. 9, pp. 11 036–11 060, 2015. [Online]. Available: <http://www.mdpi.com/2072-4292/7/9/11036>.
- [19] J. S. Kauffman and S. P. Prisley, "Automated estimation of forest stand age using Vegetation Change Tracker and machine learning," *Mathematical and Computational Forestry & Natural Resource Sciences*, vol. 8, no. 1, p. 4, 2016. [Online]. Available: <http://search.proquest.com/openview/82cdfdb6e400aae13d34aef72dad5e85/1?pq-origsite=gscholar&cbl=2030384>.
- [20] M. Fagan, D. Morton, B. Cook, J. Masek, F. Zhao, R. Nelson, and C. Huang, "Beyond edge effects: Landscape controls on forest structure in the southeastern US," San Fransisco, Dec. 2016.
- [21] M. Fagan, D. Morton, B. Cook, J. Masek, F. Zhao, H. Andersen, R. Nelson, and C. Huang, "Mapping pine tree plantations in the southeastern US using structural, spectral, and temporal remote sensing data," (*to be submitted*), 2017.
- [22] L. J. Sugarbaker, E. W. Constance, H. K. Heidemann, A. L. Jason, S. D. Lukas Vicki, and J. M. Stoker, "The 3d elevation program initiative—A call for action," *US Geological Survey Circular*, vol. 1399, p. 35, 2014. [Online]. Available: <http://pubs.usgs.gov/circ/1399/>.
- [23] C. Huang, S. N. Goward, J. G. Masek, N. Thomas, Z. Zhu, and J. E. Vogelmann, "An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks," *Remote Sensing of Environment*, vol. 114, no. 1, pp. 183–198, 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425709002685>.

- [24] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend, "High-Resolution Global Maps of 21st-Century Forest Cover Change," en, *Science*, vol. 342, no. 6160, pp. 850–853, Nov. 2013, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1244693. [Online]. Available: <http://www.sciencemag.org/content/342/6160/850>.
- [25] W. A. Bechtold and P. L. Patterson, *The enhanced forest inventory and analysis program: National sampling design and estimation procedures*. US Department of Agriculture Forest Service, Southern Research Station, 2005. [Online]. Available: http://www.srs.fs.usda.gov/pubs/gtr/gtr_srs080/gtr_srs080.
- [26] R. E. McRoberts, "The effects of rectification and Global Positioning System errors on satellite image-based estimates of forest area," *Remote Sensing of Environment*, vol. 114, no. 8, pp. 1710–1717, 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425710000854>.
- [27] R. P. Schultz, "Loblolly pine: The ecology and culture of loblolly pine (*Pinus taeda* L.)," 1997. [Online]. Available: <http://www.treesearch.fs.fed.us/pubs/734>.
- [28] U. Diéguez-Aranda, H. E. Burkhart, and R. L. Amateis, "Dynamic site model for loblolly pine (*Pinus taeda* L.) plantations in the United States," *Forest Science*, vol. 52, no. 3, pp. 262–272, 2006. [Online]. Available: <http://www.ingentaconnect.com/content/saf/fs/2006/00000052/00000003/art00006>.
- [29] D. B. Lobell, G. P. Asner, J. I. Ortiz-Monasterio, and T. L. Benning, "Remote sensing of regional crop production in the Yaqui Valley, Mexico: Estimates and uncertainties," *Agriculture, Ecosystems & Environment*, vol. 94, no. 2, pp. 205–220, 2003. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S016788090200021X>.
- [30] V. F. Quirino, "Evaluating the Potential for Estimating Age of Even-aged Loblolly Pine Stands Using Active and Passive Remote Sensing Data," 2014. [Online]. Available: <https://vtechworks.lib.vt.edu/handle/10919/71309>.
- [31] P. R. Seaber, F. P. Kapinos, and G. L. Knapp, *Hydrologic unit maps*. US Government Printing Office, 1987. [Online]. Available: <http://pubs.usgs.gov/wsp/wsp2294/>.
- [32] J. M. Omernik, "Ecoregions of the conterminous United States," *Annals of the Association of American geographers*, vol. 77, no. 1, pp. 118–125, 1987. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.1987.tb00149.x>.
- [33] R. T. Belote, S. Prisley, R. H. Jones, M. Fitzpatrick, and K. de Beurs, "Forest productivity and tree diversity relationships depend on ecological context within mid-Atlantic and Appalachian forests (USA)," *Forest Ecology and Management*, vol. 261, no. 7, pp. 1315–1324, Apr. 2011, ISSN: 0378-1127. DOI: 10.1016/j.foreco.2011.01.010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378112711000235>.
- [34] A. Noormets, M. J. Gavazzi, S. G. McNulty, J.-C. DOMEQ, G. E. Sun, J. S. King, and J. Chen, "Response of carbon fluxes to drought in a coastal plain loblolly pine forest," *Global Change Biology*, vol. 16, no. 1, pp. 272–287, 2010. [Online]. Available: <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2009.01928.x/full>.
- [35] U. NRCS, *National soil survey handbook*. 2007. [Online]. Available: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054242.
- [36] —, "Land resource regions and major land resource areas of the United States, the Caribbean, and the Pacific Basin," *US Department of Agriculture Handbook*, vol. 296, 2006.
- [37] T. R. Fox, E. J. Jokela, and H. L. Allen, "The development of pine plantation silviculture in the southern United States," *Journal of Forestry*, vol. 105, no. 7, pp. 337–347, 2007.

- [38] R. J. Norby and D. R. Zak, "Ecological lessons from free-air CO₂ enrichment (FACE) experiments," *Annual review of ecology, evolution, and systematics*, vol. 42, no. 1, p. 181, 2011. [Online]. Available: <http://www.annualreviews.org/eprint/wJxDadSqMz2VupKBxZay/full/10.1146/annurev-ecolsys-102209-144647>.
- [39] R. Murthy, P. M. Dougherty, S. J. Zarnoch, and H. L. Allen, "Effects of carbon dioxide, fertilization, and irrigation on photosynthetic capacity of loblolly pine trees," *Tree Physiology*, vol. 16, no. 6, pp. 537–546, 1996. [Online]. Available: <http://treephys.oxfordjournals.org/content/16/6/537.short>.
- [40] W. H. Carmean, J. T. Hahn, and R. D. Jacobs, *Site index curves for forest tree species in the eastern United States*. 1989. [Online]. Available: <https://www.treesearch.fs.fed.us/pubs/10192>.
- [41] E. Naesset, "Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data," *Remote Sensing of Environment*, vol. 80, no. 1, pp. 88–99, 2002. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425701002905>.
- [42] D. Riano, E. Meier, B. Allgower, E. Chuvieco, and S. L. Ustin, "Modeling airborne laser scanning data for the spatial generation of critical forest parameters in fire behavior modeling," *Remote Sensing of Environment*, vol. 86, no. 2, pp. 177–186, 2003.
- [43] C. Hopkinson, L. Chasmer, K. Lim, P. Treitz, and I. Creed, "Towards a universal lidar canopy height indicator," *Canadian Journal of Remote Sensing*, vol. 32, no. 2, pp. 139–152, 2006. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.5589/m06-006>.
- [44] S. Subedi, T. R. Fox, and R. H. Wynne, "Determination of fertility rating (FR) in the 3-PG model for loblolly pine plantations in the Southeastern United States based on site index," *Forests*, vol. 6, no. 9, pp. 3002–3027, 2015. [Online]. Available: <http://www.mdpi.com/1999-4907/6/9/3002/htm>.
- [45] C. Véga and B. St-Onge, "Mapping site index and age by linking a time series of canopy height models with growth curves," *Forest Ecology and Management*, vol. 257, no. 3, pp. 951–959, 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378112708008025>.
- [46] T. R. Fox, K. H. Kyle, L. J. Andrews, W. M. Aust, J. A. Burger, and G. H. Hansen, "Long-term effects of drainage, bedding, and fertilization on growth of loblolly pine (*Pinus taeda* L.) in the Coastal Plain of Virginia," *Southern Journal of Applied Forestry*, vol. 29, no. 4, pp. 205–214, 2005. [Online]. Available: <http://www.ingentaconnect.com/content/saf/sjaf/2005/00000029/00000004/art00006>.
- [47] T. J. Albaugh, T. R. Fox, H. L. Allen, and R. A. Rubilar, "Juvenile Southern Pine Response to Fertilization Is Influenced by Soil Drainage and Texture," *Forests*, vol. 6, no. 8, pp. 2799–2819, 2015. [Online]. Available: <http://www.mdpi.com/1999-4907/6/8/2799/htm>.
- [48] G. R. Corral and H. E. Burkhardt, "Quantifying and mapping spatial variability in simulated forest plots," 2016. [Online]. Available: <http://www.treesearch.fs.fed.us/pubs/50612>.
- [49] B. C. Larson and C. D. Oliver, *Forest stand dynamics*. John Wiley & Sons, 1996.
- [50] M. G. R. Cannell, "Physiological basis of wood production: A review," *Scandinavian Journal of Forest Research*, vol. 4, no. 1-4, pp. 459–490, 1989. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/02827588909382582>.

Chapter 5

Conclusions

In this work, I have outlined three avenues by which a large area discrete return lidar dataset could be used to generate forest cover related products over Southeast USA. In chapter two, I have outlined how a large-area, disparate-lidar-dataset canopy height model could be parametrized by using field data from a standardized and robust national inventory system (i.e., the FIA). I showed that a fair model (accuracy = 62%, kappa = 0.23) for predicting understory presence over varying forest types is possible in chapter 3. Moreover, I proposed several metrics as good correlates of such understory presence. In chapter 4, I extend on the work of chapter 1. That is, we combined a canopy height model (based on our findings in chapter one) with a Landsat-based stand age model to estimate site index over large areas of planted pine the southeast. The results highlight the power of large-area lidar collected for mostly non-forestry purposes in helping to monitor forest productivity levels.

5.1 Future work

I identified several avenues to extend the present work:

1. An effort should be undertaken to generate forest canopy height maps over the entire area of lidar coverage, keeping in mind that caveats outlined in chapter 2. Such maps are especially useful over natural forest areas (non-plantation) to serve as baselines (for future change analysis) and for better land use classification, species classification, forest fuel load modeling and forest ecology studies.
2. More work needs to be done to establish better methods to map understory over large areas. Improvements in field data collection protocols (larger sampling plots, more plots in forested areas, good number of plots in each forest type, improving the repeatability of measurements) and lidar data (sensors with less dead-time between returns, acquisitions designed to affect more returns from the understory layer) should be tried out.
3. The importance of higher point density in affecting greater sampling of understory presence has been mentioned by several previous articles (see chapter 3). Hence, a systematic controlled study should be undertaken to assess the effect of point density. Higher point density in the understory can be achieved by choosing a sensor with high pulse frequency, keeping the flying altitude low and keeping the scanning angle low (typically less than 15 degrees).
4. The forest stand age product used in chapter 4 (eVCT) seems to have the potential to be improved, so that the uncertainty of site index estimates are reduced. For example, a strong relationship (R^2

= 77.5%) could be developed for plots with age estimates having better agreement. Initial analysis points to some issues in the underlying VCT algorithm (for example, over-detection of disturbances at the beginning of the time stack, i.e., the year 1984). More work is needed to first establish a set of plots (FIA or otherwise) with associated robust stand ages spread over the entire Southeast USA, and over all possible management regimes to better quantify the accuracy of the eVCT stand age product. More work is also needed to partition the error between the underlying disturbance detection algorithm (VCT) and the algorithm that classifies them as stand-clearing (i.e., a clearcut) or not.

5. The site index map should be extended over a larger area so that true region-wide baselining and analysis can be achieved. It is important to produce robust estimates of current (as of 2010s) site index so as to better monitor for trends related to global changes in land management, atmospheric carbon dioxide concentrations and climate.

The work outlined in this dissertation showcases the effectiveness of combining lidar data from disparate projects along with national level forest inventory data to generate forest parameter maps over large areas. These maps would be helpful for a variety of applications such as wildfire risk assessment, carbon accounting (related to climate change), aiding in better forest type classification and forest biodiversity management. Ideally, the models developed in this work for canopy height and understory presence could be applied to future lidar acquisitions without substantial field calibration work, as numerous field plots over a large variety of lidar acquisition types, forest types and physiographic conditions were used for their development. This work thus adds value to large-area lidar collections to address critical environmental and ecological issues facing us in the 21st century.

Appendix A: Code and data repository

VTechData (<https://data.lib.vt.edu/>) is a portal that helps publicize and preserve the scholarly work of Virginia Tech faculty, students, and staff. Significant code and data developed as part of this dissertation can be found here: <https://data.lib.vt.edu/collections/765371353>. The associated digital object identifier (doi) is 10.7294/W41Z429P. The relevant policies for usage of this code and data are available at <https://data.lib.vt.edu/about/>.