LiDAR-Assisted Extraction of Old Growth Baldcypress Stands Along The Black River of North Carolina

Weston Pierce Murch
University of Arkansas, Fayetteville

Follow this and additional works at: http://scholarworks.uark.edu/etd

Part of the Geographic Information Sciences Commons, Natural Resources and Conservation Commons, and the Physical and Environmental Geography Commons

Recommended Citation
http://scholarworks.uark.edu/etd/1628

This Thesis is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, ccmiddle@uark.edu.
LiDAR-Assisted Extraction of Old Growth Baldcypress Stands Along
The Black River of North Carolina

A thesis submitted in partial fulfillment
Of the requirements for the degree of
Master of Science in Geography

By

Weston Pierce Murch
Brigham Young University
Bachelor of Science in Geography, 2014

August 2016
University of Arkansas

This thesis is approved for recommendation to the Graduate Council

Dr. Jason A. Tullis
Thesis Director

Dr. David Stahle
Thesis Committee Member

Dr. Jackson Cothren
Thesis Committee Member
Abstract

The remnants of ancient baldcypress forests continue to grow across the Southeastern United States. These long lived trees are invaluable for biodiversity along riverine ecosystems, provide habitat to a myriad of animal species, and augment the proxy climate record for North America. While extensive logging of the areas along the Black River in North Carolina has mostly decimated ancient forests of many species including the baldcypress, conservation efforts from The Nature Conservancy and other partners are under way. In order to more efficiently find and study these enduring stands of baldcypress, some of which are estimated to be more than 1,000 years old, LiDAR remote sensing and geospatial analysis techniques can be employed. Promising results have been discovered correlating LiDAR-derived metrics and known stands of old growth baldcypress. A number of percentile height metrics and other composite metrics like canopy cover and density were extracted from LiDAR data collected across North Carolina. Along with the metrics, locations of known stands of old growth were used as training data for a supervised classification with the C5.0 decision tree algorithm. C5.0 was used to condense the patterns found across the training data into a set of rules that could then be applied to other areas within the study site or anywhere else across the LiDAR data. Both existing stands and new areas were selected by the machine learning rulesets indicating that the use of machine learning is valid to identify stands of ancient trees along the Black River. Overall C5.0 accuracies of approximately 98.5% (based on training data) and 88.6% (based on independent test data) were achieved. More than 8 km² of predicted old growth forests, outside of available in situ reference areas, were also identified within the Black River site.
Acknowledgments

I would not have been able to complete this thesis without the tireless efforts of my wife and family. Many late nights were spent away from my wife and son while my parents-in-law watched and supported both of them. Thank you, Lynn and Gretchen. My wife’s sacrifice of time and comfort and peace of mind while I completed this thesis has been one of the greatest displays of love I have ever seen. She never gave up on me, even when I wanted to give up myself. My father making a special trip to support and keep me on track is also deeply appreciated. A special thank you goes to him. All of my family’s support and constant encouragement will forever be treasured. My committee, Dr. Jason Tullis, Dr. Dave Stahle, and Dr. Jack Cothren, have also all contributed to the culmination of this study. Special thanks go to my advisor, Dr. Tullis, whose patience and advice have been so helpful. The assistantship provided by the Geosciences department here at the University of Arkansas has been so valued and accommodating as I furthered my education in preparation for a bright career. Thank you all!
# Table of Contents

1. Introduction and Background .............................................................................................................. 1
   1.1 Baldcypress Characteristics ................................................................................................. 3
   1.2 Baldcypress Conservation .................................................................................................... 4
   1.3 LiDAR Remote Sensing ...................................................................................................... 4
   1.4 Hypothesis .......................................................................................................................... 8

2. Literature Review ............................................................................................................................... 8
   2.1 LiDAR and Forestry Studies ............................................................................................ 8
   2.2 Machine Learning ............................................................................................................. 11

3. Methodology ........................................................................................................................................ 15
   3.1 Study Site ........................................................................................................................ 15
   3.2 Geodata Collection .......................................................................................................... 17
      3.2.1 In Situ Data .......................................................................................................... 17
      3.2.2 Small-footprint LIDAR ........................................................................................ 20
      3.2.3 National Agricultural Imagery Program (NAIP) ................................................ 28
   3.3 Geodata Processing ............................................................................................................ 28
      3.3.1 Extraction of LiDAR Metrics Using R ................................................................ 30
      3.3.2 LAScanopy and ArcGIS Approach ........................................................................ 35

4. Results and Discussion ....................................................................................................................... 38

5. Conclusion ........................................................................................................................................ 49

6. Bibliography ..................................................................................................................................... 51

7. Appendix ......................................................................................................................................... 57
   7.1.1 Final Ruleset based on LAStools metrics and fine-tuned training data ............... 57
    7.2 Ancillary Tables ................................................................................................................. 59
    7.3 Scripts and Workflow models ......................................................................................... 62
      7.3.1 Final Workflow Model ............................................................................................. 62
      7.3.2 Raster to See5 format data Script ............................................................................ 64
      7.3.3 Ruleset to Selection for ArcGIS ................................................................................. 67
1. Introduction and Background

This study is an endeavor to use light detection and ranging (LiDAR) combined with machine learning and geographic information system (GIS) techniques to automatically identify previously unknown stands of potential old growth baldcypress. This work builds upon previous efforts to identify old growth baldcypress stands along the Black River, North Carolina using a combination of in situ observation and manual interpretation of aerial photography. This project leverages a relatively new LiDAR dataset that was captured for the Black River area by the North Carolina Floodplain Mapping Program to create detailed elevation/hydrology models, and aerial imagery readily available from a variety of public domain sources. The approach of LiDAR-assisted old growth characterization is desirable because of its potential applicability in identifying other areas of ancient remnant forests across the Southeast United States and beyond. The machine analysis of LiDAR data, combined with manual photointerpretation and in situ data collection, can facilitate and accelerate the surveying of larger forest areas that may be difficult or even impractical to survey in situ. LiDAR remote sensing also provides a systematic way to quantitatively characterize the canopy structure of old growth baldcypress forest, which can contribute to the study of habitat quality for various species that rely upon these ancient biological treasures.

The old growth forests of the Eastern United States have been severely decimated by the long unchecked logging for commercial timber and creation of farm lands. The early European settlers cut down and burned much of the previously untouched forests for colonization of North America. Much old growth timber that wasn’t already gone by the 1800’s was quickly harvested by industrialists, intent on profiting from the vast forests of the entire Eastern US. The federal government formed the US Forest Service in 1905 to protect the forests that were left and to
improve the approach by which forest resources were handled. However, even this federal agency cracked under pressure from the commercial powers of the timber industry by allowing unchecked logging to continue (Davis, 1996). The past 300 years has seen extensive deforestation along the Eastern seaboard, from New York to Georgia. Along with the loss of the old growth trees themselves numerous habitat areas for sensitive species of birds and other forest animals have been largely fragmented. This diminishes the likelihood of these species’ survival (Smith et al., 1993). The trees left behind from extensive deforestation are not commercially valuable and very non-traditional in terms of physical characteristics, and this may help explain why they were passed over to begin with. Regardless, these trees constitute some of the oldest and best remaining specimens of ancient forests.

Much value old growth forest is in the biodiversity, environmental and ecological trove of information that they contain. The dendrochronologies of the ancient trees, especially from a dominant species known as baldcypress (*Taxodium distichum*), has been previously used to produce proxy climate data over North Carolina (Stahle et al., 1988) and to reconstruct spatial and temporal patterns of moisture variability across North America (Cook et al., 2007). The significance of finding baldcypress along the Black River, and in the future among other areas of the United States is that it is the longest-lived tree species along the Eastern sea board and among the oldest in the Eastern United States. This longevity provides the potential for extending the existing dendrochronologies a few hundred years past the oldest sampled tree at 1,652 years (Stahle et al., 1988).

An atmospheric benefit of conserving old growth trees is due to their capacity for carbon sequestration. A recent study by Stephenson et al. (2014) that spans multiple species and continents has concluded that aboveground tree mass growth rates in old growth forests actually
accelerate over time. Carbon accumulation in tree species benefits the environment by helping to mitigate carbon emissions and greenhouse gases. Trees experiencing a diameter increase of 10 times the original size will likewise experience between 50 and 100 times the original leaf area (Stephenson et al., 2014). This indicates that a stand of newly planted trees, over the same amount of time, say 60 years, would not amass as much carbon from the atmosphere as a stand of the same area and species of already mature trees. This finding is not dependent on the level of competition within stands or the species of tree. Most trees continue to grow indefinitely and accumulate or sequester carbon at an accelerating rate despite inefficiencies that many old tree stands experience.

1.1 Baldcypress Characteristics

The species *Taxodium distichum*, commonly known as baldcypress, is a deciduous conifer. Baldcypress’ characteristic of having cones and needles but senescing in the fall and losing needles annually is not common for coniferous trees. Baldcypress trees have been known to attain a height of up to 45 meters and have trunk diameters at breast height up to 3.5 meters (Wilhite and Toliver, 1990). At 200 years old, baldcypress height growth peaks, and the proportion of heartwood is sufficiently high for harvesting (Wilhite and Toliver, 1990).

Mature baldcypress trees transition from the pyramidal shaped crown more characteristic of youthful trees, to a more flat-topped crown. As the trees get older these crowns become heavier and more asymmetrical. This is due to environmental (e.g., storm) damage and heart rot that affects many of the known remaining old baldcypress stands (Stahle et al., 2012). Baldcypress trees, especially in environments like the Black River that are of interest to this study, grow in riverine swamps and flood plains or directly in the river channel itself. As stated by Wilhite and Toliver (1990), “[d]rainage, therefore, may be more important than rainfall in determining site
suitability for baldcypress”. Nutrient poor, flooded environment at elevations below 30 meters above sea level harbors as much as 90 percent of the baldcypress stands. The climates most common to the environments where baldcypress are found are humid, moist sub-humid and dry sub-humid. Warm climates receiving between 31 and 44 inches of annual rainfall and intermittent flooding tend to be the best environment for baldcypress growth (Wilhite and Toliver 1990).

1.2 Baldcypress Conservation

Conservation efforts targeted at protecting baldcypress are headed by The Nature Conservancy (TNC). In the area of the Black River, much land is privately owned by corporations like Boise Cascade, a wood products company (TNC, 2016). The Nature conservancy has been acquiring land and easements for almost 20 years including the 1092 hectare (2,700 acre) Roan Island, and another 2,200 acres in the Black River preserve. TNC also manages easements up and down both the Black River and the Cape Fear River. A number of partners like local land owners and the Boise Cascade Corporation also work with the Nature Conservancy to protect sensitive lands and species. These entities collectively control more than 14,500 acres that are designated for protection (The Nature Conservancy, 2016).

1.3 LiDAR Remote Sensing

Used as early as 1963 for cloud measurements (Goyer, 1963), light detecting and ranging (LiDAR) now provides many opportunities to add additional information to that gained through in situ or traditional forestry methods of inventory. Based on its frequency within both the popular and scientific remote sensing literature, LiDAR has become one of the foremost technologies in remote sensing. The LiDAR investigated in this study works by sending a pulse
of laser energy away from a sensor and measuring the time it takes to record a return from the reflected pulse. The LiDAR data used in this study is small footprint aerial LiDAR. Equipment for this type of LiDAR remote sensing can cost hundreds of thousands of dollars and are often rented or contracted to minimize costs. Small footprint LiDAR sensors uses a laser tuned to a single precise wavelength typically in the green (532 nm) or near-infrared (exactly double at 1,064 nm) electromagnetic spectrum to produce the energy pulse. The specific wavelength chosen depends on the target and its ability to reflect or absorb that wavelength. In forestry and other land applications, the near-infrared wavelength (1,064) is typical and is also safer in terms of potential impact on the human eye. The footprints of these sensors, where the pulse hits the target, are between 20 and 40 cm in diameter depending on the height of the sensor above the ground or targets such as trees or buildings (Thenkabail, 2015).

A given LiDAR dataset, or “point cloud” as it is frequently termed, typically contains a large number of \(x, y, z\) (3D) coordinates produced which can be quite large and dense. The nominal post spacing (the horizontal spacing between the points) depends on the specific sensor system and its configuration for the desired application. A variety of distinct LiDAR sensors exist and have been adapted to specific areas of study varying from space exploration (Ring, 1963) to fish population surveys (Pittman & Brown, 2011). As applied to forestry, small footprint LiDAR can be used to extract very detailed digital terrain models (DTMs), models of tree/canopy height, crown cover, vertical and horizontal forest structure, and a multitude of associated metrics. Post-processing of the LiDAR data can produce results for volume, biomass, density and foliage projected cover. With a high enough density LiDAR point cloud, metrics at even the individual tree level can be attained (Kwak et al., 2007).
Airborne LiDAR Systems (ALS) that operate from an aerial platform (e.g., fixed wing aircraft, helicopter, or unmanned aerial vehicle) are increasingly common, and depending on their configuration, may be suited for collection over entire states or small countries. Common aerial LiDAR systems produce 30 cm nominal post spacing data from small footprint pulses (Thenkabail, 2015). Terrestrial LiDAR systems (TLS) can generate better resolutions of understory canopies and ground points by conducting multiple scans at the same location (Omasa et al, 2007). TLS is able to capture metrics such as basal area and other features that are more difficult to gather from aerial LiDAR. Point clouds can be processed into 2D and 3D models showing a variety of surface details including vegetation, rocks, ground, and water. Other attributes in addition to bare 3D coordinates can be captured with each point such as LiDAR return intensity, scan direction, and return number (to be explained below).

Common LiDAR systems use a linear (i.e. zig-zag, or parallel) scan pattern and capture either a single or a finite number of discrete returns generated from each original laser pulse. Each laser pulse emitted covers a footprint that the electromagnetic energy is evenly distributed across. Multiple objects such as trees, buildings, birds and bare ground can reflect a portion or the entire laser footprint. When more than one object reflects a portion of the laser footprint more than one return is measured for that single pulse. Hence, multiple returns are possible from a single laser pulse – an important feature for old growth forest identification (to be discussed further in this study). Where vegetation is present, often the first return is reflected from the top of the canopy and the last return is reflected from the ground or from very low ground cover. Intermediate returns can also be reflected from parts of the inner canopy, trunk, and vegetation making unique vertical distributions of point clouds.
Several variations of LiDAR technology exist, each with its own strengths and weaknesses. Geiger mode LiDAR by Harris Geospatial (2016) has recently been developed that is capable of capturing very high point densities and have the advantage of illuminating a given location or its vicinity from multiple angles. This works by using lower powered sensors than linear arrays that cover a much wider angle and can send multiple pulses at once. This LiDAR is only used currently by government agencies and contracted solely by Harris Geospatial. Full waveform (FWF) systems have a much larger laser footprint than small footprint LiDAR (and therefore may not be suitable capturing some spatial details). However, they capture a continuous stream of return information (as opposed to a few discrete returns) as the reflected pulse returns to the sensor. The FWF LiDAR systems may be particularly useful to help distinguish between diverse materials (Lin, 2015), and while not as common, may be applicable to forestry metrics and tree size (Reitberger et al., 2009). Multispectral LiDAR is available to some degree and has produced valuable results. Some benefits of multi-wavelength LiDAR are in estimating biomass and detecting plant stresses (Gong et al., 2012). This type of sensor allows for a combination of traditional spectral analysis and 3D point cloud generation of LiDAR; however, the intensity of the spectral return is subject to not only the reflectance of the target material, but also the quantity of photons returning to the sensor.

As has been pointed out, LiDAR systems can generate huge volumes of point cloud data. For example, the dataset used in this project covered 2,553 km², with 11 billion points at an average point density of 4.3 points/m² and has a storage size on disk of 349 GiB. As the pulse rate of LiDAR systems continues to increase, this will continue to affect data volumes. This trend can be a real concern for many studies as storage and processing resources may be costly or limited (including the costs for geospatial experts to work with this data). Development of automated
techniques for processing the LiDAR, which is the focus of this project, is therefore of particular relevance in old growth forest conservation.

1.4 Hypothesis

The following hypothesis was addressed in this study: “There is no significant relationship between in situ measurements of baldcypress stand age and small-footprint LiDAR-derived metrics.” To address this hypothesis, a geospatial workflow was developed to 1) extract metrics from the LiDAR, 2) develop machine learning decision trees for predicting whether a site may contain old growth forest, and 3) independently validate the quality of the resulting predictions.

2. Literature Review

2.1 LiDAR and Forestry Studies

The most common LiDAR study techniques used for forestry metrics, species identification, and tree detection are based on canopy analysis. Existing aerial LiDAR systems have inherent limitations where the surfaces of the canopies get the most pulse returns but underneath the canopies the number of returns is insufficiently representative of the understory structure. For this reason, observing and measuring the canopy surface has been the preferred method for studies of tree stand height metrics and canopy structure. Altering point densities by various practices such as changing the sensor height and scan angle have noticeable effects on resolving trees (Goodwin et al., 2006).

Studies like Takashi et al. (2005) found successful identification of tree heights even on steep slopes and rugged topographies when point counts were at least 8.8 points/m². Plant studies that use LiDAR are plentiful and range in scope from single weed identification (Weiss et al., 2010) to tropical forests heights and biomass (Dubayah et al., 2010). Airborne LiDAR used to measure
tree height and related metrics (e.g., Suarez et al., 2004; Riggins et al., 2009; Kim, 2007) are key examples that provide a basis of this project in discovering relationships between old growth stands of the Black River and the LiDAR point cloud metrics that can be derived from aerial LiDAR scans.

Once the LiDAR data is collected and stored on a computer system, there are many possible tools and approaches to extract meaningful information from the point cloud. Lefsky et al. (1999) studied the implementation of FWF LiDAR in predicting basal area and aboveground biomass in deciduous trees along the coastal plains of Annapolis, MD. Such measurements as basal area, aboveground biomass (Pederson, 2010), and crown height (Kalliovirta et al., 2005) are common metrics used to calculate tree age based on known species characteristics. Canopy height profiles were made with LiDAR measurements and correlated against in-field measurements of numerous plots of trees. The indices determined between the LiDAR and field measurements were maximum canopy height, mean canopy height, median canopy height, and quadratic mean canopy height. Maximum and mean canopy heights were also included in the current study as metrics to glean from the LiDAR point clouds. While the LiDAR data from Lefsky’s (1999) study was much more inclusive, the results indicated that high correlations were found between the height indices and tree stand structure based on regression analysis.

Popescu and Zhao (2008) used discrete return LiDAR to determine similar metrics of trees that were used to determine individual crown height models (CHM) and vertical height structures. These vertical structures were derived by height bins, grouping points by range of height and looking at distribution of points within the bins, for characterizing canopy cover and leaf area index. These measurements from the LiDAR data were tested against ground data for tree location and how accurately the crown height and other biophysical. Popescu’s (2008)
approach to measuring attributes of trees with binned LiDAR returns and CHM performed at high accuracy for both pine trees and deciduous hardwoods. The indication that measuring percentile heights or binning LiDAR returns vertically through the canopy can distinguish useful characteristics of trees validates the addition of percentile heights to the current study in characterizing old growth baldcypress areas.

Another study using canopy structure derived from LiDAR states that, “canopy structure is a fundamental property of forest ecosystems that strongly influence their characteristics” (Hansen, 2014) The structure of tree canopies are defined by canopy height, total canopy cover, and the distribution of canopy cover across the forest (Franklin and Van Pelt, 2004). Undisturbed forests like the ideal areas along the Black River according to Hansen (2014) display canopy heights up to 30% more and canopy cover up to 15% more than disturbed forest areas in the same region. The use of LiDAR to capture these metrics was instrumental in characterizing trees and determining stand size, as well as estimating stand age within around 10 years along the Appalachian Mountains for deciduous hardwood forests (Hansen 2014).

Canopy coverage as a metric for estimating tree age is a viable measurement that is available with the data covering the study site in North Carolina. The evidence that metrics derived from LiDAR have been used successfully to determine forest characteristics useful for estimating species type, biomass and other important forest inventory indicators gives the basis for the metrics used here. With a combination of the metrics used in these referenced studies such as percentile heights, mean height, canopy cover, vertical point distributions, this study on LiDAR assisted old growth prediction was prepared for a machine learning process to discover patterns in the metrics that applied to old growth stands of bald cypress.
2.2 *Machine Learning*

Machine learning is a process of artificial intelligence dating to the 1950’s that works by various methods to capture knowledge and construct predictive algorithms based on training data. Knowledge-based systems described in Quinlan’s *Induction of Decision Trees* (1986), introduce the methods employed by the C5.0 algorithm (Rulequest, 2015) utilized in this project to develop rules for predicting old growth. Machine learning can take the place of simple statistical techniques in modelling the relationships between many variables in geospatial data (Young, 2013). Machine learning decision tree techniques utilize a learning tree to break down a series of decisions that progressively sort cases into bins for classification. The C5.0 algorithm used in this study to classify old growth is an application of a decision tree. Another type of machine learning called support vector machines (SVM), used in many of the reference studies on forestry and LiDAR, is also of interest because it is comparable to the C5.0 decision tree algorithm (Golmah, 2014; Pandya and Pandya, 2015) to discover relevant patterns within the LiDAR metrics. “A C5.0 model is based on the information theory. Decision trees are built by calculating the information gain ratio. The algorithm C5.0 works by separating the sample into subsamples based on the result of a test on the value of a single feature. The specific test is selected by an information theoretic heuristic”. With two dimensions or variables to assess characteristics, classification would be similar to the figure below. The goal of classifiers is to find the function that has the least difference between predicted and observed values of the target variable and at the same time be able to classify new data into the same classifications (Zhao et al., 2011).
Figure 1. Possible ways to classify between two classes shown by the green lines drawn. (OpenCV, 2016). Relevant variables can be identified (training the machine learning program with known cases of the target and non-target classes) and the remaining cases are processed for relationships with the relevant variables.
Machine learning is one possible category of techniques for recognizing patterns in metrics derived from remote sensing to identify target characteristics within a study area. For example, both Weiss et al. (2010) and Zhao et al. (2011) employ support vector machine learning methods to classify test data of different singular plants or large stands of mixed species trees. The study by Weiss et al. (2010) to determine crop species at the single plant level for weed identification utilized these techniques to analyze and “learn” what characteristics individual plants did exhibit that differentiated them from other plants. A moving terrestrial LiDAR sensor was used to scan individual plants and create point clouds that accurately depicted plant features including single leaves and stems. While Weiss’s (2010) study was a much smaller scale approach to LiDAR-based identification than this study, such techniques were successful. With 6 different types of plants, based on training data fed to machine learning, a 98% success rate for plant identification in laboratory conditions was achieved. Another study on characterizing forest canopy with
LiDAR and machine learning is very closely related to the work in this study. Zhao et al. (2011) describes the increased success of LiDAR/machine learning techniques over traditional classifiers like maximum likelihood and linear regression models. Zhao et al.’s (2011) study on characterizing canopy structure follows a very similar approach to the methods in this study, although they extracted more complex composite metrics such as Lorey’s height for stands, biomass and leaf area index. Zhao et al. (2011) reported a higher accuracy on determining tree species with SVM (83.06%) versus a traditional maximum likelihood classifier (82.27%). With reported success on a combination of LiDAR metrics and machine learning techniques using SVM it is sufficient to believe that acceptable results from a similar process within the Black River study site should be attainable. Therefore, a workflow consisting of LiDAR metrics and SVM machine learning algorithms was adopted to find projected old growth baldcypress.
3. Methodology

3.1 Study Site

The Black River is an 80 km tributary of the Cape Fear River in southeastern North Carolina. The specific study site is a portion of the Black River approximately 30 km long between the borders of Bladen and Pender Counties. The elevation at the study site is very low, with the river channel on average 3-5 m and the surrounding valley no higher than 10 m above sea level. This area is fairly remote and in many locations limited to boat access along the river. This inaccessibility is a major factor in the necessity for remote sensing techniques to be employed for baldcypress stand inventory. The winding river course, many small streams and back water swamps produce ideal conditions for baldcypress stands to grow. Individual tree core samples, used to assess tree rings for age estimation, as well as stands of known remnant old growth baldcypress and other areas expected to contain more old growth trees all fall within the study site.
Figure 3. The study site along the Black River in North Carolina, covering 19 LiDAR tiles highlighting the portion of the Black River (in blue) and other nearby rivers (in purple).
3.2 Geodata Collection

A variety of \textit{in situ}, ancillary, and remote sensor data collected were used for various purposes. These were incorporated in a number of ways to identify stands of baldcypress. A focus of the methodology was on how to utilize small-footprint LiDAR instead of traditional spectral remote sensing techniques for detection of old growth tree stands based on point distribution metrics. Visual photo interpretation of high spatial resolution aerial photography of the study site and surrounding area was recently performed by Burns (2015). The purpose of the LiDAR-assisted method developed in this research was to automatically identify areas characteristic of old growth stands using GIS so that new not previously recorded \textit{in situ} can be further investigated. All geodata utilized in the project was freely downloaded or donated for the purposes of this and future projects concerning the study area.

3.2.1 In Situ Data

A total of 68 cored baldcypress samples that range from 100 to 1,600 annual tree rings were collected in-situ. Stands of old growth were determined in the field by DBH and tree ring measurements along with visual inspection of characteristics like under-fit canopies, storm damage and heart rot. Data from Burns (2015) field collection efforts for individually cored trees (Table 1 in Appendix) verified areas known to be old growth baldcypress stands either previously or through Burns’ study. These verified stands of old growth baldcypress (see Figure 4) were incorporated for the training and validation of machine learning methods developed in this study. In that photographic interpretation can be done to reasonably identify possible old growth stands, these boundaries served as a starting point for classification of LiDAR-derived metrics. Without \textit{in situ} verification, it would be difficult to ascertain whether old growth is indeed present. Preliminary validation can be done on any additional areas of predicted old
growth characteristics based on proximity and visual similarities to the regions already identified.
Figure 4. Areas of reference old growth along the Black River determined by in-field studies from Burns (2015).
3.2.2 Small-footprint LiDAR

The focus of the study was on the information content in small-footprint LiDAR collected in 2014 over much of North Carolina. The aerial LiDAR data vendors used both Leica ALS-70HP-II and Optech Pegasus HA500 sensors in this process (Morgan, 2015). This project was a collaborative effort between the North Carolina Emergency Management, North Carolina Geodetic Survey, North Carolina Department of Transportation, the United States Marine Corps, the United States Geological Society (USGS), and the National Resources Conservation Service (NRCS). These partners commissioned the data to update existing data from 2000-2005, utilizing up-to-date sensor systems. The aim of the ongoing project is to produce up-to-date terrain models, hydro-flattened break lines and digital elevation models (Morgan, 2015). The date of LiDAR collection occurred at a specific time of year to account for leaf-off conditions where the ground would deliver the most returns consistent with the objective of the project of getting the best models of the ground and hydrology systems for flood plain mapping.

The LiDAR data for the area of interest (within the study site) were collected in December of 2014. All data was projected in North Carolina State Plane, horizontal datum NAD83 (2011; also known as WKID 2264), and a vertical datum of NAVD88 (Geoid 12a) (Morgan, 2015). Both systems use United States (US) survey feet as do the methods for processing and measuring the LiDAR data. The flight paths for collecting LiDAR are no longer than 50 km (31 mi) long and each swath of LiDAR data is approximately 1.55 km (5,100 ft) across. A large amount of overlap between flight paths was used to calibrate the point clouds (Morgan, 2015).

A total of 1,126 files collected in LAS version 1.3 standard format, commonly referred to as tiles, cover portions of Bladen, Pender, Columbus, Brunswick, and New Hanover counties (see Figure 9). The entire dataset consists of 11,037,642,823 points with an average point density of...
4.3 points/m². A total of 19 tiles from this larger dataset cover the study site where known old
growth baldcypress has been located. Each tile represents a 5,000 × 5,000 ft. (1,524 ×1,524 m)
extent (see Figures 5 and 6).
Figure 5. The 19 LiDAR tiles covering the study site, approximately 30 km of the Black River.
Figure 6. One LiDAR tile in the study site covering approximately 2.6 km² (1 mi²). Red represents the highest elevation points all the way through a color gradient to blue, the lowest elevations.

Approximately 27% of the LiDAR returns are in overlap areas and approximately 20% were classified as ground returns in non-overlapping areas. The remaining 53% of the returns are above ground returns, which are essential to this study. The classifications (see Table 1) for LiDAR returns were identified in the metadata for the North Carolina Floodplain mapping program (Morgan, 2015). The LiDAR point cloud contains sufficient spatial detail that individual trees are visible across the tiles (Figures 7 and 8).
Table 1. LiDAR Classifications as designated within LiDAR metadata.

<table>
<thead>
<tr>
<th>Code</th>
<th>Return Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unclassified</td>
</tr>
<tr>
<td>2</td>
<td>Ground</td>
</tr>
<tr>
<td>3</td>
<td>Low vegetation</td>
</tr>
<tr>
<td>4</td>
<td>Medium vegetation</td>
</tr>
<tr>
<td>5</td>
<td>High vegetation</td>
</tr>
<tr>
<td>6</td>
<td>Building</td>
</tr>
<tr>
<td>7</td>
<td>Low noise</td>
</tr>
<tr>
<td>9</td>
<td>Water point</td>
</tr>
<tr>
<td>10</td>
<td>Breakline proximity point</td>
</tr>
<tr>
<td>11</td>
<td>High noise</td>
</tr>
<tr>
<td>13</td>
<td>Road surface</td>
</tr>
<tr>
<td>14</td>
<td>Bridge points</td>
</tr>
<tr>
<td>17</td>
<td>Overlap default</td>
</tr>
<tr>
<td>18</td>
<td>Overlap ground</td>
</tr>
</tbody>
</table>
Figure 7. A profile view of a younger baldcypress on the left and an ancient baldcypress in the middle of the image. Maximum height of points above ground within in the image is 28.3 m.
Figure 8. Canopies can be seen for individual trees along with the general shape and understory growth.
Figure 9. LiDAR tiles covering the Black River study site (in red) and the entire LiDAR dataset from the North Carolina Floodplain mapping program (in blue).
3.2.3 **National Agricultural Imagery Program (NAIP)**

National Agricultural Imagery Program (NAIP) imagery with a nominal spatial resolution of \(1 \times 1 \) m and four spectral bands (blue, green, red, and near-infrared) was utilized for visual reference and for interpretation of the machine learning results. The imagery was collected between March and August of 2014 and was the most recent NAIP imagery available at the time of this study. The LiDAR data was collected in December of the same year which coincides well with the NAIP imagery. This small gap between collection of the NAIP and LiDAR helps ensure the least amount of disturbance of land within the study site. The near infrared band was especially helpful for referencing areas of senescent vegetation. This highlighted the baldcypress’ tendency to senesce and drops their needles earlier in the season (around early August) than other deciduous trees.

3.3 **Geodata Processing**

The geodata processing portion of the study began with consolidation and grouping of the data using Environmental Systems Research Institute’s (Esri’s) ArcGIS for Desktop products (ArcGIS Pro and ArcMap) referred to as ArcGIS in the rest of the methodology.

A valuable tool for organizing and editing the LAS data is the LAStools package from Martin Isenburg (2016). The default setup of LAStools installs a custom toolbox that can be added to the ArcGIS ModelBuilder graphic block programming interface which was used extensively in this work. “Boundary”, “canopy”, “ground”, “las2las”, and “height” were the principal LAStools tools used in this project. The boundary tool transforms the bounding boxes of LAS files into Esri Shapefiles, while las2las rewrites LAS files into other formats and can exclude LiDAR returns of certain classes. The ground tool uses last returns of the LiDAR to determine ground points and create ground surface models while the height tools works together
with ground to calculate the height above the ground surface for each return. The canopy tool worked on the height above ground result and was the most important to the methodology of this study. Common forest metrics that the tool is capable of producing such as canopy cover and percent heights were the essential input for the machine learning. The methods for calculating these metrics as performed by LAScanopy are explained in section 3.3.2.

A general workflow diagramed in Figure 10 shows the process of the LiDAR data conversion and processing for analysis. The exploration process began by combining the 19 raw LiDAR LAS files, of the 1,126 available that overlap the study site into a single LAS dataset (LASD). An LASD references binary LiDAR data in the form of LAS files and combines them into a readable, displayable data format for ArcGIS. Because LAS files can be very large (e.g., hundreds of GBs) only a portion of the data is displayed at any time. The tiled indexing approach of the LASD helps limit the graphics and processing strain on ArcGIS software. The LASD created covered the Black River basin between Bladen and Pender counties in North Carolina and included the known baldcypress old growth stands. The LASD was visually examined for estimation of possible characteristics that old growth stands exhibit. Possible characteristics were the vertical distribution of the points across the canopy, the density of the points in or below the canopy, and the shape of the canopies.

Though the LAS files were tagged internally as conforming to ASPRS LAS file format version 1.3 (ASPRS, 2011), though upon inspection it was discovered that some points were classified in a manner that deviated from the 1.3 standard. The most significant classification errors were points classified as 17 and 18. LAS standard 1.3 specifies these as “bridge deck” and “high noise”, respectively. The metadata provided by the data vendor indicates the points classified as 17 and 18 to be “overlap default” and “overlap ground”, respectively. The overlap
occurs as the sensor collects data in paths that cross each other. Multiple collections over the same area create denser point clouds within the overlapping areas. Since overlap points would increase the total point count for a cell, and total point count is a metric considered in the analysis, these points were eliminated from the LASD. Data points reported as noise by the project metadata, classes 11 (high noise), and 7 (low noise) were also removed. Noise occurs from errors in measuring the returns or extraneous objects like birds. The remaining data point classes were left intact as received in the original dataset.

![Generalized workflow design for transforming LiDAR into training data for C5.0 classification.](image)

Figure 10. Generalized workflow design for transforming LiDAR into training data for C5.0 classification.

### 3.3.1 Extraction of LiDAR Metrics Using R

The next step was to generate multiple single-layer rasters from the LiDAR tiles each containing a single metric value at each given cell. This was accomplished in two ways using
different software packages. One approach to create these LiDAR metric rasters incorporated the R statistics package and a modified script for hardwood biomass estimation adapted from Xue (2015). In supporting the R script requirements, the LAStools package was used to extract the tile boundaries for each tile and convert the LiDAR data files to comma-delimited text files. A modified version of the LASboundary tool was created to extract a shapefile representing the boundary of the LiDAR data. The tile boundaries were then copied into an Esri Enterprise Geodatabase Feature Dataset to standardize the projection of each tile. This enterprise geodatabase contains data and workflow models for multiple projects that can be used in collaboration. The re-projected boundaries, along with the LiDAR data files converted to text, became inputs to the R script.

The original R script (Xue, 2015) was designed to create a grid of cells across the extent of a LiDAR tile and uses a neighborhood function to collect different metrics within the cell and surrounding radius of LiDAR points (the neighborhood). These metrics include: point count, percent of vegetation, and percent heights (see Figure 11) for the 5th through 95th percentile in intervals of 10 (i.e. 5th, 15th, 25th… 95th).
Figure 11. Percentile height ranges vary with the distribution of returns through the canopy (Isenburg, 2015).

Including all of the original percentile heights proved to be too time-consuming for script to reasonably produce all the needed data so the original code was reconfigured for this study to return only the 25th, 50th, 75th and 99th percentiles for height per cell. These raster layers contained 15.2 m cells of LiDAR metrics across a 1,524 x 1,524 m LiDAR tile. Each generated raster had therefore at least 10,000 cells plus some additional border cells created by the R script.

A point feature class was created for the centroid of each cell in each one of the 19 LiDAR tiles in the first study area. The point feature class was much more accessible in terms of selection and analysis based on the rulesets later determined than the original rasters generated. This created a total of 192,907 points, each point representing a 15 m (50 ft) LiDAR point cloud. The six attributes were extracted then (point count, vegetation ratio, and the four height
attributes) to the cell centroid layer. An intersection selection was then run between the points layer and the polygon layer of known old growth baldcypress. This selection allowed the designation of another attribute “old growth” which had values of 1 or 0, representing yes or no, respectively. Having this vector point layer allowed for easy conversion of the attributes of the points representing cells to a text file. This resulting text file contained the point identification and the calculated metrics from the R script, along with the $x$ and $y$ coordinates of each cell and the Boolean attribute for old growth.

This text file was searched for missing values and cleaned of empty cells where no LiDAR data was collected. These empty cells and cells with only several LiDAR points were present around some of the borders of the tiles or over open water where near-infrared LiDAR pulses are much less likely due to water’s properties to absorb near-infrared and infrared wavelength energy.

The remaining 191,113 cells were processed using C5.0 (Quinlan, 2009) through the Windows GUI See5. The C5.0 options were set to allow for winnowing of the attributes and to generate rulesets. These rulesets as the output of the C5.0 classification were systematically turned into custom python scripts that run selections on the point layer to predict old growth based on the machine learning. The points from all 19 LiDAR tiles were merged into one LAS file to avoid having to iterate through the selection script multiple times. This produced a visual representation of the predictions that the C5.0 machine learning made based on the attributes. Several different iterations were repeated to change the combinations of attributes included to find the best recipe for predicting the old growth characteristics.

One possible metric for prediction included $x$, $y$ coordinates (Huang, 1997). However, this could unintendedly restrict results to the existing known old growth bounding boxes. Including
the geographic location could produce too many linear areas of selected points in the data where a better selection would not be limited to just these search areas within the larger LiDAR tiles (Figure 12).
3.3.2 **LAScanopy and ArcGIS Approach**

The restrictions that the R script had with the inability to process data quickly and that it could only produce raster with certain metrics at an unacceptable spatial resolution, the study took a turn to other methods to accomplish the processing faster with more metrics. The LASground and LAScanopy tools and the native ArcGIS geoprocessing functions became the tools of choice for the second step. These tools combined for a targeted solution that could be automated with shorter processing times. Processing of a single tile was reduced to an average of 3 minutes.

The new process began again with an LAS dataset of the original 19 LiDAR tiles covering the study area. The same cleaning process was undertaken on the dataset to remove overlap points and noise points. This gave a better baseline for calculating height above ground. Using LAStools it would have been possible to separate the overlap points into ground and non-ground classifications, and re-add the overlap ground points to create a more complete ground surface. This would have added 5 million points to the ground classification. However, as stated above, this would have skewed the calculation of total point count for cells in the overlap areas.

Including the overlap ground points had the effect of erroneously identifying bands of old growth stands at tile boundaries. The LASheight tool was used to recalculate the relative height above ground for each non-ground point. Points that fell 0.6 m (2 ft) or more below the ground surface or more than 38 m (125 ft) above the ground were eliminated as assumed noise regardless of their classification. These ranges below 0.6 m or above 38 m were derived by a manual scan of the data in LAStools for outliers that were likely misclassified noise points (birds, return errors, etc.). In subsequent applications of this process to nearby rivers in North Carolina or other possible study sites these ranges will have to be derived based on the specific datasets used in the
workflow. This is due to the arbitrary nature of noise points above or below the rest of the dataset.

Operating on the cleaned dataset, LAScanopy handled the calculation of all metrics except total point count per cell. LAScanopy was used to include as many metrics as were thought useful for discerning old growth stands. The 25th, 50th, 75th, and 99th percentile heights calculated in the first iteration were joined by the 10th and 90th percentile heights, as well as minimum, maximum, average and standard deviation of height. LAScanopy was also used to generate the canopy cover and canopy density per cell. The canopy cover is the number of first returns above a height cutoff, typically 1.3 to 1.4 meters (the same height used for measuring trunk diameter) divided by the number of all first returns in that cell. The canopy density metric is similar with the number all returns above the cutoff height divided by the number of all returns in the cell. The point count per cell metric was calculated by the LAS point statistics to raster tool in ArcGIS.

With 13 layers of LiDAR metrics for each tile, composite rasters were created. As noted previously, each raster was converted to a point feature class representing the centroids of each cell. The final attribute added to each cell was the old growth status is indicated by the reference old growth areas. Different portions of the reference areas were used such as only choosing some of the areas known to be old growth or using all the known areas of old growth as training data. By training on a variety of subsets of the reference areas it was possible to verify that the resulting rulesets correctly identified know old growth stands. The process of converting the feature dataset into text for the See5 program to process was the same as previously done for the raster from the R script workflow. The default decision tree was transformed into rulesets by See5 for ease of conversion into SQL queries in ArcGIS. The machine learning algorithms
produced rulesets that reported 1.5% errors or 13 misclassifications out of 860 training cells and 11.3% error or 95 total errors on 838 test (control) cells. 8.9 km² of newly predicted cells that are likely old growth were also identified by the machine learning classifier. A final selection was made from the predicted cells to exclude cells far outside (more than 100 m) the floodplains because of the baldcypress species preference to grow along the banks of the river channels and in the waters of backwater streams. However, the buffer did actually contain almost all of the predicted cells for previously unidentified old growth trees (see Figure 15).
4. Results and Discussion

The first ruleset that was generated by the machine learning algorithms (C5), using data from the R script, produced rectangular selections that closely followed the bounding boxes of known stands of old growth trees. Unfortunately, the R script approach was found to have the following limitations: long processing times (in some cases several hours per tile), alignment issues within the LiDAR bounding boxes, not able to produce cell sizes smaller than 15 m, and inflexibility to creating or adding additional parameters as needed. The use of the R script allowed refinement of the workflow for processing LiDAR data into selectable and usable data for both machine learning and visualization of the results. The processing of the R script in this case however was time consuming. While the R script was useful, it had to be adapted beyond its original purpose. The result was also inadequate for this study because training data attributes included the geographic $x$ and $y$ coordinates of the LiDAR returns. This created the effect of finding lots of existing old growth and almost no new areas of potential old growth.

This ruleset would not produce any useful results outside the original 19 tiles that covered the study site because it was restricted to the bounding boxes of the LiDAR tiles and the known old growth stands themselves (see Figure 12). In subsequent iterations the $x$ and $y$ coordinates were excluded from the list of attributes that the machine learning algorithms processed to generate the rulesets. The resulting ruleset created a better selection for what should be expected of predicted old growth stands (see Figure 13). Many of the selected cells fell within the existing old growth stands. The majority of the cells outside the existing old growth stands exhibited two characteristics that showed the machine learning program’s ability to predict old growth proximity to both other old growth and water. The close proximity to known stands, as well as to water, especially smaller streams, which is characteristic of old growth baldcypress.
locations of the new predictions and the corresponding visual characteristics seen in high resolution false color near-infrared imagery such as sporadic canopy cover and lower spectral signatures in the near-infrared band also indicated successful predictions.

Figure 12. See5 generated rulesets with $x$ and $y$ coordinates included in rulesets relied too heavily on the bounding boxes of the reference data.
Figure 13. See5 prediction results without $x$ and $y$ coordinates in the rulesets expand the areas of predicted old growth along the flood plain of the Black River.
The LiDAR metrics for the first two rulesets were generated by the R script which included the vegetation ratio (the number of LiDAR returns classified as low medium or high vegetation divided by the total number of points in each cell). Possible errors in the original classification of the LiDAR were noted, so it was decided that this metric did not represent the vegetation ratio metrics accurately enough for inclusion in the analysis of old growth stand characteristics. While the R script approach was ultimately rejected for the quicker and more automated workflow created by LAStools and ArcGIS, the first two rulesets produced informative results that confirmed that the machine learning system could be useful for predicting areas that should be old growth by highlighting both known areas old growth and new areas that closely followed the floodplain of the river where old growth baldcypress would be expected.

With the flexibility of the new LAStools/ArcGIS workflow, it was then possible to execute many different rulesets focusing on different metrics of the LiDAR data. The final metrics that were used are described previously in the methodology. Besides the metrics generated from the LiDAR itself, the neighborhood cell size was also a factor in the ruleset generation. While 3, 6, 7.6 and 15 (10, 20, 25 and 50 ft.) cells were tested; 15 m cells gave the most promising results. Using metrics generated at cell sizes smaller than 15 m seemed to severely under-predict old growth areas. Before the 15 m cells were used, the machine learning rarely selected more than 15-20% of the known stands. With the optimal metric cell size of 15 m, more than 50% of the known old growth was initially selected. Further configuration of the machine learning algorithm adjust the parameters of the decision tree increased the classification to just about 98.5% of the known old growth. With this optimized ruleset almost 39,000 cells or 8.9 km² of new area was predicted along the Black River study site. The newly predicted areas exhibit similar visual and
metric characteristics to known old growth. Comparisons between the near-infrared imagery and these areas confirm similar results to the first tests with initial results produced from the machine learning.

The final ruleset generated 13 rules, 4 of which classify predicted and known old growth forest areas. The remaining 9 rules classify non-old growth forest areas. Only 9 attributes of a total 13 were used across the entire ruleset (see Table 2).

Table 2. Table of metrics used in the final machine learning classifier

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Ruleset variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average height</td>
<td>avg</td>
</tr>
<tr>
<td>Canopy Cover</td>
<td>cov</td>
</tr>
<tr>
<td>Canopy Density</td>
<td>dns</td>
</tr>
<tr>
<td>Maximum return height above ground</td>
<td>max</td>
</tr>
<tr>
<td>Minimum return height above ground</td>
<td>min</td>
</tr>
<tr>
<td>10&lt;sup&gt;th&lt;/sup&gt; percentile height (p.h.)</td>
<td>p10</td>
</tr>
<tr>
<td>25&lt;sup&gt;th&lt;/sup&gt; p.h.</td>
<td>p25</td>
</tr>
<tr>
<td>50&lt;sup&gt;th&lt;/sup&gt; p.h.</td>
<td>p50</td>
</tr>
<tr>
<td>75&lt;sup&gt;th&lt;/sup&gt; p.h.</td>
<td>p75</td>
</tr>
<tr>
<td>90&lt;sup&gt;th&lt;/sup&gt; p.h.</td>
<td>p90</td>
</tr>
<tr>
<td>99&lt;sup&gt;th&lt;/sup&gt; p.h.</td>
<td>p99</td>
</tr>
<tr>
<td>Standard deviation of return height</td>
<td>std</td>
</tr>
<tr>
<td>Return count</td>
<td>pt_ct</td>
</tr>
</tbody>
</table>
The canopy density was the most influential metric being used in 10 of the 13 total rules. This indicates that some metrics, while useful in other studies for estimating various forest characteristics, are less influential than others to the prediction of old growth across the study site. The full ruleset used can be found in the appendix. A cursory look at the statistics of the metrics between the entire study site, known old growth, and newly predicted areas of possible old growth show very different distributions as seen Figure 14. As seen in Figure 15 there are many areas highlighted in green (newly predicted old growth) between the known stands in blue. There are also swaths of green that branch off from the river path that actually follows small streams and backwater floodplains. Within the floodplain many large areas of forest are highlighted as predicted old growth. The machine learning seems to over predict in many instances but upon visual interpretation of the results prediction areas, the patterns do follow visible patterns of growth in the forest that are reasonable to understand as old growth baldcypress.

Figure 17 illustrates the ruleset applied to areas along the river outside of the original study site but close in proximity the known old growth. Areas directly along the river banks and across the wider floodplain were selected by the machine learning. Again, visual inspection and alignment to the known old growth stands confirm that these are very likely more old growth. Figure 18 presents an area far away from the river channel that is currently unknown as to whether any old growth exists, but would be expected to have little or no old growth. This shows that the final machine learning ruleset could be applied to the full area of the LiDAR dataset with a relatively high level of confidence. It is yet to be seen how much new area across the entire dataset could be classified by the machine learning as possible old growth. However, the analysis
on the test data outside the 19 study site tiles shows the realization that the machine learning has indeed found valid rulesets that identify old growth.

Figure 14. These distribution curves represent the canopy density (in percentages) for the entire study, the known old growth areas (reference data) and the newly predicted areas of old growth, respectively. The similarities between the known and new old growth show that the machine learning has discovered patterns in the metrics such as canopy density.
Figure 15. The predicted old growth areas in green as classified by the See5 rulesets along with the 100m floodplain buffer
Figure 16. Existing and new predicted old growth by See5 machine learning rulesets using refined training data that fall within the floodplain buffer.
Figure 17. Tests using the final ruleset outside the original study site show many promising areas with similar patterns of predicted old growth.
Figure 18. Data away from the Black River with large areas of bare earth and farmlands show only 6 cells of predicted old growth.
Future work to more efficiently and accurately find old growth bald cypress should take two distinct steps. The most important and first step of future work would be field validation. The area must be further explored to see if areas that are predicted as old growth actually do fit those characteristics or if the old growth is better classified by other metrics not yet used. This would be give a better sense of how accurate the results of this study area and provide more training data for subsequent iterations of the machine learning. The second step would be to continue to refine and evaluate the C5.0 rulesets or determine if another machine learning approach would yield better results.

5. Conclusion

Reviewing the results produced from of all the rulesets, and especially the final ruleset, created by the machine learning program indicates a correlation exists between metrics computed from the LiDAR and known stands of old growth bald cypress trees. This evidence refutes the null hypothesis that there is no correlation between the metrics and the old growth stands. The rulesets even predict areas outside the known stands that are reasonably expected to be old growth stands. This is confirmed by the visual inspection of the predicted areas against the reference NAIP imagery. Further in situ work would need to be done to verify the findings to ground reference but the initial results are promising.

Data classification errors in the vendor provided data were a factor in being able to process the LiDAR for use in the C5.0 (or comparable) machine learning approaches. Overlapping data at the edges of the captured swaths and misclassification were the biggest hindrances to the metrics being correctly calculated.
While the data from the North Carolina Floodplain Mapping Program provided an adequate point spacing and density for LiDAR metrics to be calculated, higher density LiDAR point clouds could open up new possibilities for more precisely identifying old growth along the Black River and other areas of interest. With sufficiently dense LiDAR data trunk diameter could be determined. The denser the point cloud the more complete a model of trees even at the individual tree level can be developed. Besides increasing the collection density, LiDAR collected during different times during the growth season would provide additional information about the canopies that might be able to further identify bald cypress. Comparisons between leaf-on and leaf-off conditions could prove very useful for identifying the unique characteristics of the bald cypress such as the relatively early shedding of needles.

While visually checking against aerial imaging seems to indicate a high likelihood that the results are accurate further validation with in situ measurements is warranted. Tree cores and DBH measurements along with average canopy height estimates for old growth and non-old growth would be useful for validation of the machine learning ruleset. For areas that were selected by the machine learning but are far from the river or streams and seem to be outliers or erroneous predictions in the data would merit further investigation either by in-field study or by detailed photo analysis across multiple dates in the year. These areas possibly represent hardwood forest trees of other species that were untouched by logging. In this case the ruleset might be considered a more general predictor, and further investigation into the differences between ancient bald cypress stands and other tree species could be conducted.
6. Bibliography


http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html.


http://www.nature.org/ourinitiatives/regions/northamerica/unitedstates/northcarolina/placeswepr
otec/black-river-preserve.xml#thingsToDo.


Pandya, Rutvija, and Jayati Pandya. “C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning.” *International Journal of Computer Applications* 117, no. 16 (2015). http://search.proquest.com/openview/1862f8f1a1d4806d9815325943daf4ea/1?pq-
origsite=gscholar.


7. Appendix

7.1 Final Ruleset based on LAStools metrics and fine-tuned training data

Options:
- Winnow attributes
- Rule-based classifiers
- Do not use global pruning

Class specified by attribute ‘Old’

Read 860 cases (13 attributes) from br_new_train.data

6 attributes winnowed
Estimated importance of remaining attributes:

- 705% dns
- 118% std
- 82% cov
- 68% pt_ct
- 59% p10
- 55% p99
- 14% max

Rules:

Rule 1: (202, lift 1.8)
- dns <= 69.8
- pt_ct > 496
  -> class 0 [0.995]

Rule 2: (204, lift 1.8)
- dns <= 75.4
- pt_ct > 702
  -> class 0 [0.995]

Rule 3: (169, lift 1.8)
- max <= 38.73
  -> class 0 [0.994]
Rule 4: (130, lift 1.8)
  cov > 93
  dns <= 80.9
  -> class 0 [0.992]

Rule 5: (115/1, lift 1.8)
  dns <= 80.9
  pt_ct > 905
  -> class 0 [0.983]

Rule 6: (39, lift 1.8)
  dns > 82.1
  std <= 10.85
  pt_ct <= 586
  -> class 0 [0.976]

Rule 7: (183/4, lift 1.8)
  dns <= 80.9
  p99 <= 52.85
  -> class 0 [0.973]

Rule 8: (35, lift 1.8)
  cov > 98.1
  dns <= 93.1
  pt_ct > 586
  -> class 0 [0.973]

Rule 9: (129/15, lift 1.6)
  cov <= 95.7
  p10 > 9.48
  p99 <= 75.99
  pt_ct > 586
  -> class 0 [0.878]

Rule 10: (39, lift 2.2)
  dns > 80.9
  max <= 62.27
  pt_ct > 586
  -> class 1 [0.976]

Rule 11: (151/3, lift 2.1)
  std > 9
  pt_ct <= 496
  -> class 1 [0.974]

Rule 12: (32/1, lift 2.1)
cov <= 93
dns > 75.4
dns <= 80.9
p99 > 52.85
pt_ct <= 905
-> class 1 [0.941]

Rule 13: (516/137, lift 1.6)
dns > 75.4
-> class 1 [0.734]

Default class: 1

Evaluation on training data (860 cases):

Rules
--------
No  Errors

13  13( 1.5%) <<

(a)  (b)  <-classified as
----  ----
463    7  (a): class 0
  6    384  (b): class 1

Evaluation on test data (838 cases):

Rules
--------
No  Errors

13  95(11.3%) <<

(a)  (b)  <-classified as
----  ----
431    38  (a): class 0
  57    312  (b): class 1

Time: 0.1 secs

7.2 Ancillary Tables
Table 1. Cored trees along the Black River (Burns, 2015).

<table>
<thead>
<tr>
<th>Diameter at Breast Height</th>
<th>Ring Count Estimate</th>
<th>Sample ID</th>
<th>Protected status</th>
</tr>
</thead>
<tbody>
<tr>
<td>56 cm</td>
<td>189</td>
<td>BR5 15A</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>217</td>
<td>BLK1</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>BLK5</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>253</td>
<td>BLK79</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>253</td>
<td>BLK85</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>255</td>
<td>BLK76</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>273</td>
<td>BLK86</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>275</td>
<td>BLK87</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>BLK4</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>282</td>
<td>BLK81</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>289</td>
<td>BLK8</td>
<td>Yes</td>
</tr>
<tr>
<td>46 cm</td>
<td>296</td>
<td>BR5 6A</td>
<td>No</td>
</tr>
<tr>
<td>51 cm</td>
<td>3</td>
<td>BR5 17A</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>335</td>
<td>BLK78</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>358</td>
<td>BR5 4A</td>
<td>No</td>
</tr>
<tr>
<td>86 cm</td>
<td>389</td>
<td>BR5 21AB</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>425</td>
<td>BLK72</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>474</td>
<td>BLK2</td>
<td>No</td>
</tr>
<tr>
<td>71 cm</td>
<td>477</td>
<td>BR5 22AB</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>51</td>
<td>BLK22</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>521</td>
<td>BLK3</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>539</td>
<td>BLK64</td>
<td>No</td>
</tr>
<tr>
<td>43 cm</td>
<td>628</td>
<td>BR5 5A</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>679</td>
<td>BLK13</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>695</td>
<td>BLK66</td>
<td>No</td>
</tr>
<tr>
<td>89 cm</td>
<td>714</td>
<td>BR5 3AB</td>
<td>No</td>
</tr>
<tr>
<td>Diameter</td>
<td>Number</td>
<td>Code</td>
<td>Answer</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>58 cm</td>
<td>785</td>
<td>BR5 16AB</td>
<td>No</td>
</tr>
<tr>
<td>56 cm</td>
<td>796</td>
<td>BR5 7A</td>
<td>No</td>
</tr>
<tr>
<td>843</td>
<td>BLK28</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>845</td>
<td>BLK18</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>853</td>
<td>BLK6</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>31 cm</td>
<td>854</td>
<td>BR5 18A</td>
<td>Yes</td>
</tr>
<tr>
<td>858</td>
<td>BLK61</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>BLK3</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>893</td>
<td>BLK17</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>64 cm</td>
<td>921</td>
<td>BR5 12A</td>
<td>No</td>
</tr>
<tr>
<td>952</td>
<td>BLK54</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>954</td>
<td>BLK62</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>43 cm</td>
<td>973</td>
<td>BR5 13A</td>
<td>No</td>
</tr>
<tr>
<td>996</td>
<td>BLK1</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>998</td>
<td>BLK47</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>113</td>
<td>BLK43</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>99 cm</td>
<td>126</td>
<td>BR5 2A</td>
<td>No</td>
</tr>
<tr>
<td>148</td>
<td>BLK42</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>154</td>
<td>BLK73</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>53 cm</td>
<td>167</td>
<td>BR5 8A</td>
<td>No</td>
</tr>
<tr>
<td>179</td>
<td>BLK7</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>BLK52</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>194</td>
<td>BLK34</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>BLK65</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>66 cm</td>
<td>1099</td>
<td>BR5 14AB</td>
<td>Yes</td>
</tr>
<tr>
<td>111</td>
<td>BLK56</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>1111</td>
<td>BLK26</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>1155</td>
<td>BLK2</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>BLK48</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>1215</td>
<td>BLK19</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>71 cm</td>
<td>1297</td>
<td>BR5 11A</td>
<td>No</td>
</tr>
<tr>
<td>71 cm</td>
<td>1394</td>
<td>BR5 19AB</td>
<td>No</td>
</tr>
<tr>
<td>147</td>
<td>BLK63</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>76 cm</td>
<td>1426</td>
<td>BR5 2A</td>
<td>No</td>
</tr>
<tr>
<td>1484</td>
<td>BLK67</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>BLK31</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>BLK69</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>
7.3 Scripts and Workflow models

7.3.1 Final Workflow Model
# Import arcpy module

import arcpy

arcpy.CheckOutExtension("Spatial")

from arcpy.sa import *

# Set environment settings

#env.workspace = "Y:/R lidar script/rasters/"

inRaster = arcpy.GetParameterAsText(0)

#inRaster = "Y:\R lidar script\rasters\nb50_2704.tif"

#inRaster = "nb50_2704.tif"

# Local variables:

print(inRaster)

Rast_Name = inRaster[-13:-4]

print(Rast_Name)

Pts_Loc = "I:\Black_River2\Black_River2.gdb\"+ Rast_Name +"_points"

Band_2 = "\Band_2"

Band_3 = "\Band_3"

Band_4 = "\Band_4"

Band_5 = "\Band_5"

Band_6 = "\Band_6"

Band_7 = "\Band_7"

nb_x_TableToExcel_xls = "C:\Temp\"+Rast_Name+"_TableToExcel.xls"

#print(nb_x_TableToExcel_xls)
inRasterList = [[inRaster+Band_2, "veg_ratio"], [inRaster+Band_3, "Pert25"],
[inRaster+Band_4, "Pert50"], [inRaster+Band_5, "Pert75"], [inRaster+Band_6, "Pert100"]]
#print(inRasterList)

### Process: Raster to Point

arcpy.RasterToPoint_conversion(inRaster, Pts_Loc, "VALUE")

# Process: Alter Field

arcpy.AlterField_management(Pts_Loc, "GRID_CODE", "pnt_num", "point_count", "", "8", "NON_NULLABLE", "false")

# Process: Extract Multi Values to Points

arcpy.gp.ExtractMultiValuesToPoints_sa(Pts_Loc, inRasterList, "NONE")

# Process: Add Geometry Attributes

arcpy.AddGeometryAttributes_management(Pts_Loc, "POINT_X_Y_Z_M", "", "", "PROJCS['NAD_1983_StatePlane_North_Carolina_FIPS_3200_Feet',GEOGCS['GCS_North_American_1983',DATUM['D_North_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]],PROJECTION['Lambert_Conformal_Conic'],PARAMETER['False_Easting',2000000.002616666],PARAMETER['False_Northing',0.0],PARAMETER['Central_Meridian',-79.0],PARAMETER['Standard_Parallel_1',34.33333333333334],PARAMETER['Standard_Parallel_2',36.16666666666666],PARAMETER['Latitude_Of_Origin',33.75],UNIT['Foot_US',0.3048006096012192]]")
# Process: Add Field


# Process: Calculate Field

arcpy.CalculateField_management(Pts_Loc, "old", ",?\"", "VB", ",")

# Process: Table To Excel

arcpy.TableToExcel_conversion(Pts_Loc, nb_x_TableToExcel_xls, "NAME", "CODE")
import arcpy

points = arcpy.GetParameterAsText(0)
print(points)
tile_name = points[-16:]

arcpy.management.SelectLayerByAttribute(points, "NEW_SELECTION", "pnt_num > 1391 and pnt_num <= 2540 and veg_ratio <= 0.9876 and Pert25 > 30.08 and Pert100 > 74.29")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION", "pnt_num <= 722 and veg_ratio <= 0.9735 and Pert25 > 12.97 and Pert100 > 60.74 and Pert100 <= 70.73")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION", "pnt_num <= 722 and veg_ratio <= 0.9735 and Pert25 > 12.97 and Pert100 > 60.74 and Pert100 <= 70.73")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION", "pnt_num <= 722 and veg_ratio <= 0.9735 and Pert25 > 12.97 and Pert50 > 33.04 and Pert100 > 60.74")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION", "pnt_num <= 671 and pnt_num <= 1391 and veg_ratio > 0.8682 and veg_ratio <= 0.9876 and Pert25 <= 12.97 and Pert100 > 70.73")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION", "pnt_num <= 1391 and veg_ratio > 0.7826 and veg_ratio <= 0.9876 and pert75 > 53.09 and Pert100 > 70.73")
arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 751 and veg_ratio <= 0.7826 and Pert50 > 22.02")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 1391 and veg_ratio > 0.8682 and veg_ratio <= 0.9876 and pert75 <= 36.34 and Pert100 > 60.74 and Pert100 <= 70.73")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 692 and veg_ratio > 0.7826 and veg_ratio <= 0.966 and Pert100 <= 60.74")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 2540 and veg_ratio > 0.8682 and veg_ratio <= 0.9784 and Pert25 <= 30.08 and pert75 > 52.96 and Pert100 > 74.29 and Pert100 <= 79.33")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num > 2540 and pnt_num <= 2946 and pert75 > 74.73")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 1743 and veg_ratio > 0.7826 and veg_ratio <= 0.9544 and Pert25 <= 12.97 and Pert100 > 60.74")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 2540 and veg_ratio > 0.8682 and veg_ratio <= 0.9784 and Pert25 <= 30.08 and Pert100 > 74.29 and Pert100 <= 79.33")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 1743 and veg_ratio > 0.7826 and Pert25 <= 12.97 and Pert100 > 60.74")

arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num <= 2119 and veg_ratio > 0.8682 and veg_ratio <= 0.9876 and Pert100 > 74.29")
arcpy.management.SelectLayerByAttribute(points, "ADD_TO_SELECTION","pnt_num > 1391 and pnt_num <= 2540 and veg_ratio <= 0.9876 and Pert25 > 30.08 and Pert100 > 74.29")