DETECTION AND ANALYSIS OF EXTRAORDINARY TREE HEIGHTS IN THE GREAT SMOKY MOUNTAINS NATIONAL PARK USING REGIONAL SCALE LIDAR DATA

by

CHRISTOPHER WAYNE STROTHER

(Under the Direction of Marguerite Madden)

ABSTRACT

Tall trees that are ecologically important because of their large biomass and stored carbon may be declining due to warming associated with climate change. This study demonstrates Light Detection and Ranging (LiDAR) data collected in the Great Smoky Mountains National Park can be used to detect tallest trees in a complex mixed forest. It also highlights a methodology for processing large data volumes for quantifying and visualizing vegetation structure. Ten tall tree sites within the park were identified in the LiDAR dataset. And eight sites were field inspected to measure tree heights. A subset of the park also was examined to automatically extract tree objects from the LiDAR point cloud and compute structural parameters such as height, stem diameter, and canopy width. Multivariate regression modeling was performed to determine if LiDAR-derived datasets are efficient baselines for the modeling of environmental variables associated with tree growth.

INDEX WORDS: LiDAR, tree height, Great Smoky Mountains National Park, resource management

DETECTION AND ANALYSIS OF EXTRAORDINARY TREE HEIGHTS IN THE GREAT SMOKY MOUNTAINS NATIONAL PARK USING REGIONAL SCALE LIDAR DATA

by

CHRISTOPHER WAYNE STROTHER

B.A. The University of North Carolina – Chapel Hill, 1992

B.S. Gainesville State College, 2011

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2013

DETECTION AND ANALYSIS OF EXTRAORDINARY TREE HEIGHTS IN THE GREAT SMOKY MOUNTAINS NATIONAL PARK USING REGIONAL SCALE LIDAR DATA

by

CHRISTOPHER WAYNE STROTHER

Major Professor: Marguerite Madden

Committee: Thomas R. Jordan
J. Marshall Shepherd

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

DEDICATION

This work is dedicated to my parents Larry and Letta, for their unwavering support, as well as to my wife Emily and our children – Aaron, Molly, Olivia, Katherine, and Caden.

ACKNOWLEDGEMENTS

I would like to acknowledge those who have aided in the preparation of this thesis as well as those who have inspired me to move forward in my academic journey: Dr. Marguerite Madden, Dr. Thomas R. Jordan, Dr. J. Marshall Shepherd, and Dr. Andrea Presotto at The University of Georgia, along with Dr. J.B. Sharma, Dr. Sudhanshu Panda, and Chris Semerjian, M.S. at Gainesville State College. I would like to acknowledge the support of the Center for Geospatial Research (CGR), formerly the Center for Remote Sensing and Mapping Science in the Department of Geography at The University of Georgia and the Institute for Environmental and Spatial Analysis at Gainesville State College (now the University of North Georgia). Special thanks to the members of the Eastern Native Tree Society, especially Ian Bruckheimer and Michael Davie, whose initial work on measuring the heights of tall trees in the North Carolina mountains was the inspiration for this work. Thanks also to Paul Super and Tom Colson of the National Park Service for their assistance during my field research.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

For years, tree enthusiasts, arborists, and forestry experts have searched the world's forests for trees that exhibit extraordinary characteristics. These endeavors have been perpetrated for a variety of reasons ranging from aesthetic to historical to scientific. Scientifically, the location of exceptional trees can provide information to forest researchers about the location of old growth tree communities and the conditions favorable for growth potential. These extraordinarily tall trees are ecologically important to the planet because they are stores of large quantities of biomass and carbon. They also harbor numerous other species of plants and animals and there is some evidence to suggest that the existence of tall trees is on the decline due to the general warming associated with climate change (Laurance, 2012). Trees can absorb up to forty-eight pounds of carbon dioxide per year, accumulating one ton by age forty. It is estimated that one large tree can provide enough oxygen for two people and discharge up to one hundred gallons of water per day into the air through the process of evapotranspiration (American Forests, 2013). The large amount of data now available from LiDAR (Light Detection and Ranging) collection systems that allow analysis of specific tree heights can be useful in individual tree detection and can also help to inform ecological models, which often use the maximum measured tree heights as the upper limit for species height potential.

Groups such as American Forests (America's oldest conservation nonprofit) and the Eastern Native Tree Society (ENTS) have devoted much of their time and efforts to documenting the locations of candidate tallest individual trees of various species and collecting accurate field

measurements of tree heights in the Eastern United States. The ENTS in a nonprofit organization established in 1996 to measure and record the tallest trees in Eastern North America. The members of this society use a combination of measurement methods recommended by the U.S. Forest Service to measure tree heights including field verification by climbing and measuring the tree heights with a tape drop method.

The Great Smoky Mountains National Park (GRSM) is especially suited to studies of very tall trees due to its location at the southern extent of the Appalachian mountain range, high range of elevation values, diversity of terrain aspects and moist and temperate climate. The GRSM has been allowed to thrive mostly undisturbed since its creation in 1934 to mitigate the damage created by logging activities in the 1800s and early 1900s (Houk, 2000). It is home to over one hundred species of trees and comprises the most extensive virgin hardwood forest in the eastern U.S.(Whittaker, 1956; Houk, 2000). These favorable park attributes are reflected in its high Rucker Index (RI) value (163.6) compared to other tall tree sites. The RI is a measure of the average height of the tallest examples of the ten tallest species at a site, created by Colby Rucker of the ENTS. To-date, the ENTS reports a tulip tree (*Liriodendron tulipifera*) with a height of 58.0 m as the tallest recorded tree in the GRSM, found in North Carolina in the Fork Ridge Trail area using LiDAR data obtained for N.C.'s floodplain mapping program. (Rucker, 2004). Correspondence on the ENTS website (www.ents-bbs.org) reports the tallest tree recorded to-date in the Tennessee portion of the GRSM is a 52.7-m tulip tree on Porters Creek near Gatlinburg, TN and speculation among members of the society is that the tallest tree in the eastern U.S. is in Tennessee.

LiDAR remote sensing provides three-dimensional information about the Earth's surface using emitted laser pulses. This information can be useful in examination of terrain surface

models, vegetation characteristics, and man-made features (Renslow, 2012). In February and April of 2011, acquisition of multiple return, moderate density (< 1 m point spacing) LiDAR data for the Tennessee portion of the GRSM was completed as part of a project funded by the U.S. Geological Survey (USGS) under the American Recovery and Reinvestment Act (ARRA) of 2009. The USGS provided oversight for the project conducted by the Center for Remote Sensing and Mapping Science (CRMS), Department of Geography at The University of Georgia (UGA) and the Institute for Environmental and Spatial Analysis (IESA) at Gainesville State College (GSC). A public-private partnership between faculty and students of UGA/GSC and Photo Science, Inc. created jobs and trained students involved in tasks from project management to data processing and metadata creation. The availability of these data provided the opportunity for this research, namely to process large quantities of tree canopy height data in a mixture of automated and manual processes that could potentially save hundreds of hours of field work locating potential sites of extraordinary trees. By ranking these sites by their LiDAR predicted heights, field teams can be targeted and directed to the areas where these trees grow and accurate field measurements can be taken, reducing the chances of following anecdotal evidence of tall tree site locations to fruitless results.

The large amount of data inherent in the multiple return LiDAR format, sometimes several hundred gigabytes (Gb), coupled with emerging feature extraction software, also creates an opportunity for rigorous statistical analysis to determine whether the large databases created by these extractions can be used to accurately predict environmental factors that contribute to forest canopy heights. The possibility exists to create models that show how variables such as elevation, soil type, terrain aspect, and slope can predict tree heights in these types of highly variable environments.

The Great Smoky Mountains National Park

The GRSM was established in 1934 after Tennessee and North Carolina donated the purchased lands to the Federal government (Figure 1.1). Straddling the boundary between the two states, it was established as a conservation measure to mitigate the damage caused by

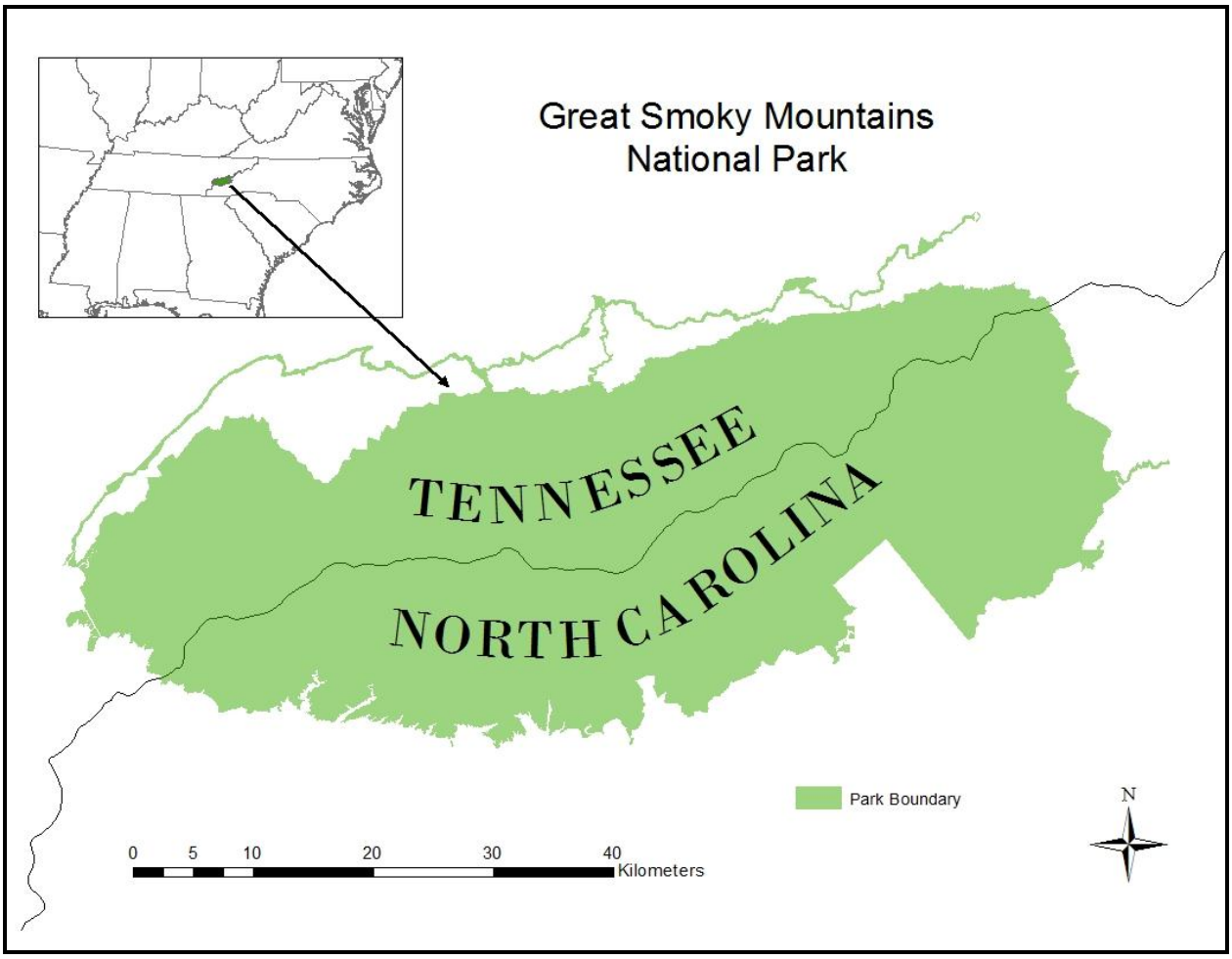

Figure 1.1 - The Great Smoky Mountains National Park

erosion and fire due to the extensive logging of the 1800s and early 1900s (Madden et al., 2004). It is estimated that up to two-thirds of the lands that comprise the park were logged or burned by the 1920s (Walker, 1991). The park encompasses approximately 209,000 hectares of forest cover, forty percent of which make up the most expansive virgin forest land on the East coast

(Houk, 2000). The park is part of the Appalachian Mountain range, and is one of the most biologically diverse areas on the planet, containing more tree species than in all of Northern Europe (NPS, 1981). The park is also the most heavily visited park in the National Park system, receiving up to 10 million visitors a year. In light of its invaluable ecological importance and potential threats from surrounding urbanization and climate change, it has been designated as an International Biosphere Reserve and a U.N. World Heritage Site (Welch et al., 2002). The GRSM has elevation that ranges from 250 m at the western border to 2025 m at Clingman's Dome. Its climate is generally temperate and humid with over 200 cm of precipitation per year. Geologically, the region is comprised of sedimentary rock formations, the result of a shallow sea that covered the area six to nine million years ago (Moore, 1988).

LiDAR Remote Sensing

LiDAR is an active remote sensing system that uses laser pulses emitted from a sensor mounted on an airborne platform, usually an airplane or helicopter, to measure features on the ground (Figure 1.2). Laser light pulses travel from the sensor to ground features and are reflected back to the sensor. The data are captured as x, y, and z coordinates in a point cloud, where the x and y values represent the geographical location on the surface and the z value represents the elevation value (Figure 1.3). Airborne LiDAR data are typically collected by scanning perpendicular to the line-of-flight with a near-infrared laser with a wavelength of 1040 to 1060 nm. The elevation of the scanned feature is determined by calculating the time it takes for the laser pulse to leave the sensor and return using the formula:

$$R = \frac{1}{2(tc)}$$

where R is the distance between the feature and the sensor, t is the time in seconds, and c is the speed of light (3×10^8 m/sec) (Jensen, 2007). On-board data collectors receive the

information from each laser pulse's return and create an x, y, and z coordinate point for each return based on information derived from on-board and ground-based Global Positioning Systems (GPS) as well as Inertial Measurement Units (IMUs) that record the attitude of the aircraft's pitch, roll, and yaw during collection. This type of active sensor has been used since the 1990s to perform cost-effective and rapid assessments of terrain characteristics (Maune, 2001).

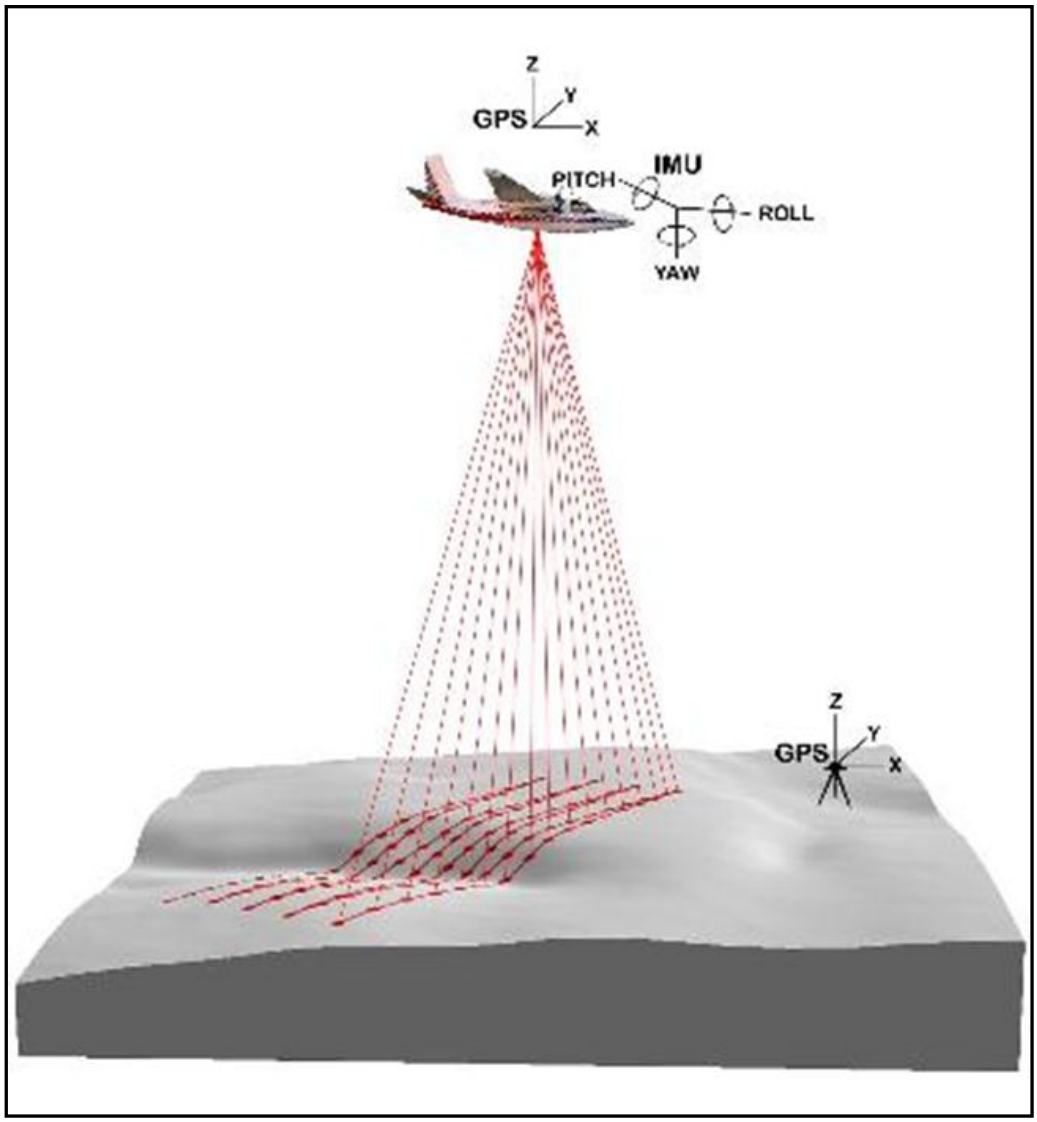

Figure 1.2 – Representation of an airborne LiDAR system

(http://forsys.cfr.washington.edu/JFSP06/lidar_technology.htm)

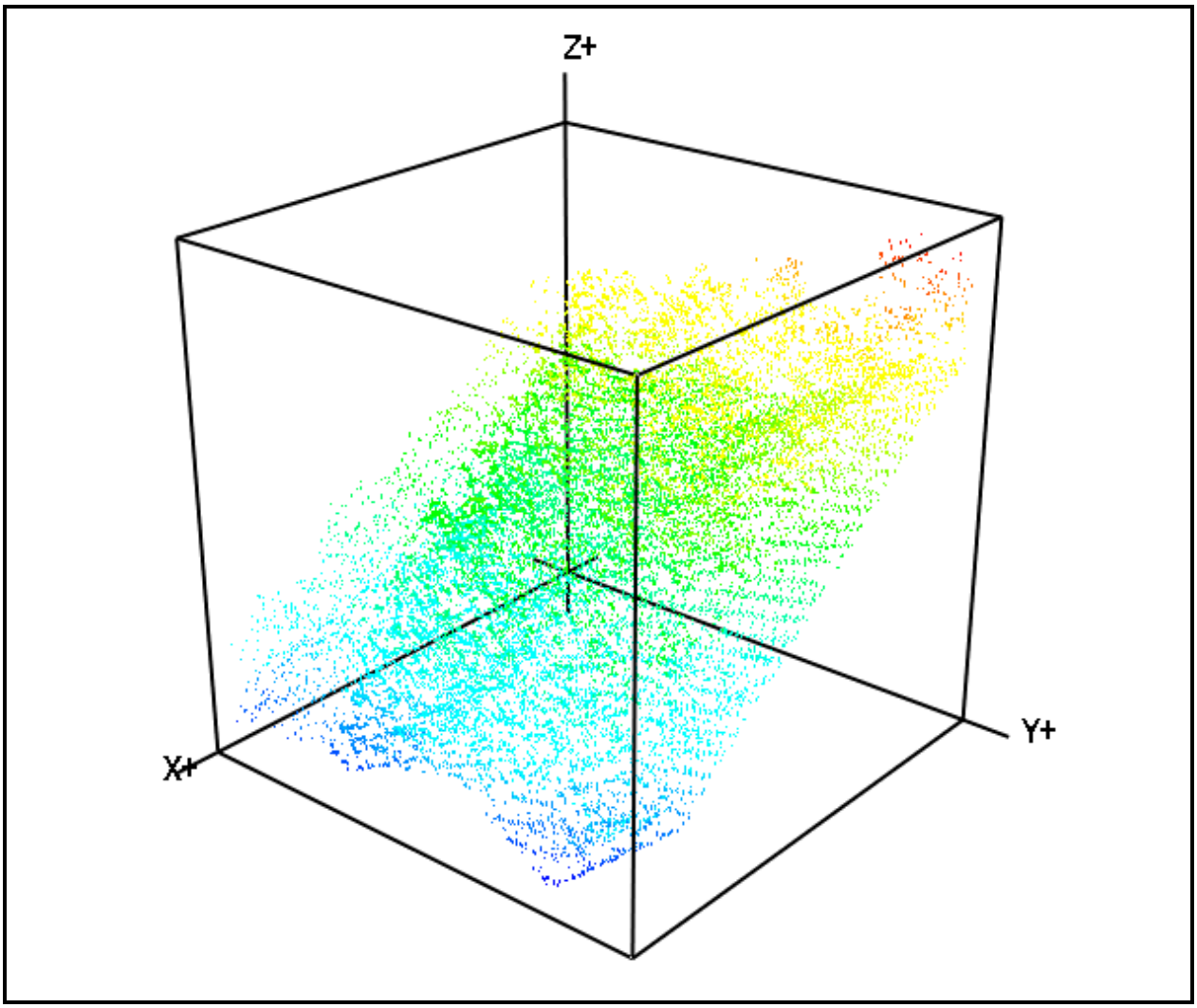

Figure 1.3 – Visualization of a LiDAR point cloud showing the x, y, and z axis.

After collection, LiDAR data sets are typically classified according to a standardized scheme created by the American Society for Photogrammetry and Remote Sensing (ASPRS). This study's data adheres to version LAS 1.2. This classification scheme classifies points into four categories: 1= Nonground, 2 = Ground, 7 = Noise, and 12 = Overlap (ASPRS, 2008). Classification is made possible by the multiple returns of energy to the sensor by a single laser pulse that can occur. First pulse returns measure the distance to the first surface contacted, such as tree canopy or rooftops, while last pulse returns measure the distance to the last surface

contacted, which includes the ground (Figure 1.4) (Renslow, 2012). The collection of multiple returns in a vegetated area is extremely valuable because it provides information about the vertical structure of the canopy. This classification process allows for the selection of last return

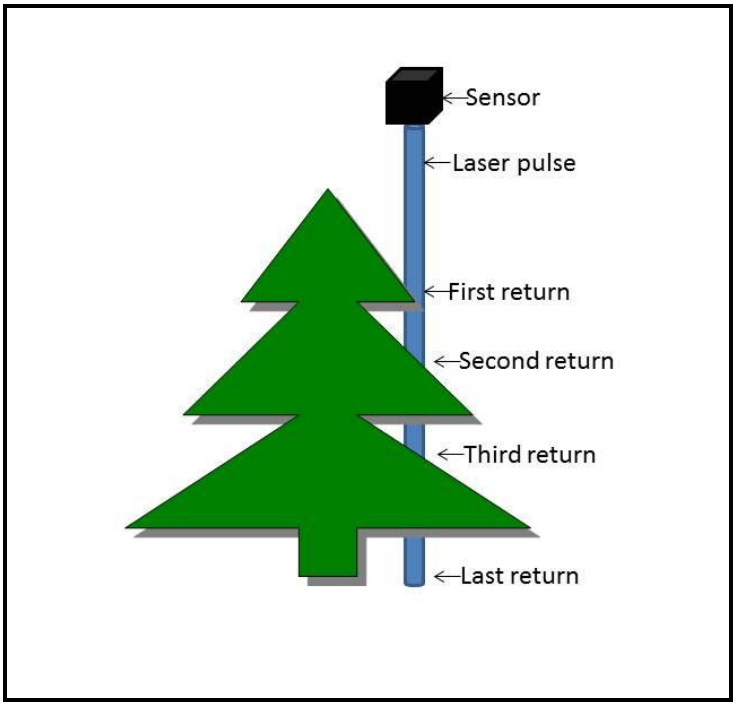

Figure 1.4 - Illustration of multiple LiDAR pulse returns from a single laser pulse

points to create digital elevation models (DEMs), which are interpolated surface models of the bare Earth with all above ground features removed, as well as digital surface models (DSMs) which model surfaces above ground. It should be noted that often the first return can also be the last return in cases where the pulse hits bare soil, grass, or even a solid surface above ground such as a rooftop. Canopy height models (CHMs) can be created to model vegetated canopy heights above the terrain by taking the difference between the DSM and the DEM and creating an interpolated surface model that represents the canopy above ground.

LiDAR data provide the remote sensing analyst with the ability to create interpolated models of Earth and feature elevations of increasingly high resolution and accuracy with

relatively low costs compared with conventional field techniques. DEM rasters created from LiDAR datasets having spatial resolutions less than 2 m is not uncommon. Class-I grade LiDAR for engineering can have horizontal accuracies with RMSE of ± 20 cm or better and vertical accuracies of ± 5 cm or better (Renslow, 2012). The vast amount of sampled points acquired in a LiDAR collection at these accuracies would be prohibitively expensive for a traditional survey crew to complete in the relatively short time (hours or days vs. days or months) it takes to fly a LiDAR mission.

The Use of LiDAR in Forestry Applications

Airborne LiDAR has been used extensively in the past two decades to obtain accurate measurements of forest structure (Nilsson, 1996; Maune, 2001; Jensen, 2007; Andersen et al., 2006). In the context of forestry, height is defined as the vertical distance between the ground and the top of the tallest tree branch (Husch et al., 1972). Research conducted by the U.S. Forest Service in western Washington State using LiDAR sensors produced sub-meter horizontal and vertical accuracies of a mountainous, forested area dominated by Douglas fir (McGaughey et al., 2004). The maximum height of trees was predicted with R^2 values between 85 and 90% in a mixed forest area of Appomattox-Buckingham State Forest in Virginia, U.S. (Popescu et al., 2002). The methodology of subtracting the DEM from the DSM to obtain the relative tree heights has been used by others in measuring forest structure. (Naesset, 1997; Zimble et al., 2003; Andersen et al., 2006). Past studies have pointed to the errors inherent in the LiDAR data format that can be created by the post spacing, or distance between height measurements that become apparent when ground measurements are made to the highest peaks visible in the tree crown (Figure 1.5).

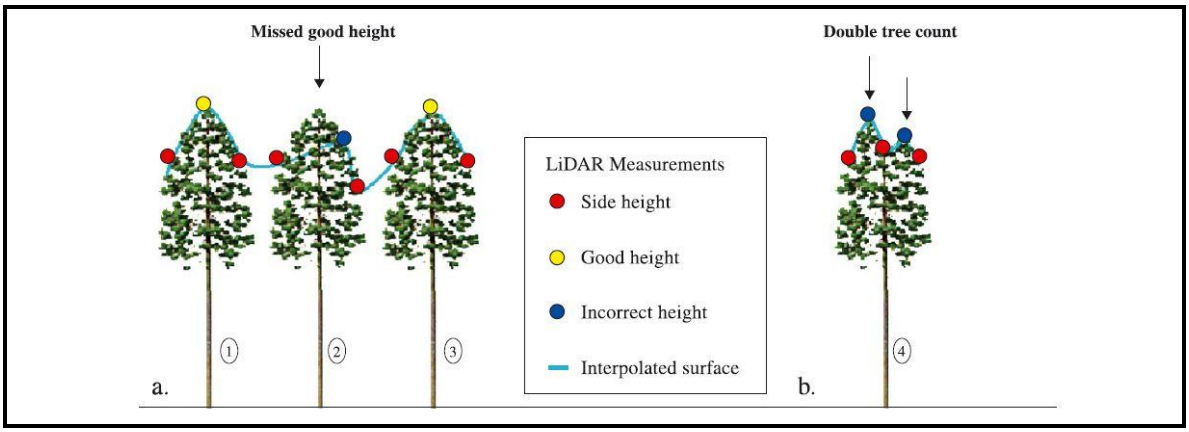

Figure 1.5 - Height variance explanation a) Trees 1 and 3 correctly measured, but Tree 2 is measured on the side of the crown. b) Tree 4 is incorrectly measured as two trees. (Zimble et al., 2003)

Ground-based Tree Height Measuring Techniques

Andersen et al. (2006) pointed out that accurate direct measurements of the heights of trees in the field are difficult. Crown overlap in dense canopy, as well as other factors such as slope can affect the *in situ* measurements. The U.S. Forest Service (USFS) indicates that the best height measurements are made using an instrument that functions as a laser rangefinder with a built-in clinometer (Figure 1.6) (USFS, 2005). This tool measures the horizontal distance (hd) to the tree from a fixed location as well as angles to the base of the tree (Θ) and the tip of the crown (ρ) (Figure 1.7). The tree height (h) is derived by the trigonometric equation:

$$h = hd(\tan \rho + \tan \Theta)$$

Figure 1.6 - Impulse 100 laser rangefinder with clinometer used in the study

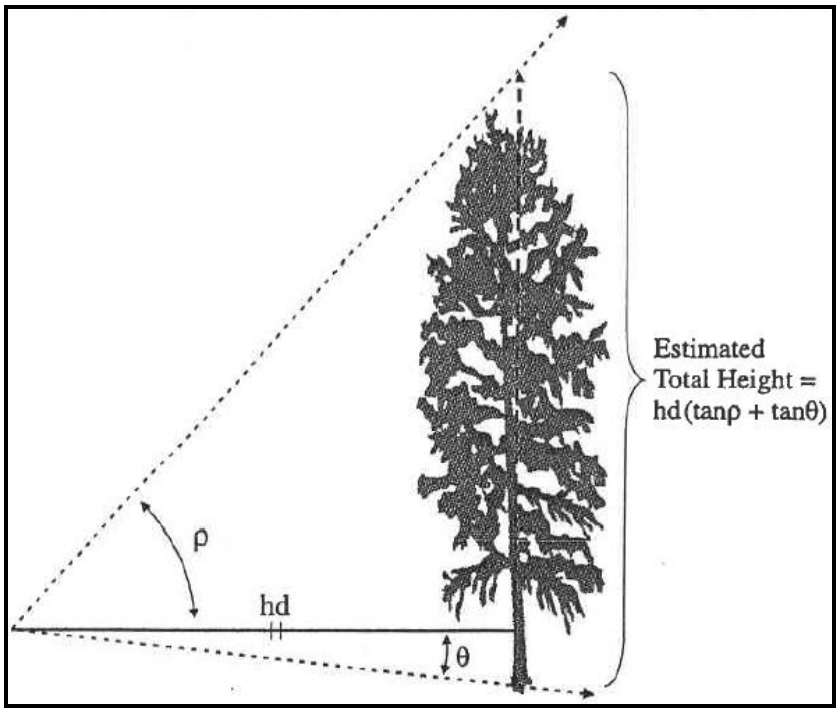

Figure 1.7 - Diagram of the principles of field-based tree height measurements using a clinometer (Andersen et al., 2006)

Thesis Objectives

This manuscript-style thesis contains two chapters that will be submitted in 2013 for peer-review and potential publication. Both articles focus on the analysis of airborne LiDAR data for tree object extraction and height measurement. The first article discusses the use of LiDAR data to determine a methodology for detecting the tallest trees in the Tennessee portion of the Great Smoky Mountains National Park within the constraints of a large volume data set. The second article provides the development of a statistical analysis of an automatically derived tree height database using LiDAR data from the GRSM and its potential efficacy.

The primary goal of this research is to investigate the use of LiDAR as a remote sensing tool for assessing vegetation structure and providing resource managers with detailed information on canopy height. Special focus will be made to challenges of data management and manipulation methodologies and specifically the challenges of working with large data sets containing billions of potential points. To achieve this goal, specific objectives include:

1. Detection of maximum tree heights in the GRSM by creating a methodology for processing a large dataset (724 tiles – each representing 225 ha in area and around 200 – 300 Mb file size) of recently acquired (2011) LiDAR data to identify potential trees of extraordinary height and to assess the environmental conditions at the top ten sites (Chapter 2).
2. The use of multivariate regression models to assess the validity of using LiDAR-derived tree height databases to predict tree heights in a highly variable forested environment (Chapter 3).

This research will provide the National Park Service with a database of the locations and heights of the tallest trees in the GRSM using publicly available data. It will provide the Park

managers with a baseline of canopy height data that can serve as the beginning of temporal analysis of canopy height. The University Consortium for Geographic Information Science (UCGIS) lists temporal analysis using remotely sensed data as its number one long term priority area of research (McMaster & Usery, 2005). While LiDAR remote sensing has been used extensively in the last decade to predict forest parameters, this is the first study of its kind to detect specific individual trees using the 2011 USGS data for the complex mixed deciduous forest of the GRSM. Upon completion of this work, the Park resource managers will have a digital database, and map visualizations of the ten tallest tree sites in the Tennessee portion of the Park, along with the environmental conditions associated with the trees. These will include the individual tree's species, its height, its surrounding overstory community, the ground slope, aspect, elevation, and adjacency to riparian areas. Multivariate statistical analysis of these parameters will provide correlation coefficients for predicting maximum tree heights in other parks. Using this information, the Park managers can effectively monitor these tree sites for their protection and further research of old growth forest areas.

CHAPTER 2

LIDAR DETECTION OF THE TEN TALLEST TREES IN THE TENNESSEE PORTION OF THE GREAT SMOKY MOUNTAINS NATIONAL PARK[1]

[1] Strother, C.W., M. Madden, T. Jordan, and A. Presotto. To be submitted to *Photogrammatic Engineering & Remote Sensing.*

Abstract

LiDAR (Light Detection and Ranging) has been shown to be an effective tool for measuring forest parameters such as canopy height. Researchers also have used LiDAR data for individual tree detection. This article describes a method for predicting the locations and heights of the ten tallest trees in the Tennessee portion of the Great Smoky Mountains National Park (GRSM) using airborne LiDAR data collected between February 15 and April 7, 2011 as part of the U.S. Geological Survey (USGS) National Geospatial Program to add imagery and elevation data to The National Map. Iterative computation tools were utilized to process the LiDAR data along with the LiDAR-derived bare Earth digital elevation models (DEMs) and digital surface models (DSMs) to create canopy height models (CHMs) for the entire Tennessee portion of the park. A height threshold of 51.8 meters was chosen as the minimum value for a tree of extraordinary height based on discussions with arborists familiar with the area. Ten potential sites containing tall trees were identified using this methodology and seven of the top ten ranking trees' heights were field measured using accepted forestry methodology. The trees detected using these methods are potentially the tallest trees ever measured on the East Coast of the United States. These methods show that extraordinarily tall trees can be successfully detected in a large area using large amounts of LiDAR data with varying accuracy. It is proposed that the variations in accuracy found in this study are a result of the severe terrain characteristics as well as variations in heights of the individual tree species.

Introduction

LiDAR in Forestry

Airborne LiDAR data has been used extensively in the past decades to obtain accurate measurements of forest structure (Nilsson, 1996; Maune, 2001; Jensen, 2007; Andersen et al.,

2006). In the context of forestry, height is defined as the vertical distance between the ground and the tip of the tree crown (Husch et al., 1972). Research conducted by the U.S. Forest Service in western Washington State produced sub-meter horizontal and vertical accuracies in a mountainous, forested area dominated by Douglas fir using airborne LiDAR data (McGaughey et al., 2004). The maximum height of tree plots was predicted with R^2 values between 85 and 90% in a mixed forest area of Appomattox-Buckingham State Forest in Virginia, U.S. (Popescu et al., 2002). The methodology of subtracting the digital elevation model (DEM) values from the digital surface model (DSM) values to obtain canopy heights has been used by others in measuring forested areas. (Naesset, 1997; Zimble et al., 2003; Andersen et al., 2006). Past studies have pointed to height measurement errors created in the LiDAR data collection process created by the post spacing, or distance between height measurements that become apparent when ground measurements are made to the highest peaks visible in the tree crown (Figure 2.1).

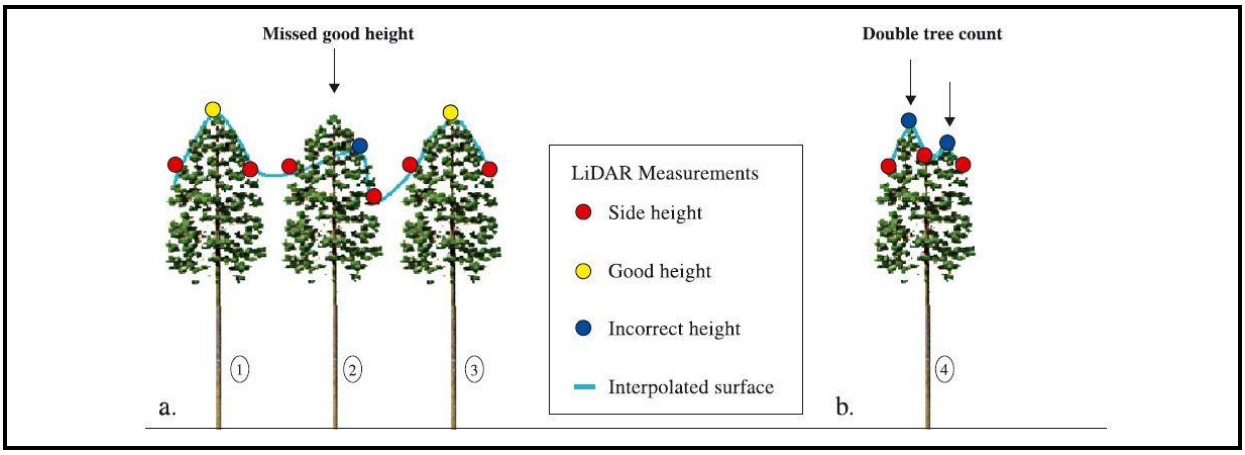

Figure 2.1 - Height variance explanation a) Trees 1 and 3 correctly measured, but Tree 2 is measured on the side of the crown. b) Tree 4 is incorrectly measured as two trees. (Zimble et al., 2003)

Ground-based Tree Height Measurement Procedures

Andersen et al. (2006) point out accurate direct measurement of trees in the field is difficult. Crown overlap in dense canopy as well as other factors such as slope can affect the ground measurements. The U.S. Forest Service (USFS) indicates that the best height measurements are made using an instrument such as a laser rangefinder with a built in clinometer (USFS, 2005). This tool measures the horizontal distance to the tree (hd) from a fixed location as well as angles to the base of the tree (Θ) and the tip of the crown (ρ) (Figure 2.2). The height (h) is derived by the trigonometric equation:

$$h = hd(\tan\rho + \tan\Theta)$$

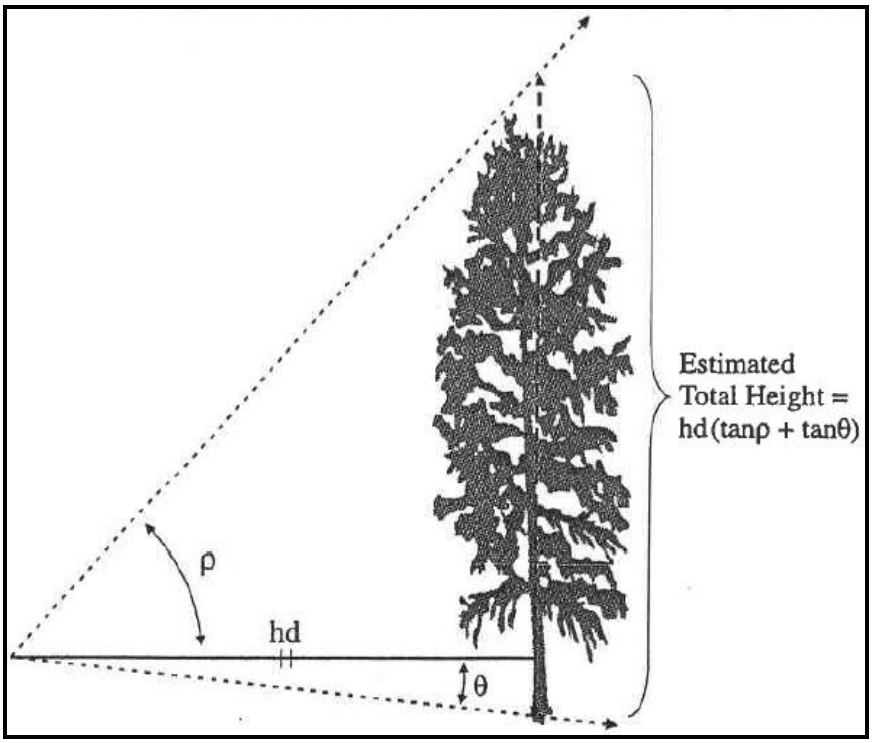

Figure 2.2 - Field tree height measurement (Andersen et al., 2006)

Study Area

The 209,000 hectares of the Great Smoky Mountains National Park straddle the border between the states of Tennessee and North Carolina (Figure 2.3). The GRSM receives over 10

million visitors a year; making it the most visited National Park in the U.S. This area contains roughly 1500 meters of relief ranging from around 250 m at the western border of the park to 2025 m at Clingman's Dome, the highest mountain in Tennessee and third largest east of the Mississippi (NPS, 2012). The park was created in 1934 from lands donated by Tennessee and North Carolina in an attempt to mitigate the devastating effects nineteenth century timber logging and subsequent erosion. The park is part of the Appalachian Mountain range, one of the oldest mountain ranges on Earth. It is also one of the most biologically diverse areas on the planet, given its unique terrain, its geologically rich sedimentary soil, and its ample precipitation. The park contains almost forty percent virgin forest cover. The extreme slopes found in the park as well as the other environmental variables create a unique opportunity for tall trees to exist undisturbed. It also complicates efforts to make accurate field measurements of those trees.

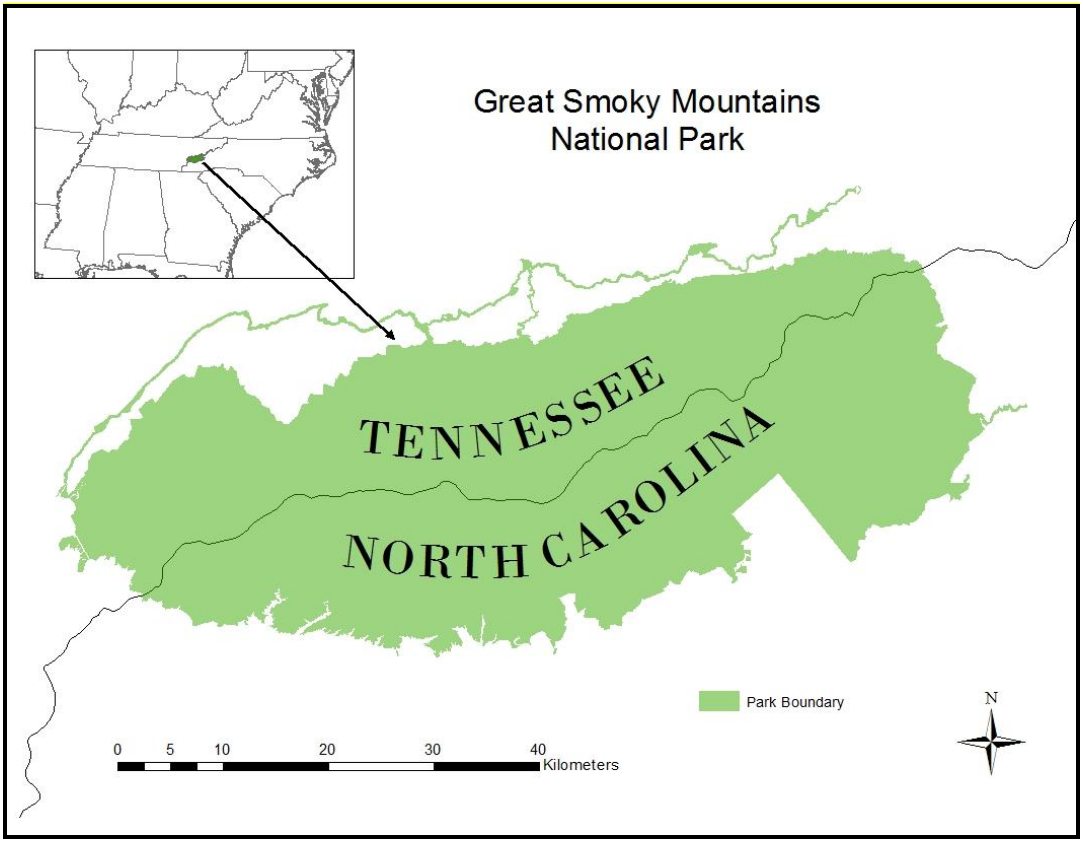

Figure 2.3 – The Great Smoky Mountains National Park

Previous Work and Objectives

Ian Breckheimer of the University of North Carolina performed a similar unpublished study of tree heights in the North Carolina portion of the GRSM in 2011 using LiDAR data obtained by the state for their Floodplain Mapping Program. His methodology focused on specific areas where arborists had given anecdotal evidence concerning potential sites, whereas this study follows a more automated and comprehensive approach to obtaining candidate sites. Breckheimer's work included a maxent ecological model for tall tree (in this case, > 54.9 meters) suitability sites which included variables that considered elevation, topographic moisture, aspect, and disturbance history (Figure 2.4).

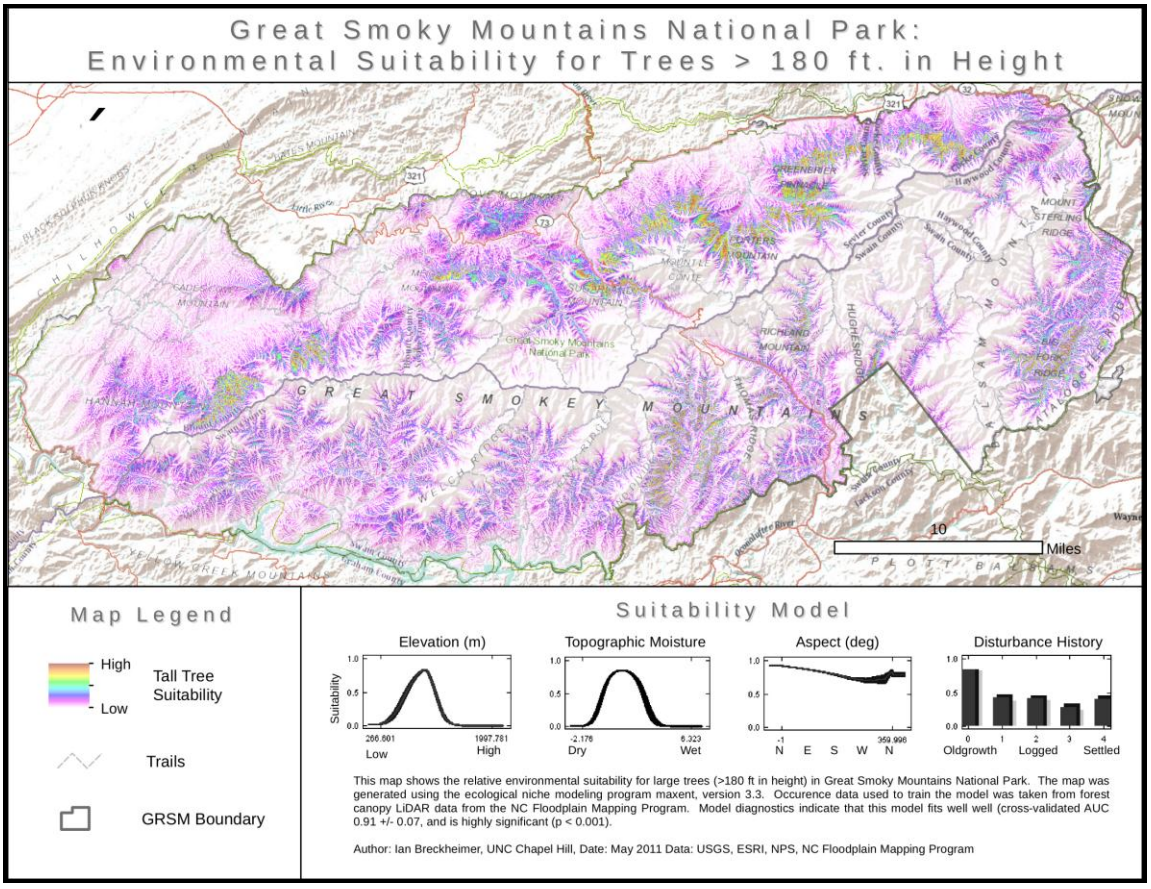

Figure 2.4 - Breckheimer's tall tree suitability map for the Great Smoky Mountains National Park (Breckheimer, 2011)

Previously, the Eastern Native Tree Society (ENTS) reported a tulip tree (*Liriodendron tulipifera*) with a height of 58.0 m as the tallest recorded tree in the GRSM, found in North Carolina in the Fork Ridge Trail area using LiDAR data obtained for the N.C. floodplain mapping program. (Rucker, 2011). Correspondence on the ENTS website (www.ents-bbs.org) reported the tallest tree recorded to-date in the Tennessee portion of the GRSM is a 52.7 m tulip tree on Porters Creek near Gatlinburg, TN.

This work will provide a unique methodology to process large volumes of recently acquired LiDAR data for the Tennessee portion of the GRSM. The specific objective of this study is to show that these data can be used to detect trees that are taller than have ever been measured in the eastern U.S.

Data

Researchers at the Center for Remote Sensing and Mapping Science (CRMS) in the Department of Geography at The University of Georgia and the Institute for Environmental and Spatial Analysis (IESA) at Gainesville State College (GSC) collaborated on a project to collect high resolution orthoimagery and LiDAR data of the GRSM and adjacent Foothills Parkway as part of the USGS Geospatial Program to add imagery and elevation data to The National Map. Funded by the American Recovery and Reinvestment Act (ARRA) of 2009, a total of 111 flight lines were used to ensure complete coverage of high resolution (30 cm) four-band orthoimagery and one meter point spacing point clouds of LiDAR data covering the study area (Figures 2.5 & 2.6). The LiDAR data used in this study were acquired by Photo Science, Inc. using both a Leica ALS-60 sensor and an Optech ALTM Gemini sensor. Both sensors use a laser with a wavelength of 1064 nm (Table 2.1). Poor weather conditions preventing leaf-off and snow free conditions created a narrow time window of data acquisition, but Photo Science was able to

successfully complete data collection in February and April of 2011. High resolution four-band digital at 30 cm spatial resolution imagery was also collected of the same study area by the same commercial vendor.

Table 2.1 – LiDAR sensor specifications

Sensor	**Optech ALTM Gemini**	**Leica ALS-60**
Altitude (AGL)	1981.2 m	Not listed
Speed	110 knots	150 knots
Scan frequency	20.2 Hz	34 Hz
Scan angle	± 16°	± 16°
Pulse frequency	50 kHz	53 kHz

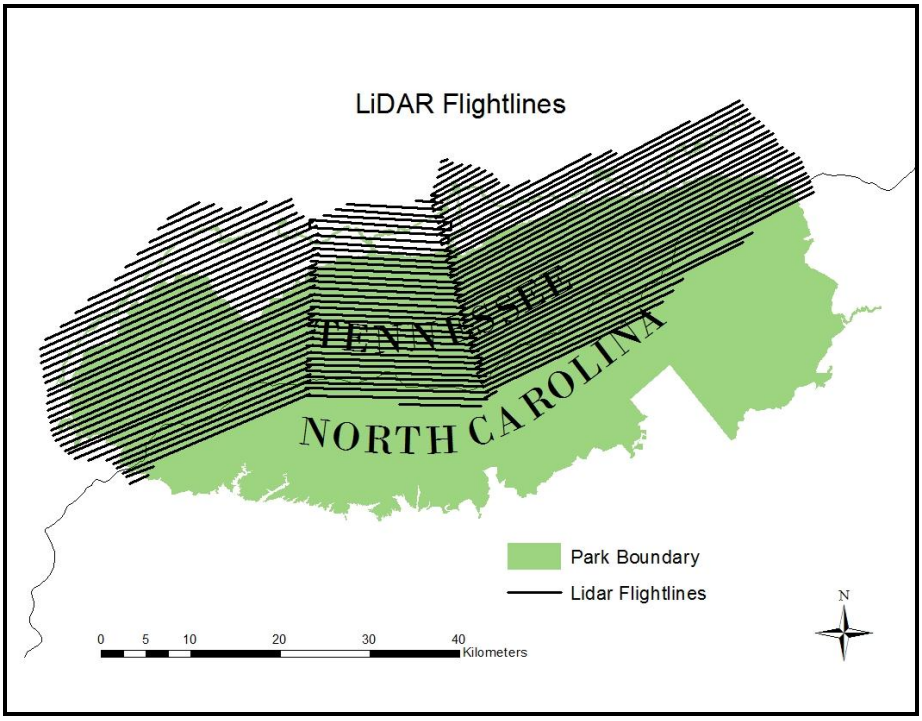

Figure 2.5 – LiDAR and orthoimage data acquisition flight lines

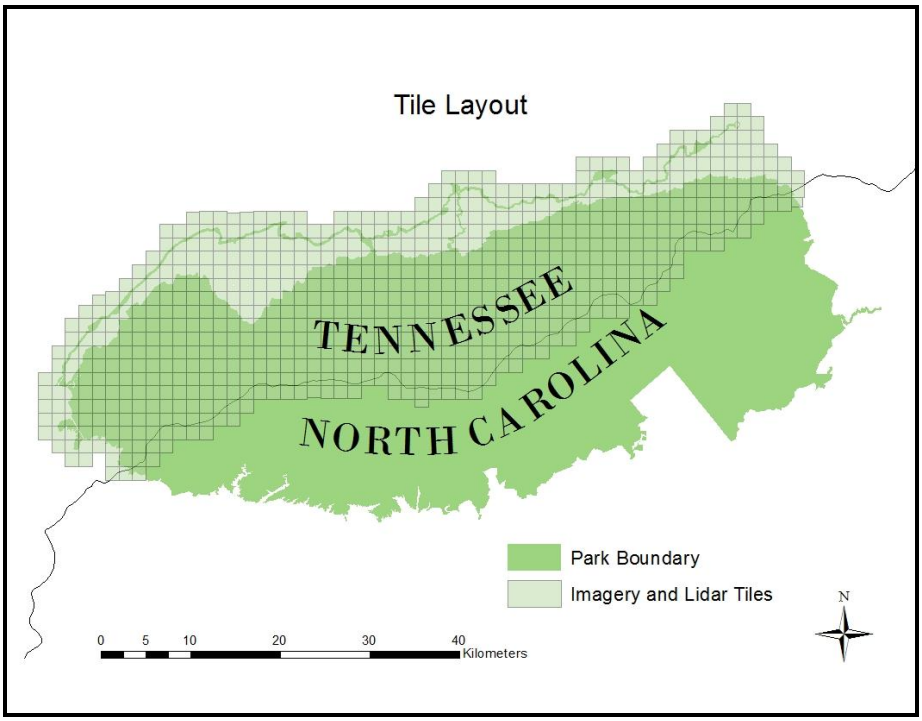

Figure 2.6 - LiDAR and orthoimage tile layout

Photo Science, Inc. performed the initial processing of the collected orthoimagery and LiDAR data and created the interpolated bare Earth DEMs used in this study. Applanix software was used to perform the GPS and inertial correction of the collected data. GeoCue, TerraScan, and TerraModeler software was used to classify the point cloud according the LAS 1.2 specifications and 1.5 m resolution bare Earth elevation models were created using the classified ground points by Esri products in an .img format (CRMS-UGA, 2011). The CRMS in the Department of Geography at The University of Georgia and the IESA at GSC were involved in the planning and post-processing of the raw image and LiDAR data and are the sources of the LiDAR data used in this study. Other layers used in the analysis were obtained from public sources listed in Table 2.2.

Table 2.2 – Description of data used in the study

Data Layer	**Source**	**Datum and Projection**	**Resolution (m)**	**Comments**
724 .tif files	UGA CRMS 2011	NAD 83 UTM 17N NAVD88	0.3	Digital color images of study area
724 .img files	UGA CRMS 2011	NAD 83 UTM 17N NAVD88	1.5	Lidar-derived DEMs
724 .las files	UGA CRMS 2011	NAD 83 UTM 17N NAVD88	0.69	Vertical RMSE of 0.165 m
statesp020.shp	nationalatlas.gov 2005	Geographic Lat/Long		Projected to NAD 83 UTM 17N
overstory.shp	UGA CRMS 2004	NAD 27 UTM 17N		Vegetation reference map
stream.shp	UGA CRMS 2004	NAD 83 UTM 17N		Stream network map
GRSM_boundary.shp	UGA CRMS 2004	NAD 83 UTM 17N		Park boundary map
Quads27block.shp	UGA CRMS 2004	NAD 83 UTM 17N		USGS Quad map
GRSM_Trail_clip.shp	UGA CRMS 2004	NAD 83 UTM 17N		Park trails map
GRSM_Major_road_hwys.shp	UGA CRMS 2004	NAD 83 UTM 17N		Road network map
Classified_LAS_Point_File_Info.shp	UGA CRMS 2004	NAD 83 UTM 17N		Lidar, DEM, and image tile index

Methods

It is assumed that by using the discrete returns from the LiDAR data, features with a certain height value above ground can be detected. Using ArcGIS® 10, several process models were created in order to handle the large amount of data processing (over 228 GB of LiDAR point cloud data) necessary to perform the analysis of areas with the highest z-values. First, a model was developed to transform the 724 .las data files (3.8 billion data points) into 724 "multipoint" .shp files (Figure 2.7). These .shp files were then converted to a raster format (.tif) using another iterative model (Figure 2.8) based on the maximum value for z within each cell, creating a digital surface model (DSM). The pixel size was selected as 1.5 m in order to correlate with the 1.5 m DEM dataset that was created by the commercial LiDAR vendor, Photo Science. ArcMap® 10 did not provide an iterative model capable of processing the normalized DSM (nDSM) that is necessary to determine z-values relative to the ground surface (heights). Therefore, a script was developed using Python programming language to perform this function, subtracting the z-values of the 724 DEMs from the 724 DSMs and creating new rasters that contained the heights relative to the ground in each cell (Figure 2.9). This script was used to create a tool within ArcMap that can be used in future processing of nDSMs (Figure 2.10). Another iterative model was created to extract pixels within the new nDSMs that contained values greater than 51.8 m (about 170 ft.) because it was determined that any trees above this height would be considered "tall" based on conversations with members of the ENTS (Figure 2.10).

The resulting rasters were mosaicked to create a single raster image of 17961 pixels that contained z-values greater than 51.8 m (Figure 2.11). This raster was then converted back to vector format (point) and queried for features that had z-values (heights) between 52 m and 60 m

as it was determined unlikely for any trees to be higher than 59 m. This height range was again based on correspondence with the ENTS. The resulting 2784 features were then selected by their location within the Park boundary, excluding points that may have been within a data tile but not within the GRSM boundary, resulting in 2178 possible height values that met the study parameters (Figure 2.12). At this point, manual interpretation was employed to remove man-made features from the processed results, including apparent power lines. The 1523 remaining values were sorted according to maximum z-values and analyzed as possible individual tree crowns. If points were arranged in a cluster of more than two points, they were considered possible tree crowns. Using the 2011 high resolution imagery, the potential tree sites were validated to ensure that the features were natural and not man-made.

The top ten sites were all located in the northwestern portion of the park (Figure 2.14). These potential tall trees were mapped according to their UTM coordinates and these coordinates were loaded onto a hand-held Garmin Etrex Vista HCx GPS unit for field verification. On March 12, 2012, six sites (Sites 5 – 10) were visited and tall trees were measured using an Impulse 100 clinometer and rangefinder based on the methodology described by Andersen et al. (2006). Three measurements were taken for each tree at each site to obtain a mean field height to be compared to the LiDAR predicted heights from the dataset. On March 13, 2012, an attempt was made to access the remaining four potential sites (Sites 1 – 4), but due to extreme topography and a flooded creek channel, only one site (Site 3) could be accessed. Ground photos were taken to document the trees as well as to show the terrain conditions of the potential tall tree sites. Once the field work was completed, further examination and verification of each tree site was performed using the US Forest Service FUSION LiDAR utility tool, which measured the tallest points in the data set at each site.

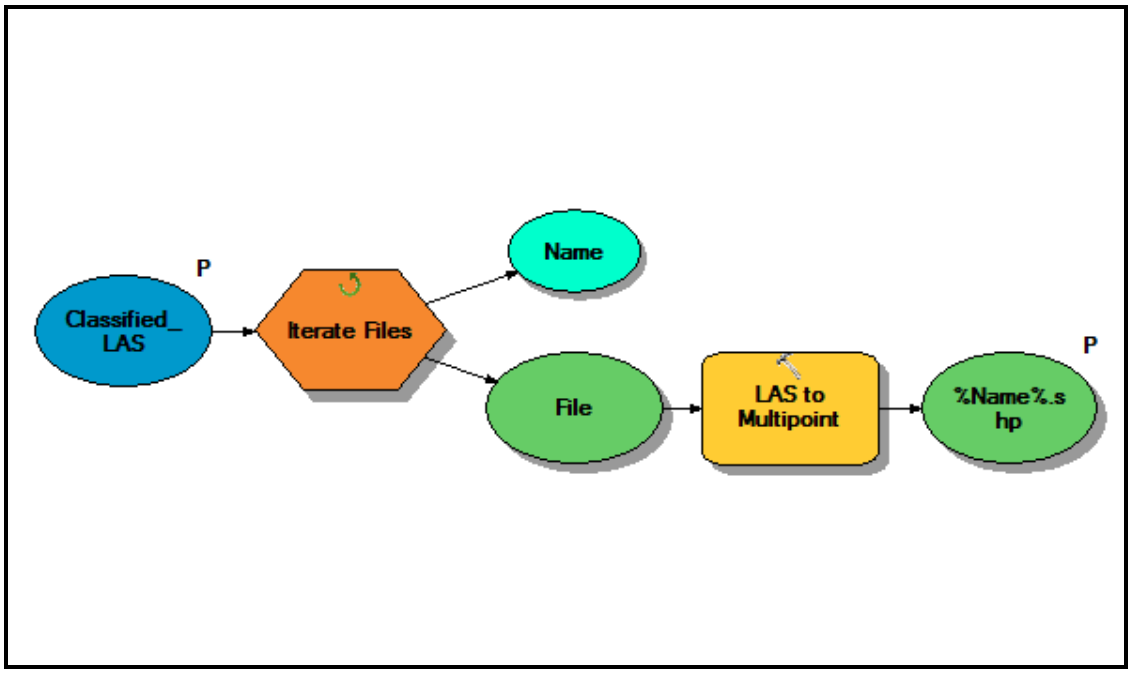

Figure 2.7 - ArcGIS Model Builder model for iterative conversion of .las files

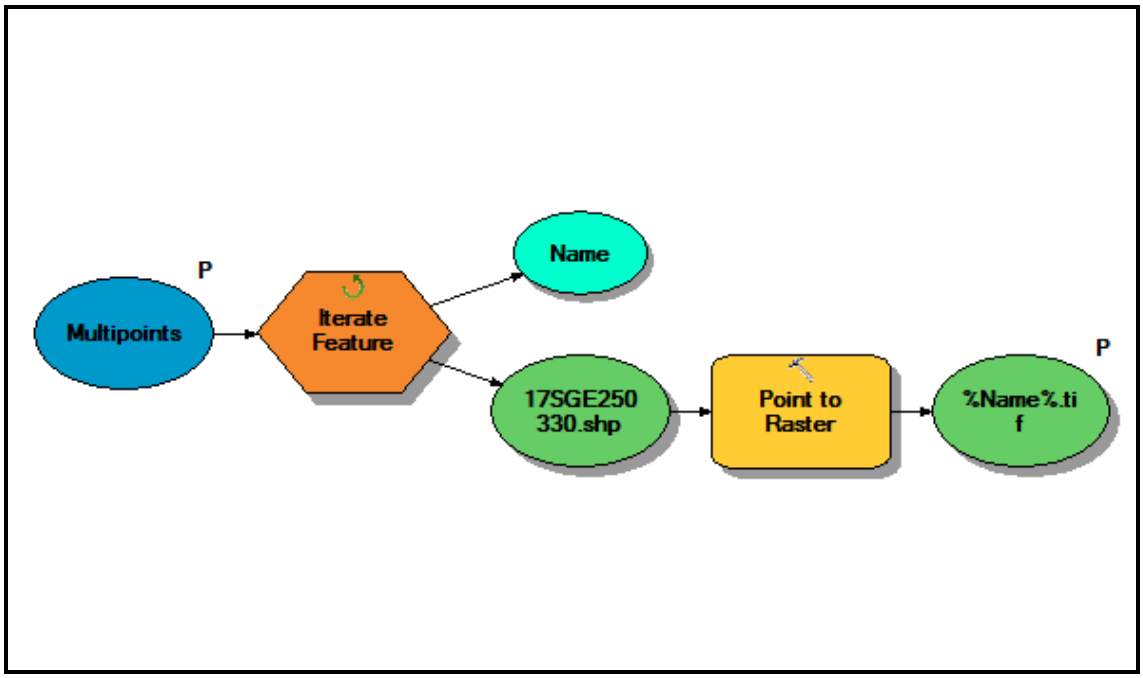

Figure 2.8 - ArcGIS Model Builder model for iterative conversion of .shp files

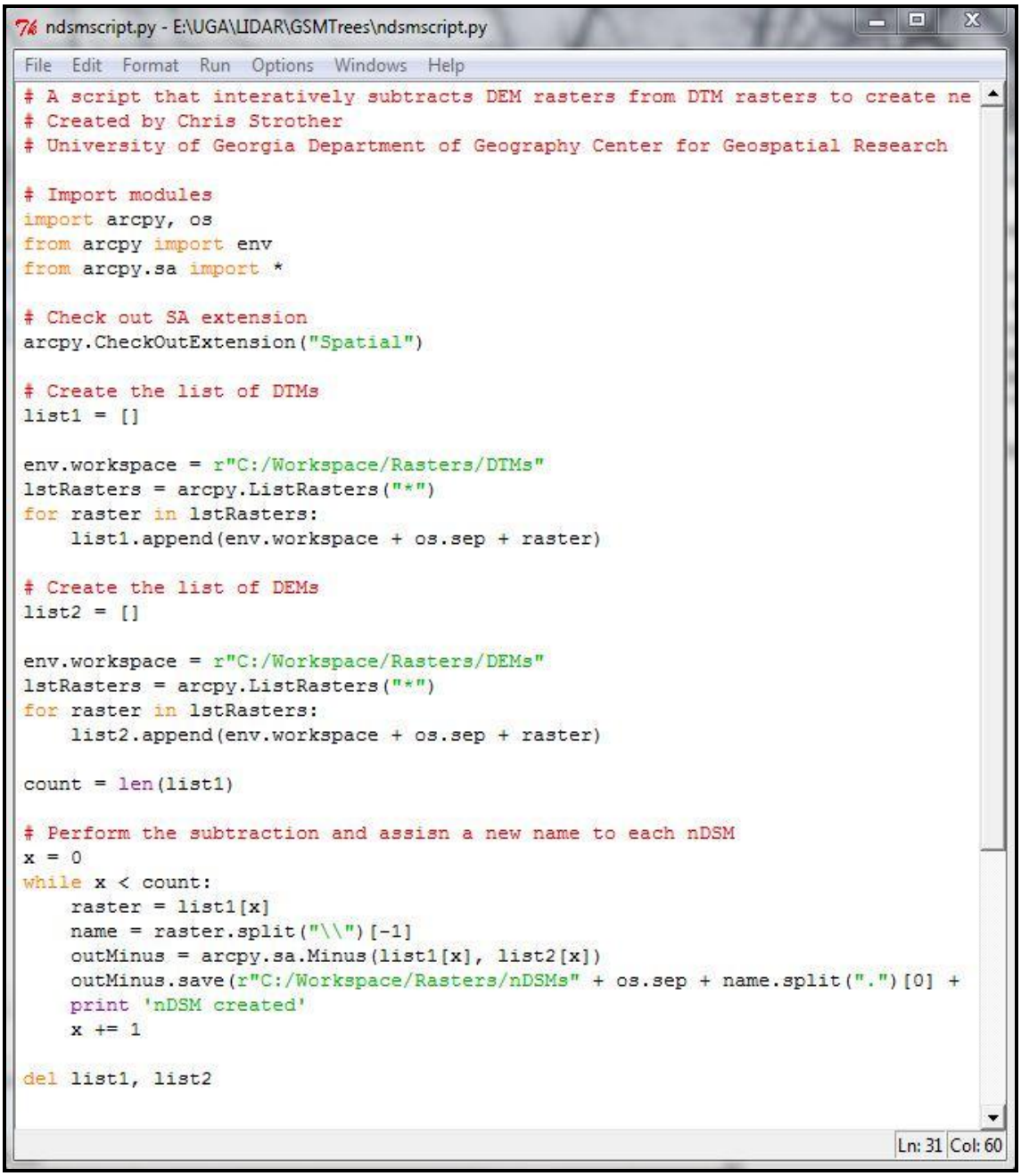

Figure 2.9 - Python interpreter window showing script created to process the subraction of 724 DEMs from 724 DTMs

Figure 2.10 - BatchnDSM software tool created using a Python script to iteratively process raster files

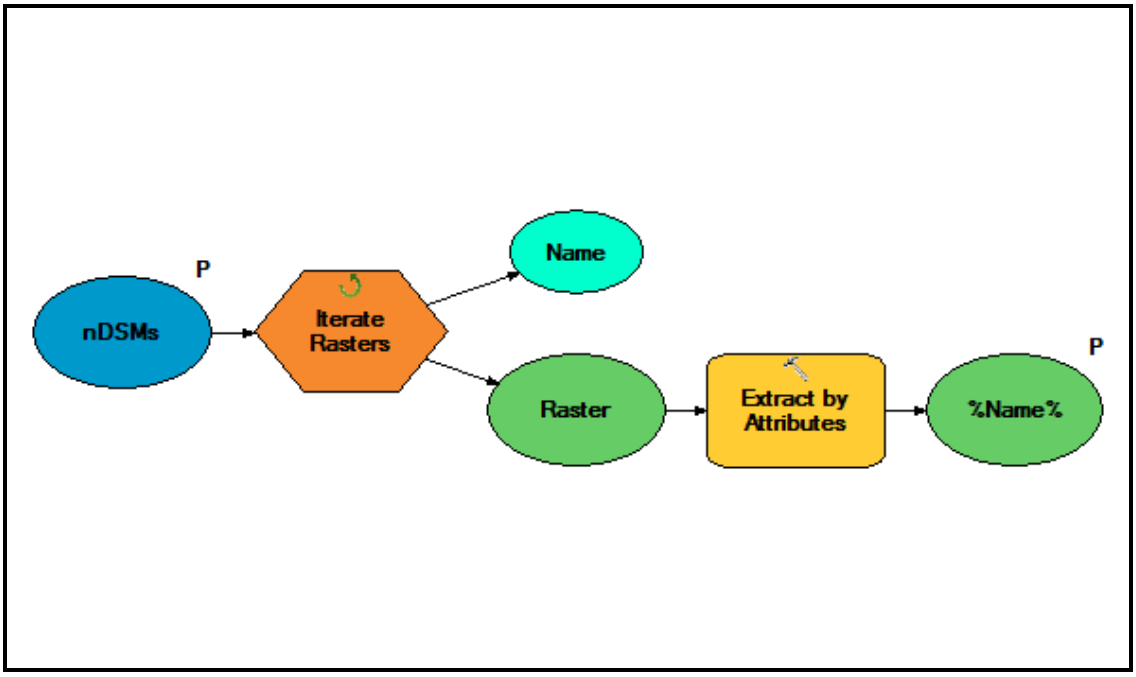

Figure 2.11 - ArcGIS Model Builder model for iterative analysis of nDSM files

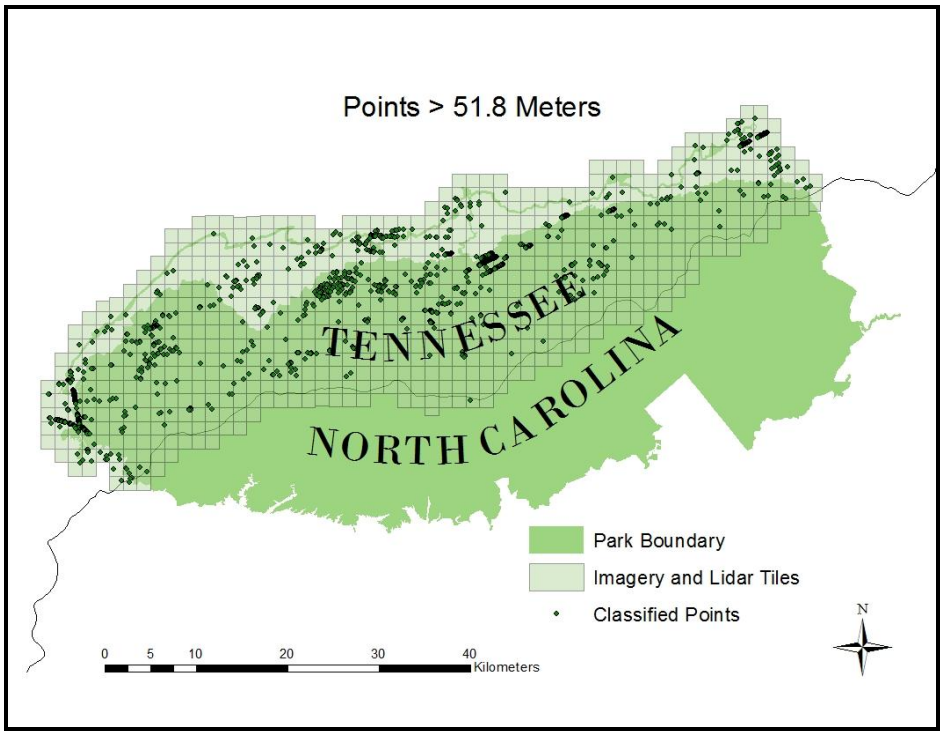

Figure 2.12 – Dataset points with heights greater than 51.8 meters

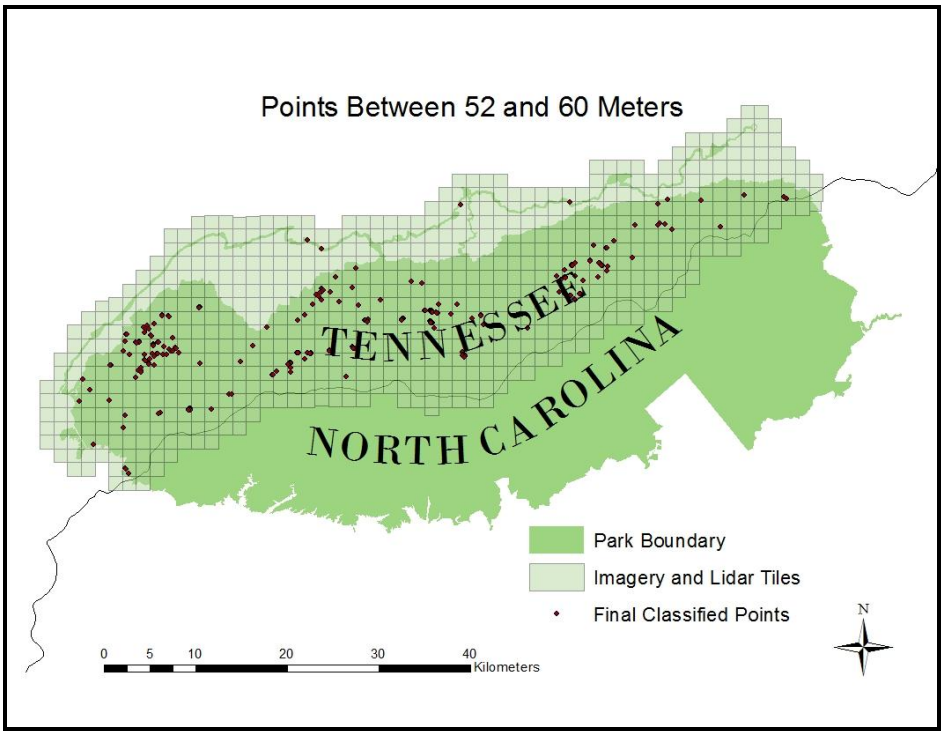

Figure 2.13 – Dataset points between 52 and 60 meters

Results and Discussion

A total of 1523 discrete points were identified from the LiDAR data with heights above ground of greater than 51.8 meters. The top ten heights were chosen and ranked based on the assumption that ten sites could be visited in a two-day field study window. All ten sites were located in the Northwestern portion of the park (Figure 2.14). It should be noted that the exact location of these trees is ecologically and historically sensitive information and can only be obtained by request to the NPS. The range in heights of trees at these ten sites is roughly 55 m to 59 m, all higher than the 52.7 m tulip poplar previously listed as the tallest tree recorded in the Tennessee portion of the GRSM as of May 2011 (Table 2.3). This finding justifies further investigation by qualified arborists for *in situ* height verification. The measurements of the ten sites conducted in the field yielded results of varying errors. The difficulties in using a hand-held instrument in such extreme conditions as well as the sometimes very extreme slope at the sites are possible sources of the error. Also, given the great heights of the objects in the field and the proximity of other trees as obstacles, it was often difficult to get an adequate distance from the measured trees to obtain the measurements. Although it is difficult to assign significant statistical importance to the errors between the field data and the LiDAR data due to the intentionally small sample size, it is worth noting that five of the seven measurements suggest that the LiDAR data underestimated the actual height which is consistent with findings of missed tree crowns described by Zimble et al. (2003).

Of the seven sites visited, none exhibited characteristics typically found in "Champion Trees". This designation is determined by a point system, calculated by adding the tree's circumference in inches to the tree's height in feet to one-fourth of the total crown spread in feet (American Forests, 2013). The trees that could be accessed did not have large trunks proportional

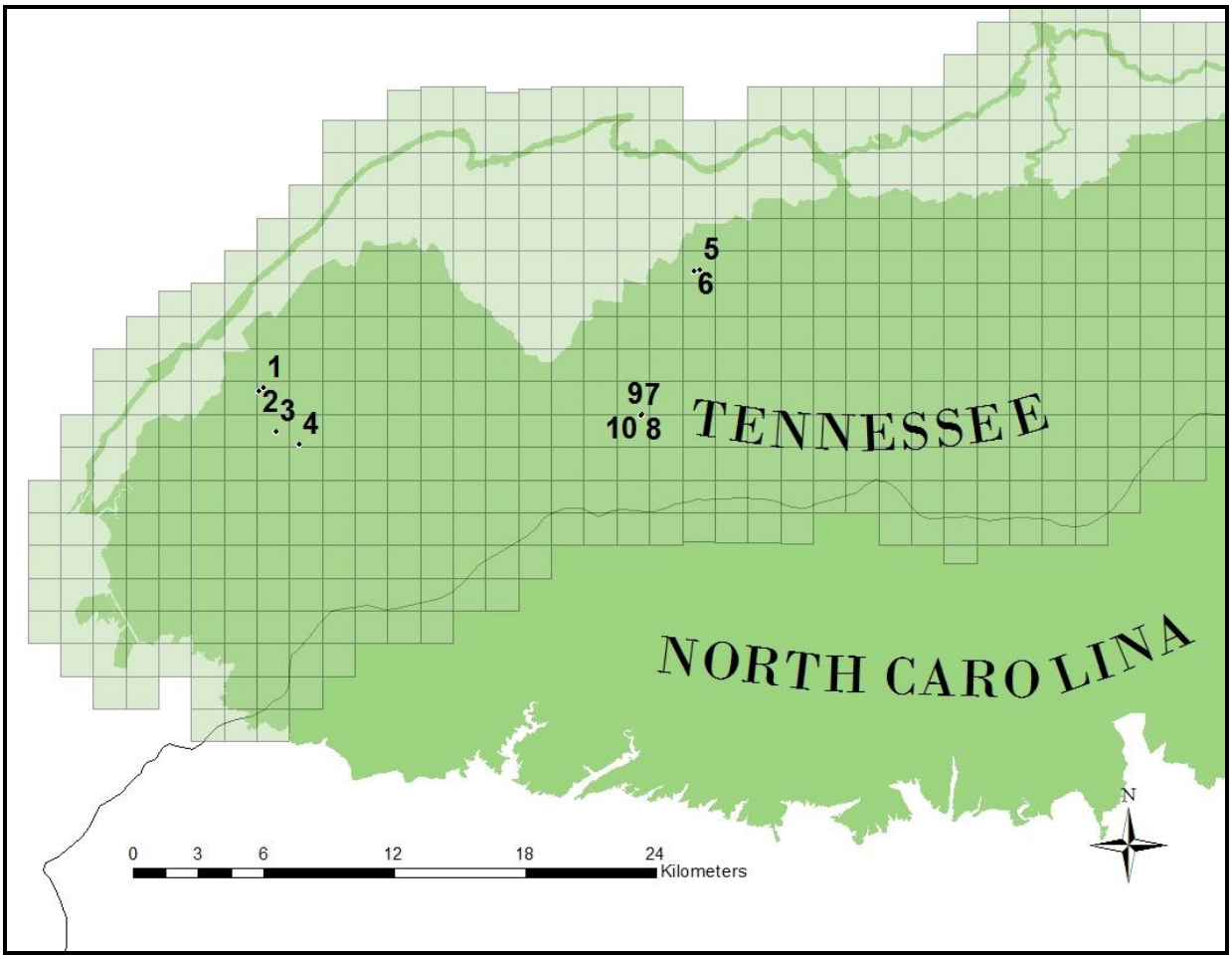

Figure 2.14 – Approximate location of ten tallest tree sites

to their height, leading to speculation that these trees are not old growth, but likely successional trees that have been growing since logging practices ended in the late 1920s and early 1930s. The favorable environmental conditions in the Southern Appalachians and geological conditions that created fertile soils in the region have allowed these trees to grow unthreatened by humans for nearly a century.

The topographical conditions of each site are also of interest. All but one site are below or very near the threshold for "low elevation" conditions described by Madden et al. (2002) as below 2500 ft. (762 m) with the average elevation at the ten sites as 554.1 m. (Figures 2.15 –

2.19). Slopes at the ten sites were varied from relatively flat conditions (10.3°) to severe (80.5°) with the average slope for the ten sites measuring 37.8°. Eight of the ten sites had a generally Northwest, North, or Northeast orientation, which is expected because these aspects present the moistest conditions for tree growth. Interestingly, the tallest potential tree site was found to have a Southwest facing aspect, usually associated with hot and dry conditions - a less favorable environment for tree growth. Based on the GRSM Vegetation Community Database created by Madden et al. (2004) as part of the USGS NPS National Vegetation Inventory program, the overstory communities at the ten sites were identified and split into three main categories: Pines (PIs = Eastern White pine successional, PI = Southern Yellow pine, PIs-T = Eastern White pine with Hemlock), Southern Appalachian Cove Hardwoods (CHx = S. Appalachian Cove Hardwoods, CHxA-T = S.A. Cove Hardwoods/Acid type with Hemlock), and Oaks (OmH/T = Submesic to mesic oak/Hardwoods with Hemlock). Five of the seven accessible sites were visually inspected and found to be hardwoods, four tulip poplar and one white oak. The remaining two trees were identified as pines. Figures 2.20 – 2.29 show point cloud representations of each tree site as visualized by the USFS FUSION/LDV LiDAR utility tool.

Table 2.3 - Tree site measurement results, terrain conditions, and measurement error

Site	Lidar Height (m)	Field Height (m)	% Error vs. Field	Elevation (m)	Degree Slope	Aspect	Overstory	Tree Type
1	59.0	Unknown	Unknown	376.3	35.1	SW	PIs-T	Unknown
2	59.0	Unknown	Unknown	358.6	51.9	N	OmH/T	Unknown
3	55.9	72.8	-30.2	494.2	10.3	E	CHxA-T	Pine
4	57.0	Unknown	Unknown	477.9	39.4	NW	PIs-T	Unknown
5	57.0	56.6	0.7	394.0	54.7	NW	PI	Pine
6	57.0	61.5	-7.9	367.4	80.5	NW	PIs	White oak
7	55.0	58.4	-6.2	785.3	28.6	NE	CHx	Tulip poplar
8	56.0	56.9	-1.6	765.0	26.8	E	CHx	Tulip poplar
9	56.0	51.6	7.9	761.1	19.0	NE	CHx	Tulip poplar
10	55.0	56.4	-2.5	761.4	31.7	N	CHx	Tulip poplar

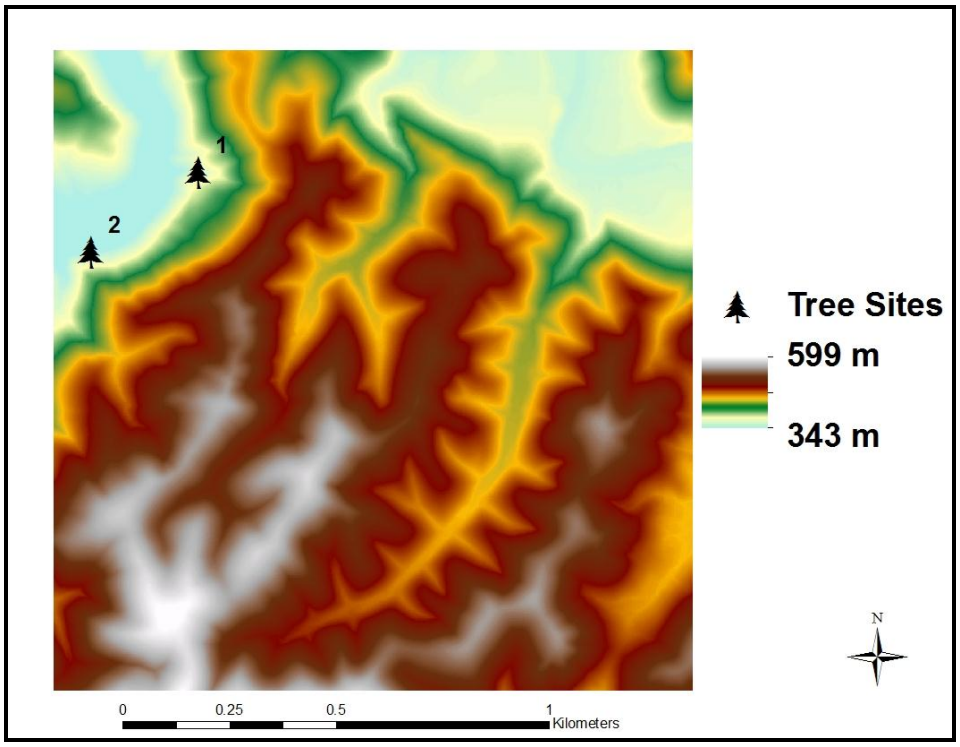

Figure 2.15 - Sites 1 & 2 elevation map

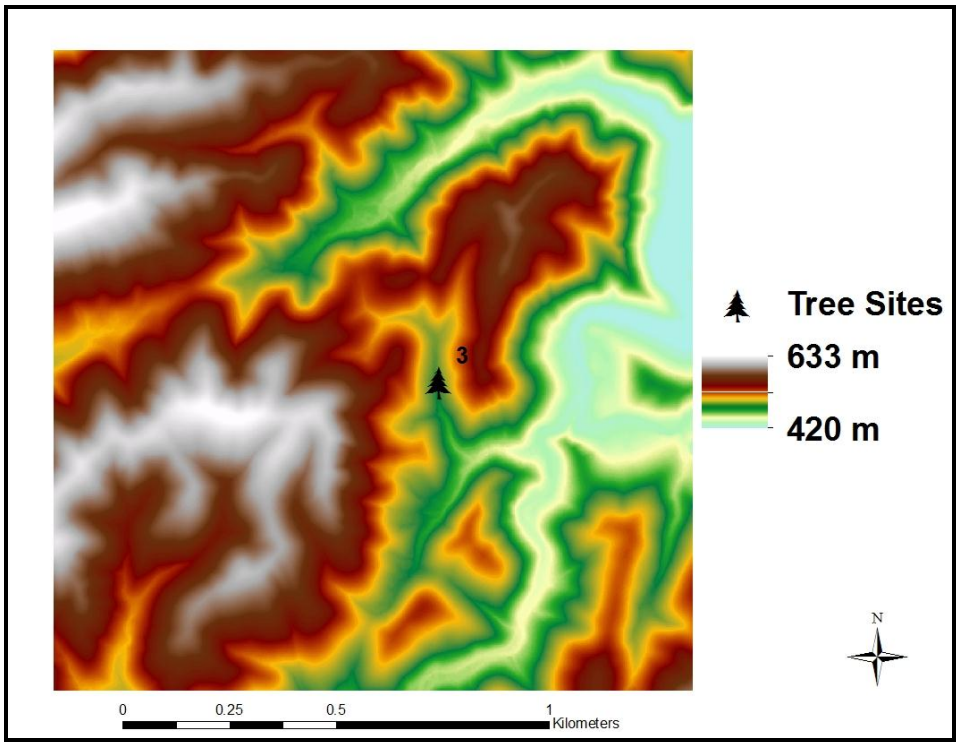

Figure 2.16 - Site 3 elevation map

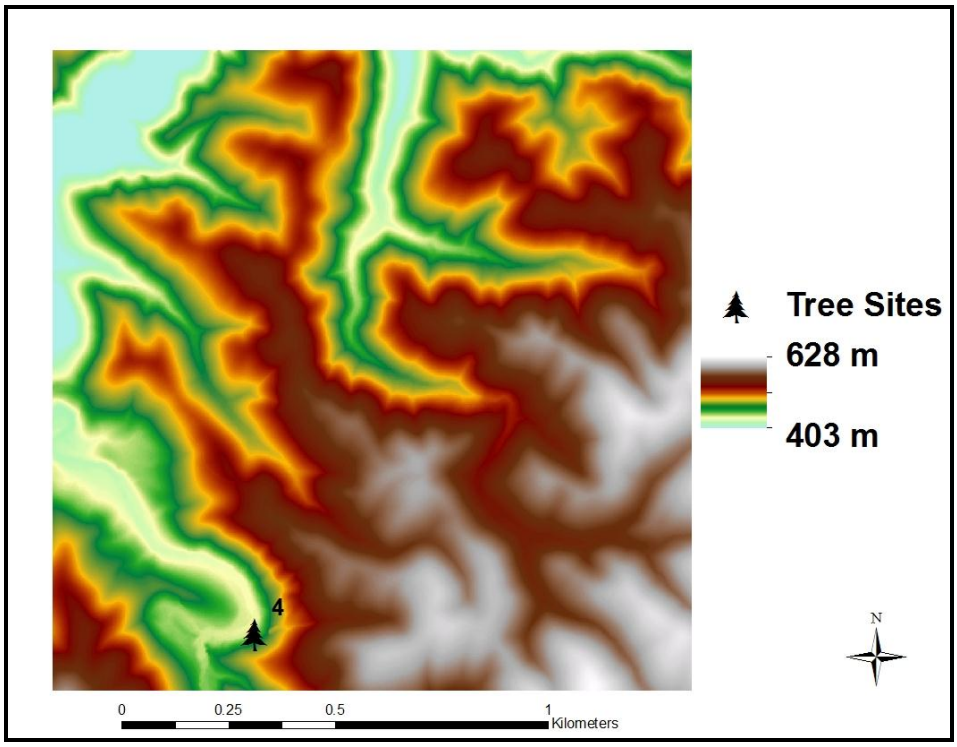

Figure 2.17 - Site 4 elevation map

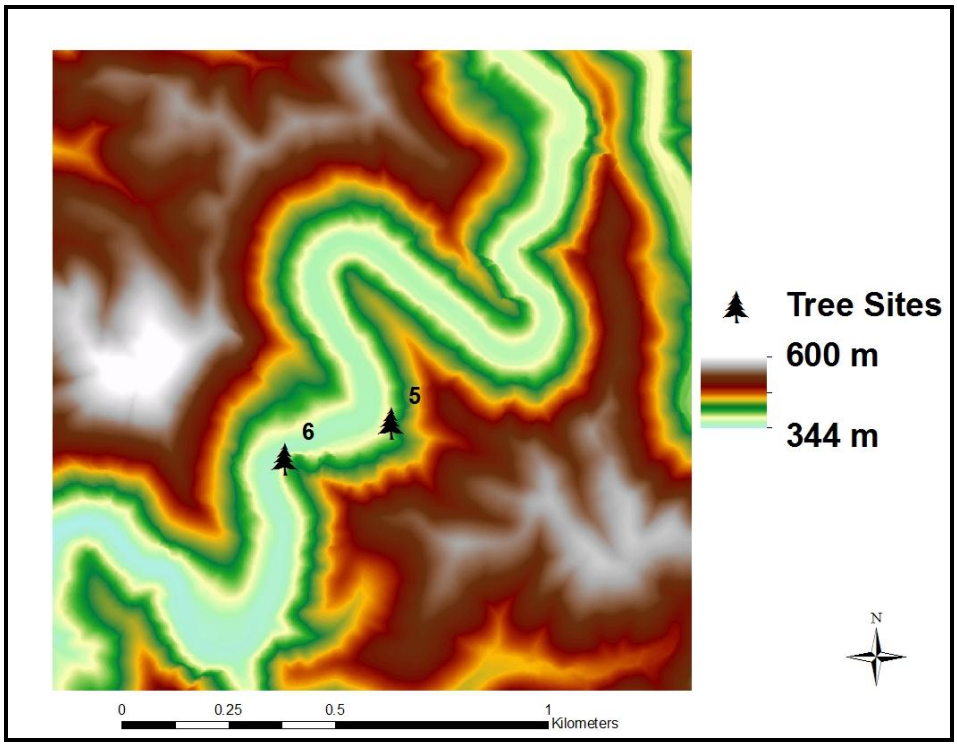

Figure 2.18 - Sites 5 & 6 elevation map

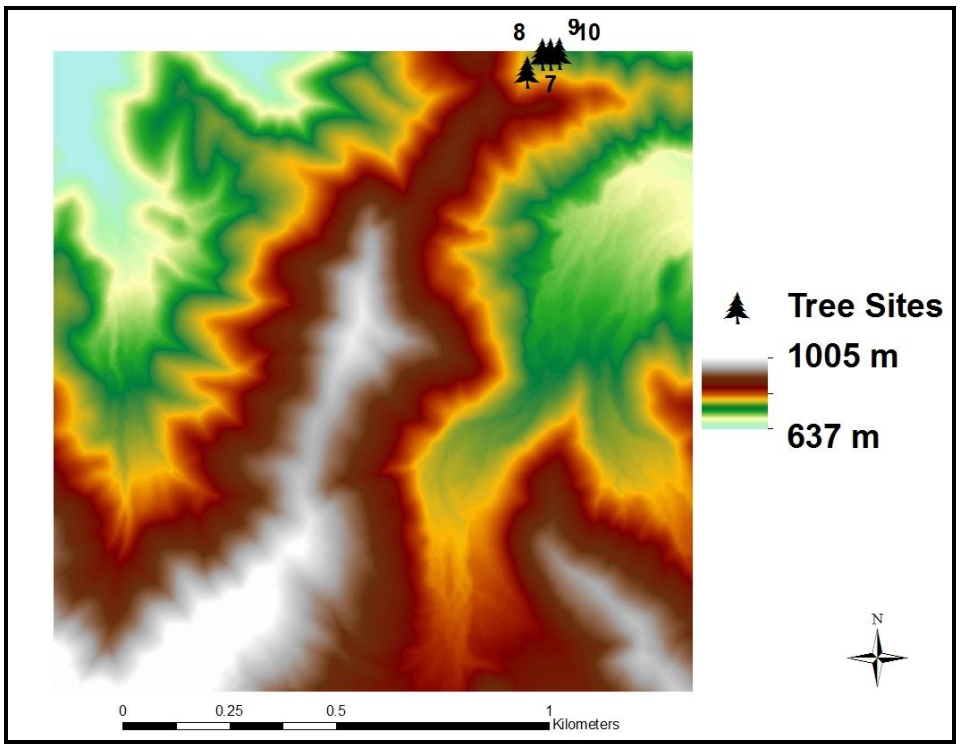

Figure 2.19 - Sites 7 - 10 elevation map

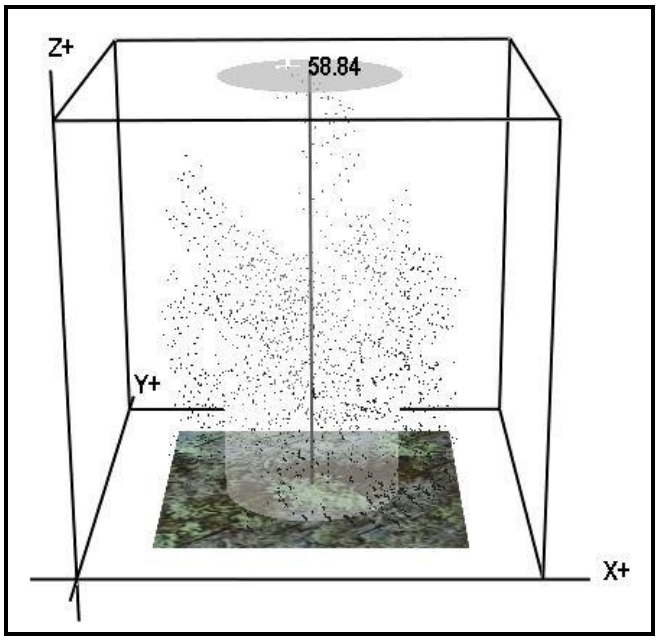

Figure 2.20 - Representation of the highest point detected in the point cloud at Site 1 with image plate (measurements are in meters)

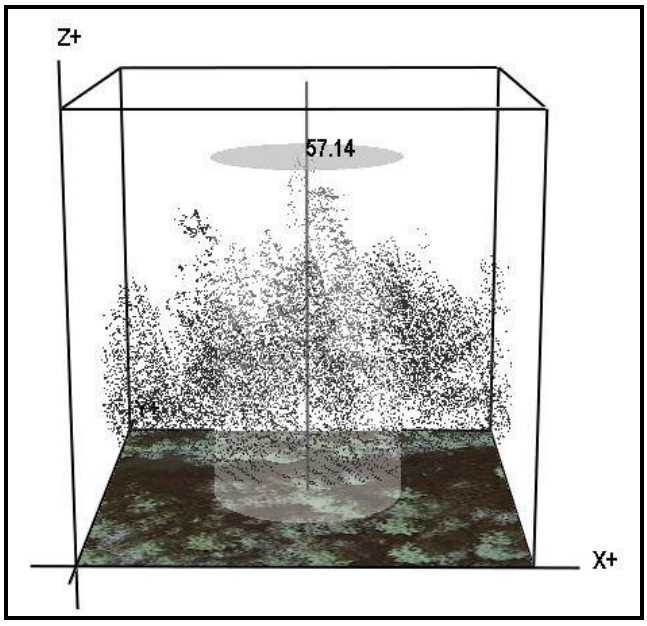

Figure 2.21 - Representation of the highest point detected in the point cloud at Site 2 with image plate (measurements are in meters)

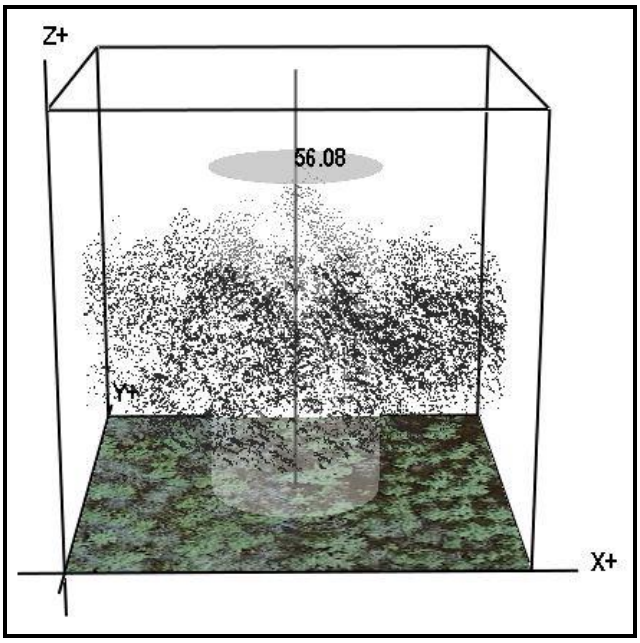

Figure 2.22 - Representation of the highest point detected in the point cloud at Site 3 with image plate (measurements are in meters)

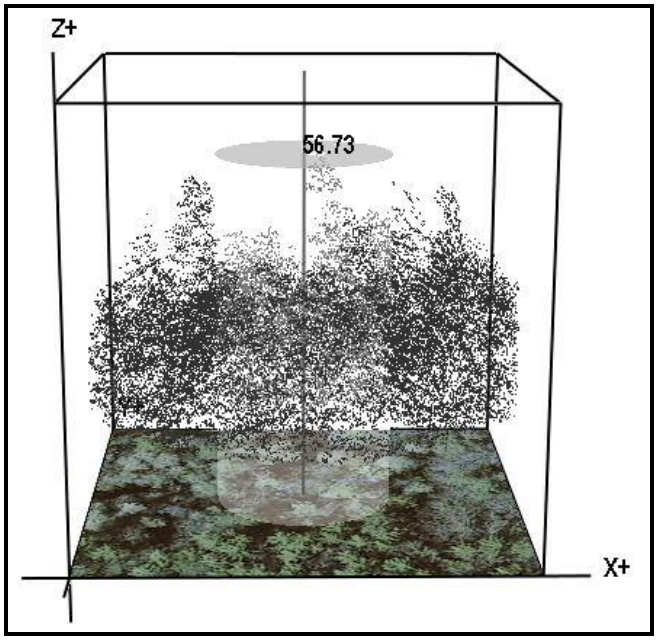

Figure 2.23 - Representation of the highest point detected in the point cloud at Site 4 with image plate (measurements are in meters)

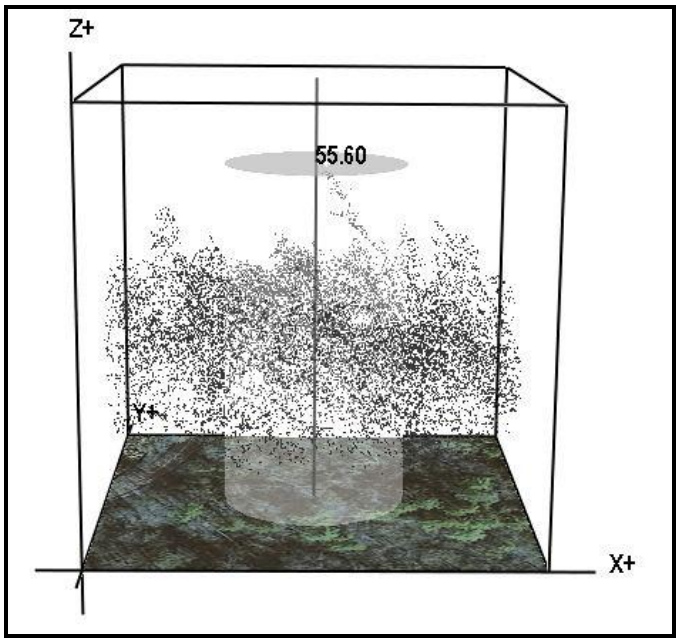

Figure 2.24 - Representation of the highest point detected in the point cloud at Site 5 with image plate (measurements are in meters)

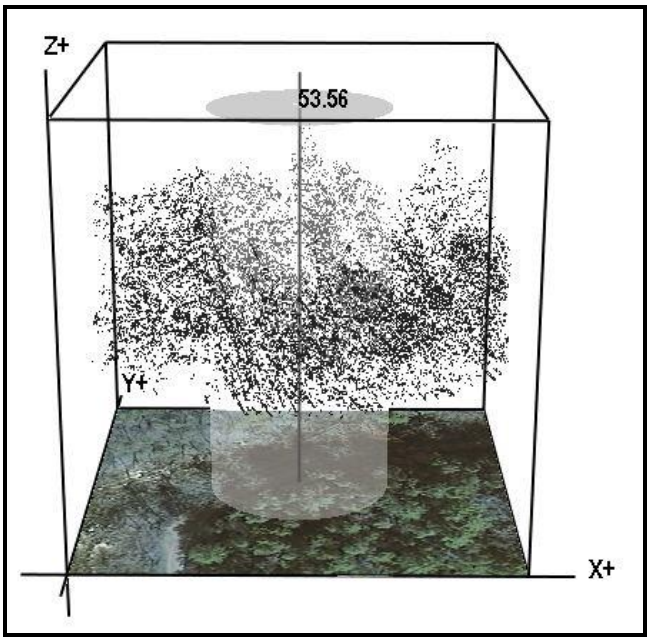

Figure 2.25 - Representation of the highest point detected in the point cloud at Site 6 with image plate (measurements are in meters)

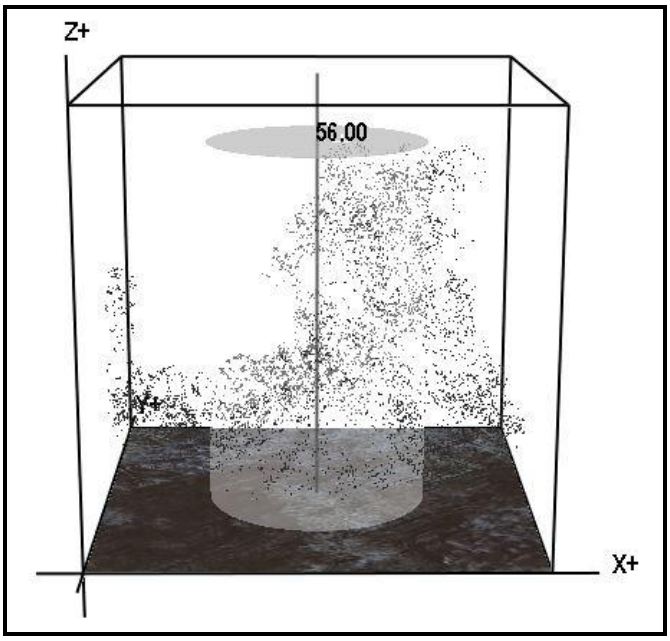

Figure 2.26 - Representation of the highest point detected in the point cloud at Site 7 with image plate (measurements are in meters)

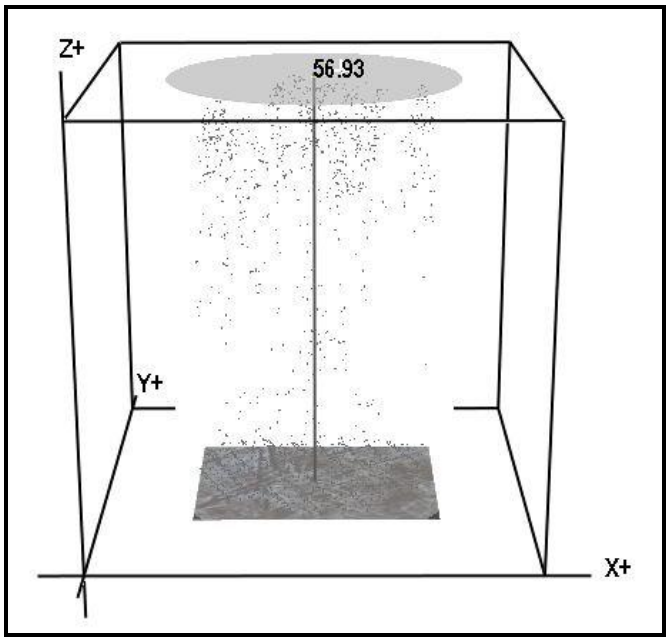

Figure 2.27 - Representation of the highest point detected in the point cloud at Site 8 with image plate (measurements are in meters)

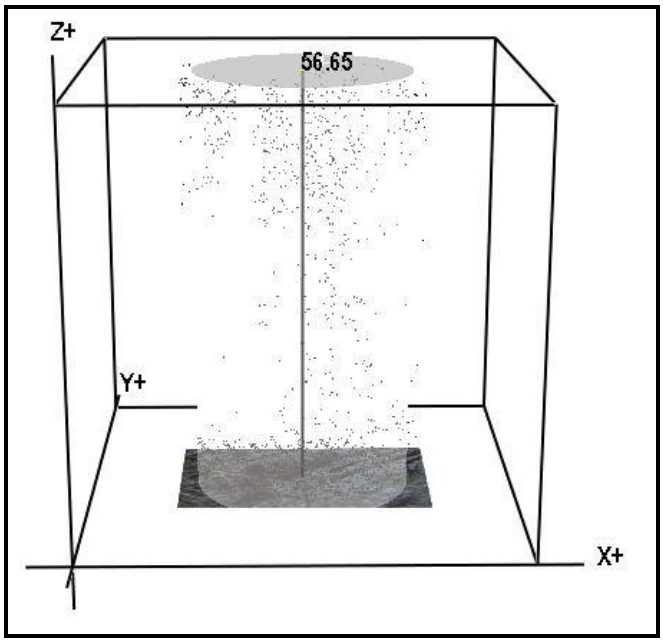

Figure 2.28 - Representation of the highest point detected in the point cloud at Site 9 with image plate (measurements are in meters)

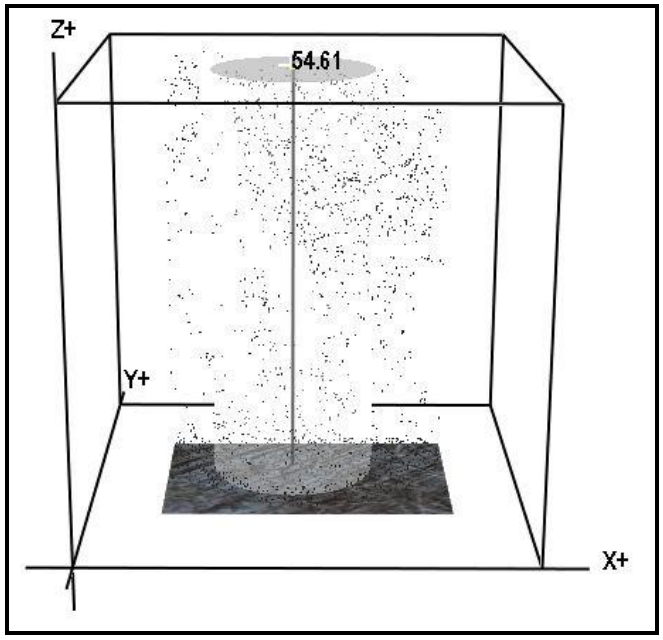

Figure 2.29 - Representation of the highest point detected in the point cloud at Site 10 with image plate (measurements are in meters)

Conclusions and Recommendations

This study demonstrates the use of LiDAR data as an effective tool for detecting individual trees of extreme height within a complex forest community in an area of rugged terrain. Using GIS and other data processing workflows, a methodology was developed to parse and query large volumes of data to obtain verifiable results. In all, ten tree sites were detected using this unique methodology with heights ranging from 55 m to 59 m. Sites 1 and 2 would be the tallest trees ever measured in the eastern U.S. if they prove to be as tall as the LiDAR data indicate (both 59 m) and all ten trees are taller than the tallest measured tree in the Tennessee portion of the GRSM (53.8 m). These ten tree sites varied in their tree species, general overstory community type and terrain slope, but had similarities in regards to elevation and terrain aspect at the tree sites. Field measurements of these individual trees were made difficult given the rugged terrain and density of vegetation at each site. Regardless, the findings of this research are sufficient to require further investigation of these tree sites, especially the three sites that could not be accessed in the field at the time of this work.

This tree site information can be passed on to arborists and park managers at the GRSM to better inform them regarding the locations of potentially historical, record-setting trees. This information can be used to promote park visitation, as well as provide specific areas for further ecological niche research. It is recommended that these ten areas be revisited by arborists skilled in accurate tree measurement techniques to verify the LiDAR derived data and further research be conducted at these sites to gather more environmental and ecological information, including soil types, soil moisture, precipitation, temperature, and understory vegetation.

CHAPTER 3

ORDINARY LEAST SQUARES ANALYSIS OF A LIDAR-DERIVED TREE HEIGHT DATABASE[2]

[2] Strother, C.W., M. Madden, T. Jordan, and S. Holloway. To be submitted to *The Professional Geographer*.

Abstract

In the examination of trees that reach extraordinary heights in Eastern forests, researchers have identified environmental factors influencing the potential maximum height that an individual tree can obtain. Access to water, sunlight, prior disturbance, topography, elevation, and available fertile soil has been hypothesized to affect the maximum tree height. This study is unique because it uses large volumes of feature data extracted from a LiDAR (Light Detection and Ranging) dataset covering a 225 hectare area of diverse Eastern Deciduous forest within Great Smoky Mountains National Park (GRSM) to correlate tree heights and environmental variables. Although relationships were significant between certain environmental variables and tree heights in the model, the variability in the dataset was not well explained. Results from ordinary least squares (OLS) regression showed $R^2 = 0.2057$, $p = 0.0000$ with a $RMSE = 7.2512$ m for the described model. This study demonstrates the use of large volumes of LiDAR derived data in showing tall trees are generally located in areas with North oriented slopes in mid-elevations that have fertile, loamy soils, and some reasonable adjacency to water.

Introduction

The increasing acquisition of airborne LiDAR data over the last two decades has created a valuable analysis tool for those interested in a variety of disciplines, including forest conservation and research (Ussyshkin and Theriault, 2011). LiDAR data provide the remote sensing analyst with the ability to create bare earth digital elevation models (DEMs) and digital surface models (DSMs) of increasingly high resolution and accuracy with relatively low costs compared to conventional field techniques (Andersen et al., 2006). These digital raster models allow the user to perform mathematical functions to determine the relative heights of objects on the Earth's surface based on the LiDAR point cloud data. The use of airborne LiDAR to estimate

forest parameters has been shown to have some advantages to traditional photogrammetry because the data do not require further orthorectification (geometric correction) and can produce results with accuracies within a few centimeters, both vertical and horizontal (Suárez et al., 2005). Through the use of software-based extractive algorithms, large amounts of data regarding canopy structure can be acquired. The increase in computer processing capabilities, along with the prevalence of the computationally large data sets that LiDAR sensors collect, gives researchers in forestry research using LiDAR data a great opportunity to increase the number of observations available for statistical analysis versus traditional field measurement techniques. However few studies have fully investigated the effects of extremely large sample sizes with traditional statistical analyses. The over twenty thousand unique observations used in this particular study would have taken field researchers months to possibly years to acquire at prohibitive costs. These observations can be collected in a database and then used to create a multivariate model to determine if this type of collection is useful in predicting tree heights based on environmental conditions.

Canopy height of a forest is a complicated metric to measure and predict given the number of factors that contribute to an individual tree's growth. It is a useful metric, however, because it provides a suitable estimator of forest carbon stock as well as contributing to natural resources management and fire fuel modeling (Kenyi et al., 2009). Among the factors shown to have an effect on tree growth are ground elevation (Coomes and Allen, 2007; Petit et al., 2011) soil characteristics, light, water, and physical characteristics (Kozlowski, 1971). Past studies using LiDAR remote sensing have focused primarily on accurately measuring maximum and mean canopy heights and modeling forest parameters such as timber volume, basal area, and

stem density from these data (Nilsson, 1996; Naesset, 1997; Maune, 2001; Jensen et al. 2006; Magnussen et al., 2010).

This work used 22187 observations of tree height values extracted from a LiDAR data set as well as other GIS layers to model the environmental variables that contribute to tree heights in a mountainous forested landscape with a high variability of overstory species. An Ordinary Least Squares (OLS) model was used because of its unbiased characteristics. It is proposed that the influence of increasing elevation will have a negative impact on tree heights in the model due to environmental conditions such as low temperatures and high winds (Madden et al., 2004). Increased distance from riparian features is also predicted to produce lower tree heights because of decreasing soil moisture. It is also proposed that increasing slope percentage will predict lower tree heights because of the difficulty of seedling establishment. The presence of Ditney-Unicoi and Soco-Stecoah soils (low in plant macro-nutrients), and a general Southern exposure (drier conditions due to sunlight) are also predicted to reduce modeled tree heights as well. Conversely, low slope percentages, lower elevations, shaded Northern exposures, close proximity to water, and the presence of highly productive Spivey-Santeetlah soils should predict taller tree heights (NRCS, 2009).

A successful multivariate, ordinary least squares (OLS) model of these parameters will allow park managers, foresters, and conservationists to predict the expected recoverable carbon stock volume in similar conditions after an event such as fire, flood, illegal logging, or destructive pest invasion. This study will provide a thorough multivariate statistical analysis to determine the effectiveness of creating such a model.

Study Area

The study area (Figure 3.1) is a 225-ha area of mountainous terrain located in the Great Smoky Mountains National Park (GRSM) in Tennessee, U.S. The park contains almost continuous forest cover and is one of the most biologically diverse forests in the world. The subset study area within the park was chosen because potentially record setting trees were known to grow there and site conditions included the presence of water sources, the large range of elevation values, the large variation in slope percentages, as well as the presence of a nutrient rich and a nutrient poor soil type. There are two predominate fluvial features in the study area known as False Gap Prong and Wooly Tops Branch, respectively. The area is bounded by the NAD 83 UTM Zone 17N coordinates: 285000 m to 286500 m Easting and 3951000 m to 3952500 m Northing and all analyses were performed using the UTM projection. The study area contains 502 m of relief between 662 m and 1164 m (Figure 3.2) and contains low to extreme slopes of 0.14 to 76 percent (Figure 3.3). The forested landscape is comprised of ten major overstory vegetation communities (Figure 3.4)– Low Elevation Mixed Pine-Xeric Oak, Montane Alluvial Forest, Montane Grape Vine Opening, Southern Appalachian Cove Hardwoods, Southern Appalachian Early Successional Hardwoods, Southern Appalachian Heath Balds, Southern Appalachian Mixed Hardwoods without Oaks, Southern Appalachian Mixed Hardwood, Acidic, Southern Appalachian Northern Hardwoods, and Submesic to Mesic Oak/Hardwoods as described by Madden et al. (2004). The area contains three major soil types – Ditney-Unicoi, Soco-Stecoah, and Spivey-Santeetlah (Figure 3.5). A three-dimensional (3D) fly-through animation of the study area can be viewed at http://strother.myweb.uga.edu by clicking on the Research tab on the home page.

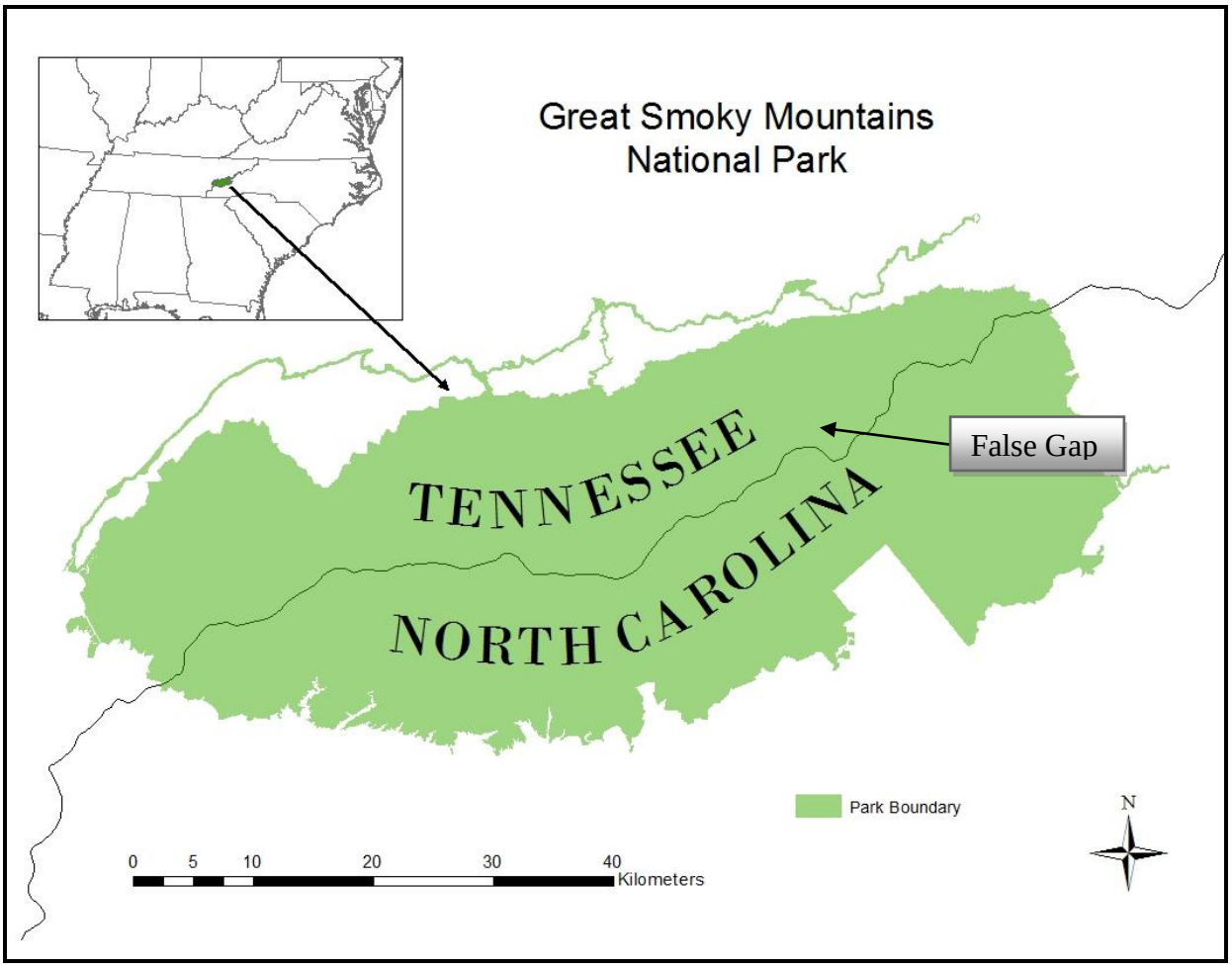

Figure 3.1 - False Gap study area within the GRSM

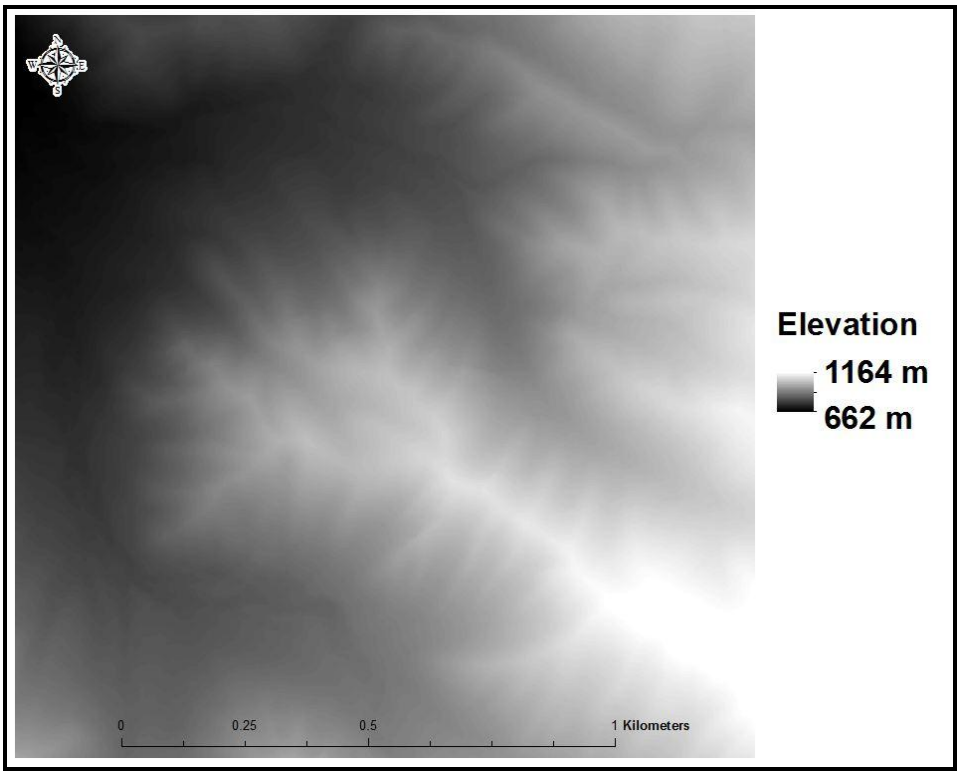

Figure 3.2 - Elevation raster of False Gap

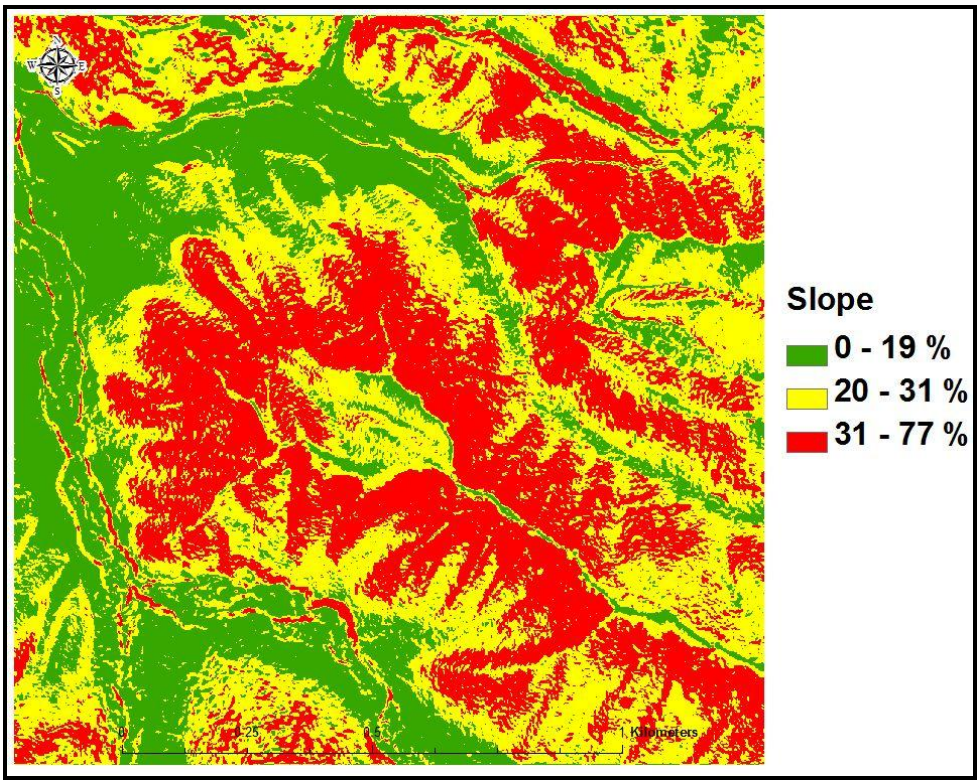

Figure 3.3 - Slope raster of False Gap

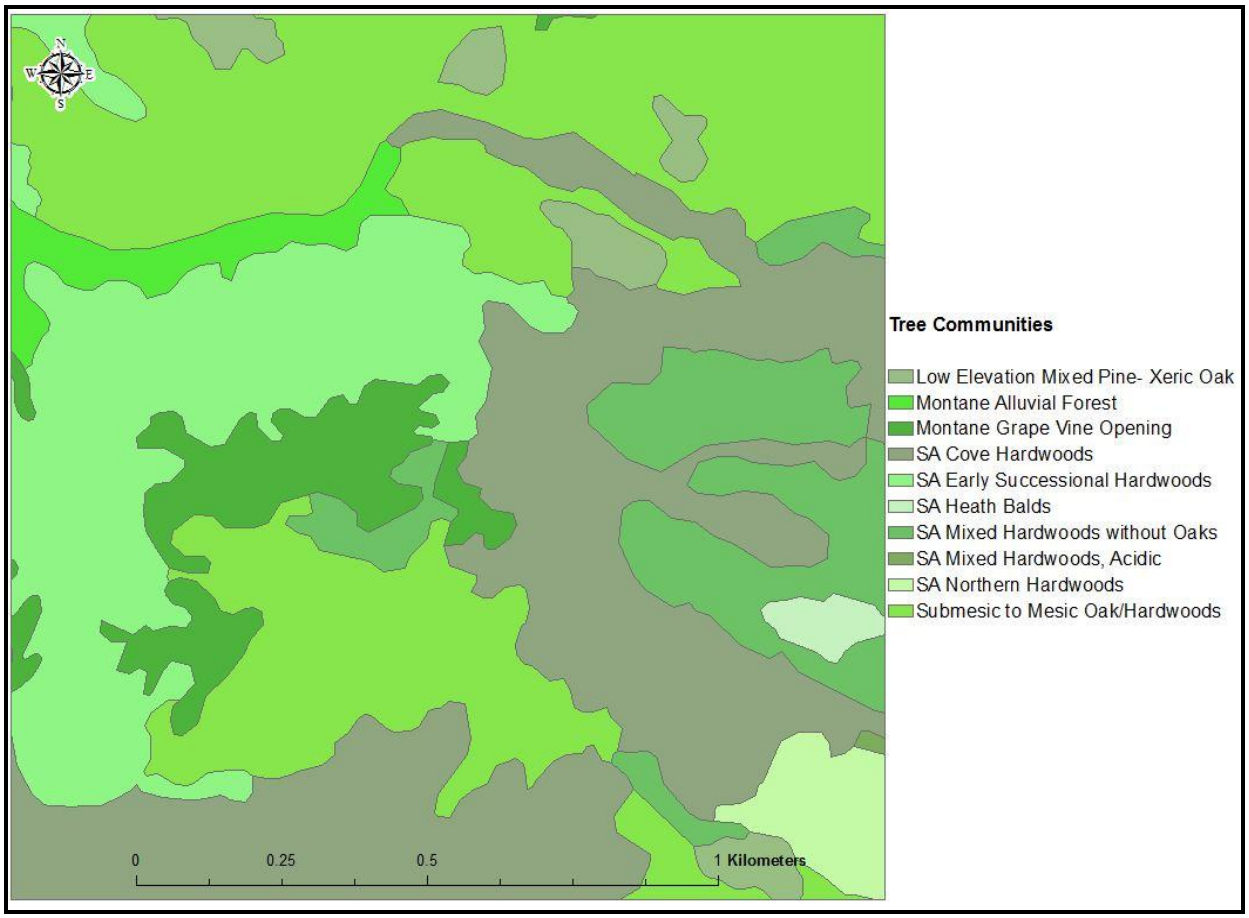

Figure 3.4 - Overstory vegetation map of False Gap (Madden et al., 2004)

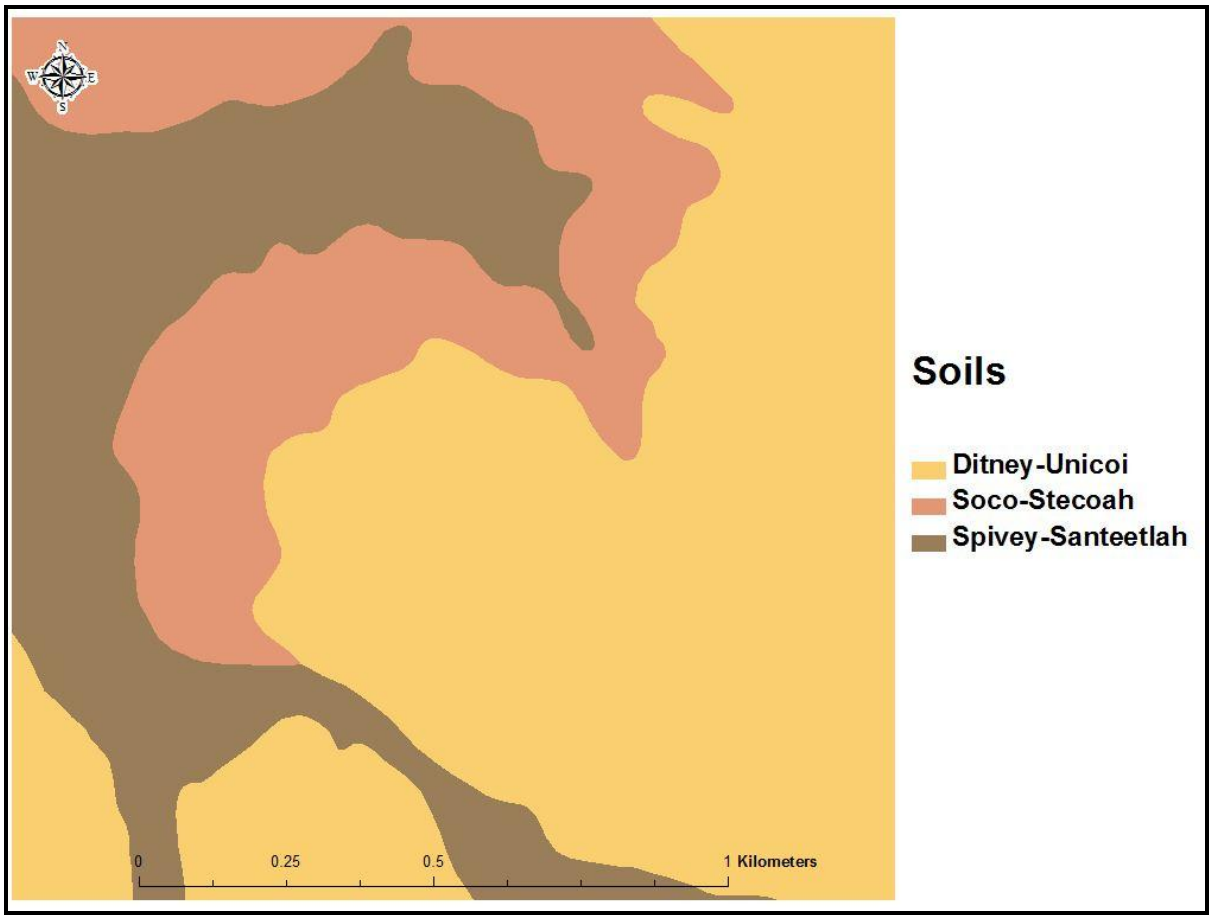

Figure 3.5 - Soil map of False Gap (NRCS, 2011)

Methods

Between February 15 and April 7, 2011, LiDAR data were collected for the U.S. Geological Survey (USGS) to be added to The National Map. The data were collected in leaf-off and snow-free conditions for the Tennessee portion of the GRSM as well as the Foothills Parkway with grants awarded to The University of Georgia (UGA) and Gainesville State College (GSC) funded by the American Recovery and Reinvestment Act (ARRA) of 2009 (Jordan & Madden, 2011) A total of 111 flight lines were used to ensure complete coverage of the study area (Figure 3.6). The LiDAR data used in this study were acquired by Photo Science, Inc. using a Leica ALS-60 sensor and an Optech ALTM Gemini sensor. Both sensors used a laser with a wavelength of 1064 nm. The flying height for the ALTM Gemini mission was reported to be

1981.2 m above ground level (AGL), with a scan angle of ±16°, scan frequency of 20.2 Hz, and a laser pulse frequency of 50 kHz. The flying height for the ALS-60 mission was not listed, the scan angle was ±16°, the scan frequency was 34 Hz, and the pulse frequency was 53 kHz (Table 3.1).

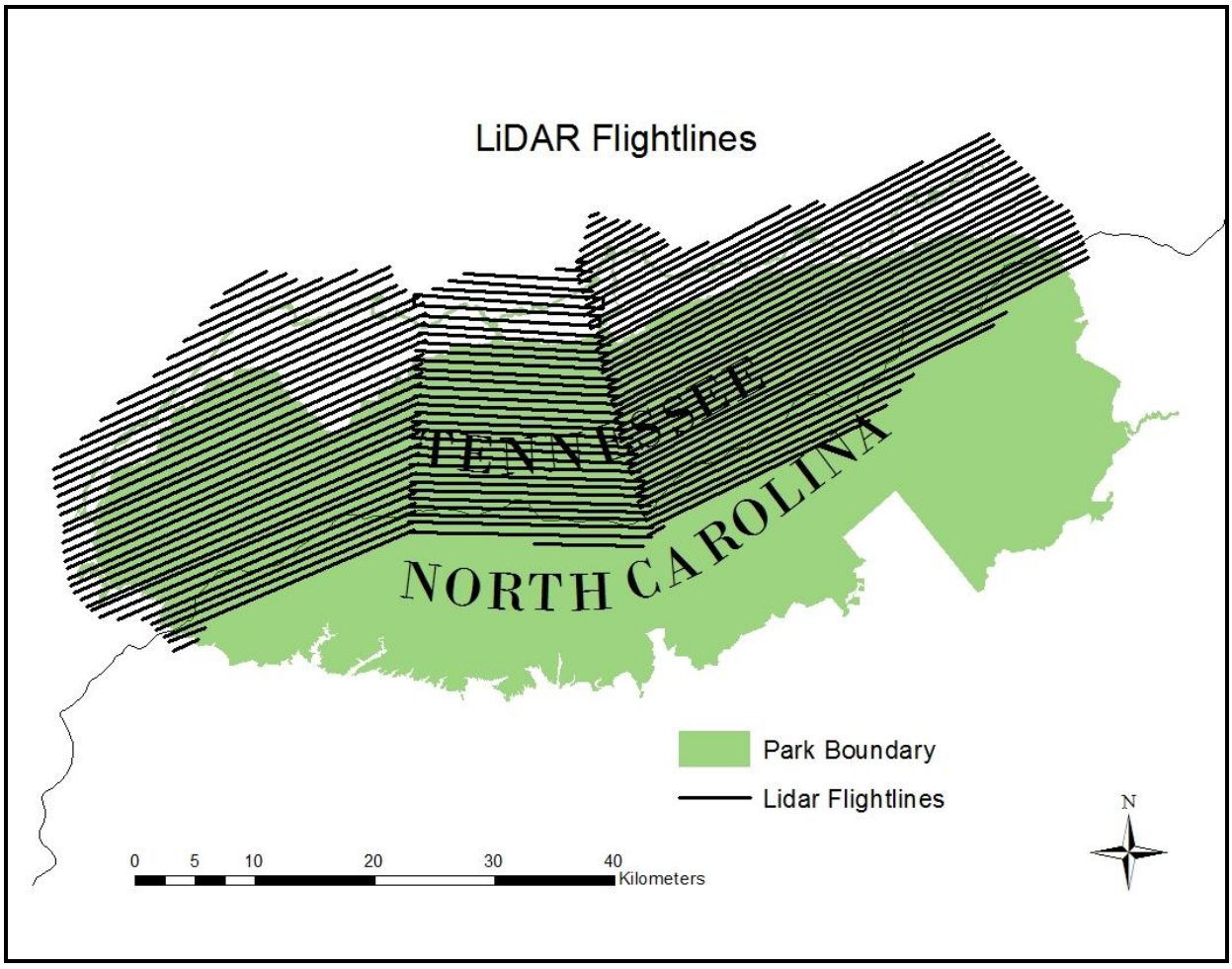

Figure 3.6 - Mission flight lines for LiDAR collection in early 2011

Table 3.1 – Photo Science, Inc. LiDAR sensor acquisition data

Sensor	**Optech ALTM Gemini**	**Leica ALS-60**
Altitude (AGL)	1981.2 m	Not listed
Speed	110 knots	150 knots
Scan frequency	20.2 Hz	34 Hz
Scan angle	± 16°	± 16°
Pulse frequency	50 kHz	53 kHz

High resolution (30-cm pixel size), four-band color infrared (CIR) imagery was also collected of the same study area by the same commercial vendor, Photo Science, Inc. The Center for Geospatial Research (formerly Center for Remote Sensing and Mapping Science - CRMS) in the Department of Geography at UGA was involved in the post-processing and quality control checking of the raw LiDAR data and orthoimagery used in this work.

The LiDAR data file used in this study contains 3,601,691 points with point spacing of 0.79 points / m^2. The vertical accuracy was reported as ± 0.165 m RMSE, which meets the standard of < ± 0.18 m RMSE required by the USGS. A 1.5-m bare Earth DEM was derived from the ground return points of the LiDAR data file and was interpolated using the Nearest Neighbor method by the vendor, Photo Science, Inc. The stream layer was sourced from the USGS National Hydrology Dataset (NHD) and the vegetation overstory layer was created by the CRMS as part of the USGS – National Park Service National Vegetation Inventory Project by manual interpretation of large scale CIR aerial photographs acquired in 1997 / 1998 (Welch et al., 2002; Madden et al., 2004). The state layer was sourced from the USGS National Atlas website and the soils layer from the United States Department of Agriculture's (USDA) Natural Resources Conservation Service (NRCS) Web Soil Survey (Table 3.2).

Table 3.2 – Description of data used in the study

Data Layer	**Source**	**Format**	**Resolution**
Bare Earth DEMs	UGA CRMS 2011	.img	1.5 m
LiDAR point cloud files	UGA CRMS 2011	.las	0.69 m
State boundary	nationalatlas.gov 2005	.shp	
Streams	USGS 2004	.shp	
Forest overstory communities	UGA CRMS 2004	.shp	
Soils	USDA 2012	.shp	

The first step in the analysis was to perform a feature extraction using a software extension for ArcGIS® 10 called Lidar Analyst. This tool required input from the user regarding the type of forest of interest. A fixed window search was recommended for dense forests. The "deciduous" forest option was chosen given the overstory designations found in the reference layer by Madden et al. (2004). The resulting points created from the algorithm yielded 22187 trees (Figure 3.7) and created an attribute table with the tree characteristics: height, crown width, and stem density based on allometric algorithms within the software (Overwatch Systems, Ltd., 2010).

Next, raster layers were created in ArcGIS to represent percent slope and aspect (Figure 3.8) using the elevation layer as the source. The stream, overstory, and soil layers were added along with the tree point layer for analysis in the GIS. Spatial data joins were used to extract the values of each layer to the tree point layer and an attribute table was created that contained the values for tree height (in meters), slope percentage, aspect azimuth, elevation (in meters), soil type, and overstory vegetation type. A Euclidean distance operation was performed on the tree

points to determine the linear distance to features in the stream layer and an attribute field was added to the tree point layer with the distance in meters.

The GIS attribute table was then exported in text format (.txt) so it could be imported into a spreadsheet for further manipulation. Once opened in a spreadsheet, a formula was created to categorize the aspect azimuth values into eight cardinal directions, plus a category for flat aspects. Cardinal directions were defined as North, Northeast, East, Southeast, South, Southwest, West, and Northwest with each category including 45° of azimuth. Next, formulas were entered to categorize the original soil designations into three main categories. This was decided after research of the qualities of the six original soil types and determining that the finer resolution in the soil types was due to difference in slope. For example, the soils of the Ditney-Unicoi (DtD and DtF) complex were identical except for the slope percentage in the area where they are found. The final spreadsheet was exported in a comma separated value format (.csv) for compatibility with the Stata IC 10 statistical analysis software package.

Once the data were imported for statistical analysis, a decision was made to exclude observations that fell into the Montane Grape Vine Opening and Heath Bald overstory categories based on information reported by Madden et al. (2004) regarding the species found there. The Heath Bald communities are general on rock outcrops and contain shrub species of Rhododendron (*R. catawbiense* and *R. carolinianum*) whose low-growing characteristics would affect the model performance. Montane Grape Vine Openings are areas where vines (*Vitis aestivalis)* grow over the canopy and weight down the overstory creating lower tree heights than would be found with their absence. The exclusion of these data resulted in reducing the sample to 20,708 tree height observations.

To create models for analysis, independent categorical variables had to be created for the aspect, overstory vegetation, and soil type categories. The dummy coded categorical values were used along with the independent continuous variables representing elevation, slope percentage, and distance to riparian features, to create the initial model for the dependent variable - tree heights. A list of the variables, their aliases, and their types is provided in Table 3.3. Decisions were made regarding which dummy variables in each category would be used as the reference variable. For Model 1, it was decided that the variable with the smallest percentage of frequency in the observations would be the reference variable (Tables 3.4-3.6). For the soils, this was the Spivey variable (s3); aspect - the East variable (a3); overstory - SA Mixed Hardwoods, Acidic variable (n6). Basic summary statistics were also created for the continuous variables (Table 3.7).

Table 3.3 – Dependent and independent variables used in Models 1 and 2

Variable alias	**Variable value**	**Type**
treeheight	Tree height (m)	Dependent; continuous
elevation	Elevation (m)	Independent; continuous
near_dist	Distance to streams (m)	Independent; continuous
slope	Slope %	Independent; continuous
s1	Ditney-Unicoi soil	Independent; categorical
s2	Soco-Stecoah soil	Independent; categorical
s3	Spivey-Santeetlah soil	Independent; categorical
a1	North aspect	Independent; categorical
a2	Northeast aspect	Independent; categorical
a3	East aspect	Independent; categorical
a4	Southeast aspect	Independent; categorical

a5	South aspect	Independent; categorical
a6	Southwest aspect	Independent; categorical
a7	West aspect	Independent; categorical
a8	Northwest aspect	Independent; categorical
n1	Low elevation Mixed Pine-Xeric Oak	Independent; categorical
n2	Montane Alluvial Forest	Independent; categorical
n3	SA Cove Hardwoods	Independent; categorical
n4	SA Early Successional Hardwoods	Independent; categorical
n5	SA Mixed Hardwoods w/o Oaks	Independent; categorical
n6	SA Mixed Hardwoods, Acidic	Independent; categorical
n7	SA Northern Hardwoods	Independent; categorical
n8	Submesic to Mesic Oak/Hardwoods	Independent; categorical

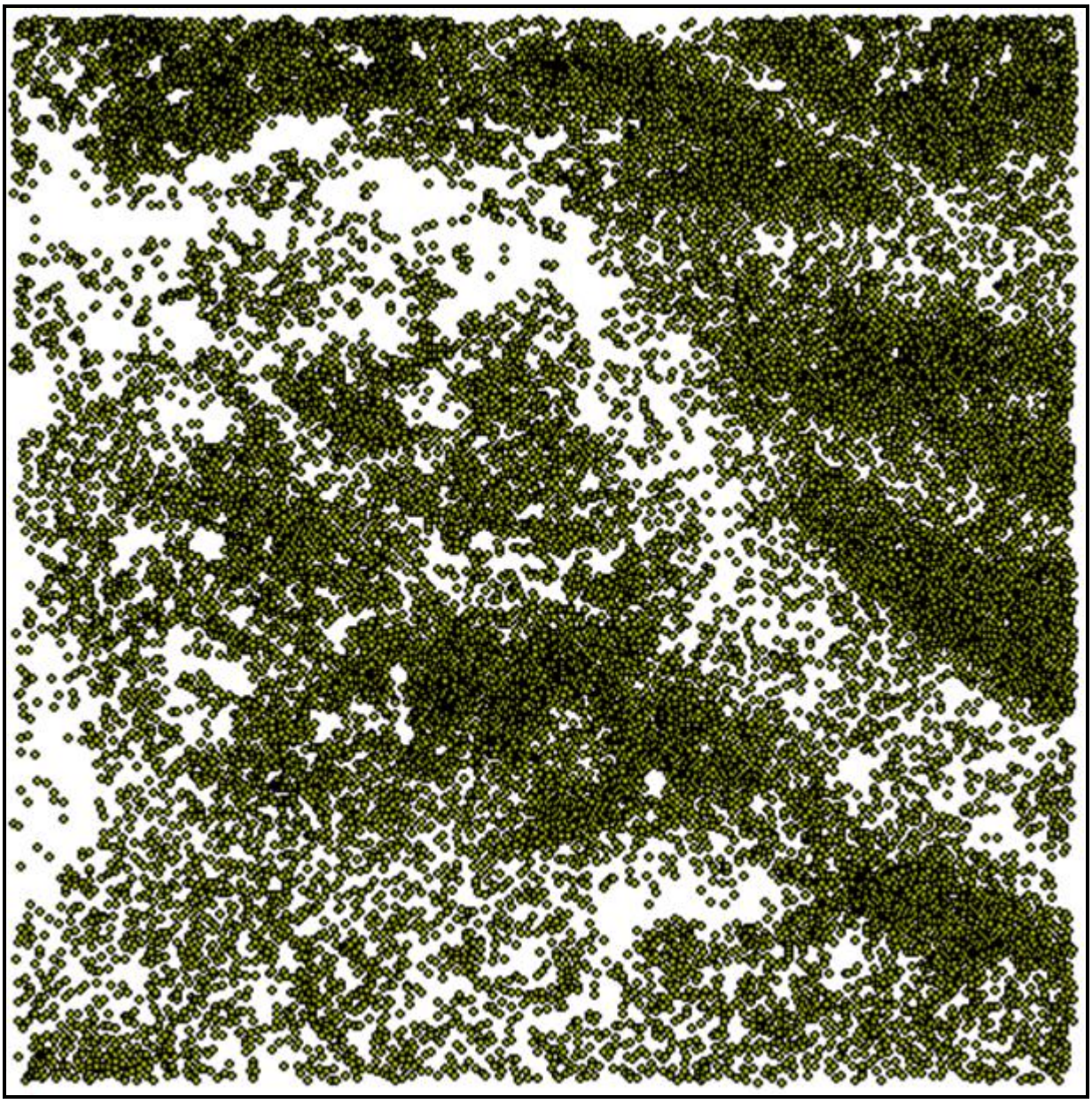

Figure 3.7 - Tree points extracted by Lidar Analyst algorithm

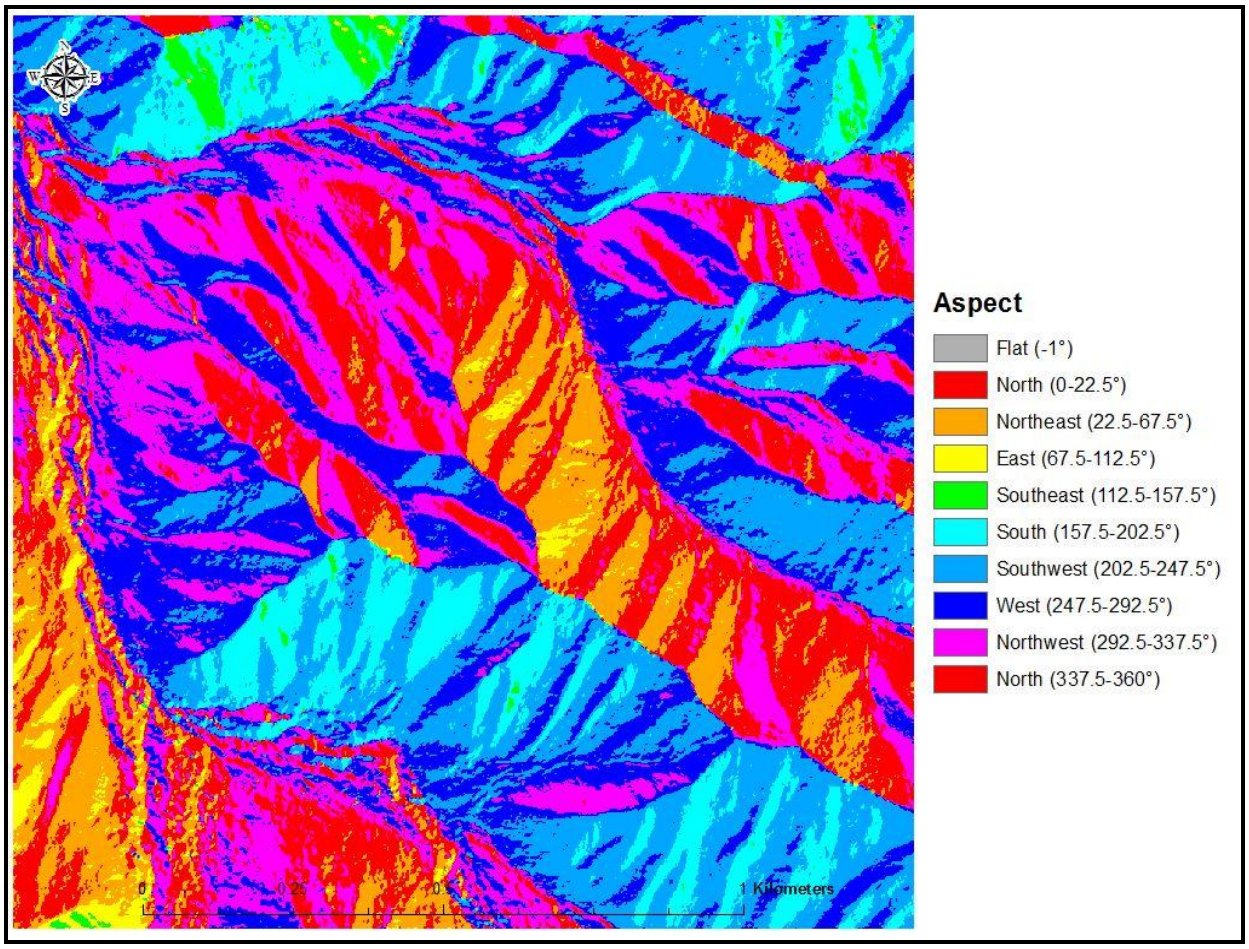

Figure 3.8 - Aspect raster of False Gap

Table 3.4 – Stata output of frequency statistics for soils in False Gap

Soil_cat	Freq.	Percent	Cum.
Ditney	11,344	54.78	54.78
Soco	6,536	31.56	86.34
Spivey	2,828	13.66	100.00
Total	20,708	100.00	

Table 3.5 - Stata output of frequency statistics for aspect in False Gap

Asp_name	Freq.	Percent	Cum.
E	137	0.66	0.66
N	2,638	12.74	13.40
NE	1,148	5.54	18.94
NW	2,883	13.92	32.87
S	2,947	14.23	47.10
SE	498	2.40	49.50
SW	6,359	30.71	80.21
W	4,098	19.79	100.00
Total	20,708	100.00	

Table 3.6 - Stata output of frequency statistics for overstory vegetation in False Gap

Name	Freq.	Percent	Cum.
Low Elevation Mixed Pine- Xeric Oak	941	4.54	4.54
Montane Alluvial Forest	218	1.05	5.60
SA Cove Hardwoods	5,809	28.05	33.65
SA Early Successional Hardwoods	2,768	13.37	47.02
SA Mixed Hardwoods without Oaks	2,294	11.08	58.09
SA Mixed Hardwoods, Acidic	2	0.01	58.10
SA Northern Hardwoods	615	2.97	61.07
Submesic to Mesic Oak/Hardwoods	8,061	38.93	100.00
Total	20,708	100.00	

Table 3.7 - Stata output of descriptive statistics for all continuous variables in False Gap

Variable	Obs	Mean	Std. Dev.	Min	Max
treeheight	20708	19.67359	8.134268	3.552734	47.67651
slope	20708	28.52869	9.350015	.237586	69.1052
elevation	20708	907.5433	109.3009	662.903	1153.98
near_dist	20708	163.3713	101.2992	0	437.5657

Results and Discussion

Examination of normality and variables

Before running the initial model using an OLS estimator, the continuous variables were visually inspected for normal distribution (Figures 3.9-3.12).

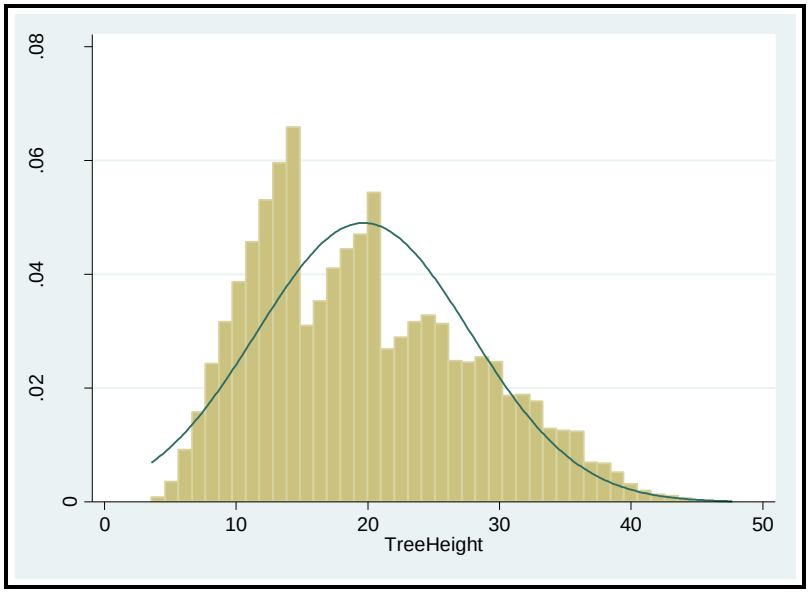

Figure 3.9 - Data distribution versus a normal curve for treeheight variable

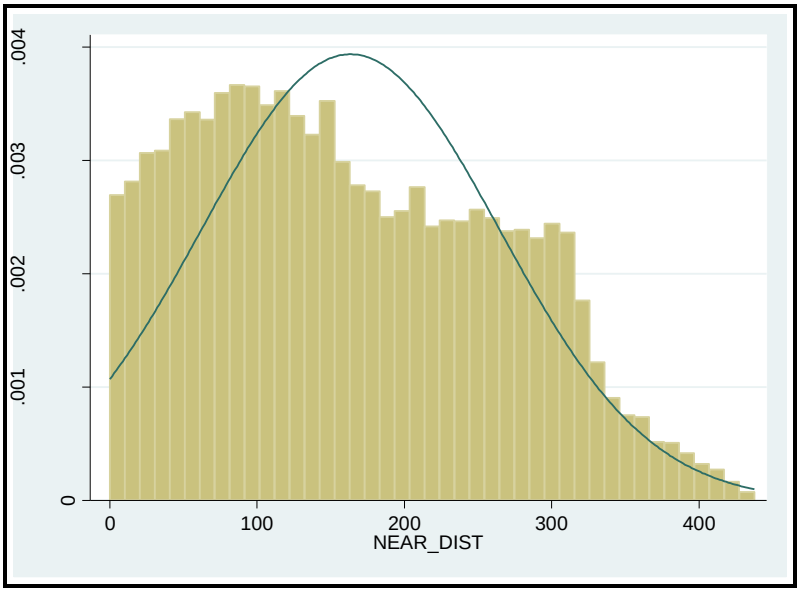

Figure 3.10 - Data distribution versus a normal curve for near_dist variable

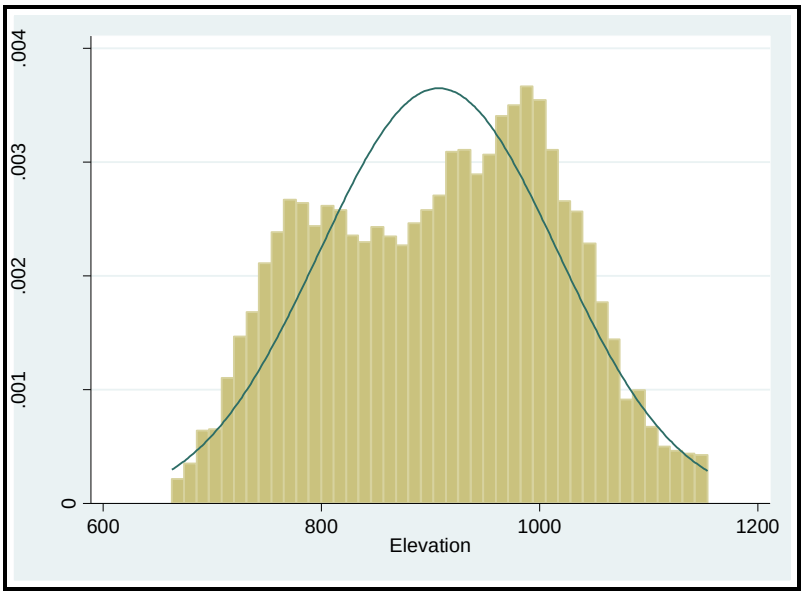

Figure 3.11 - Data distribution versus a normal curve for elevation variable

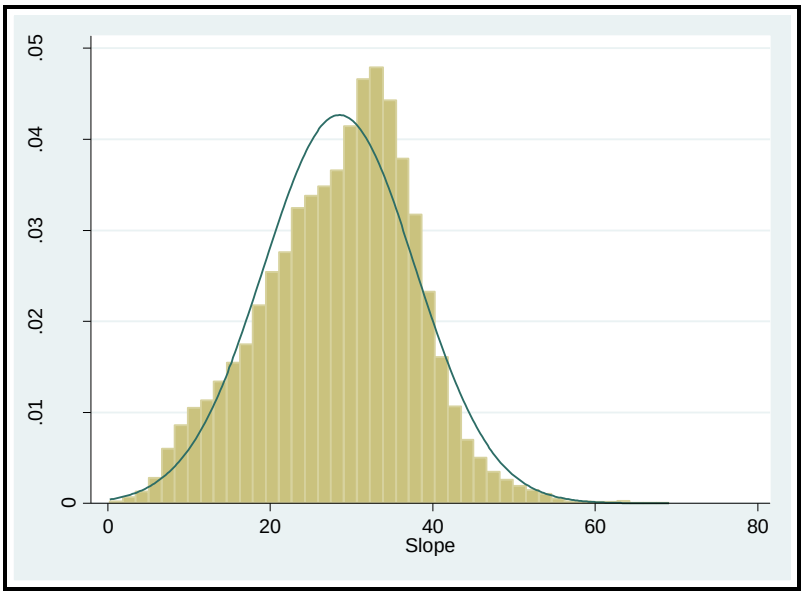

Figure 3.12 – Data distribution verses a normal curve for slope variable

The histograms showing the data distribution against a normal distribution curve show that there is non-normality in the elevation, distance, and tree height data. The slope variable is closest to a normal distribution. The same trends can be seen in the boxplots of the same variables (Figures

3.13-3.16). Although normality is not apparent in all of the variable data, this is not necessarily crucial to the model's success because the distribution of the variables' residuals display a normal distribution.

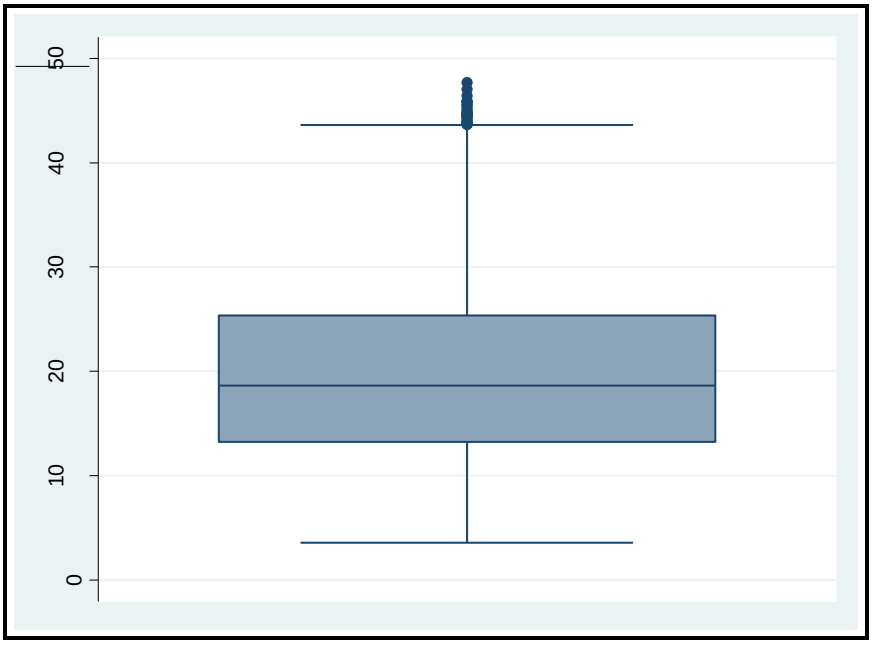

Figure 3.13 – Boxplot of treeheight variable data

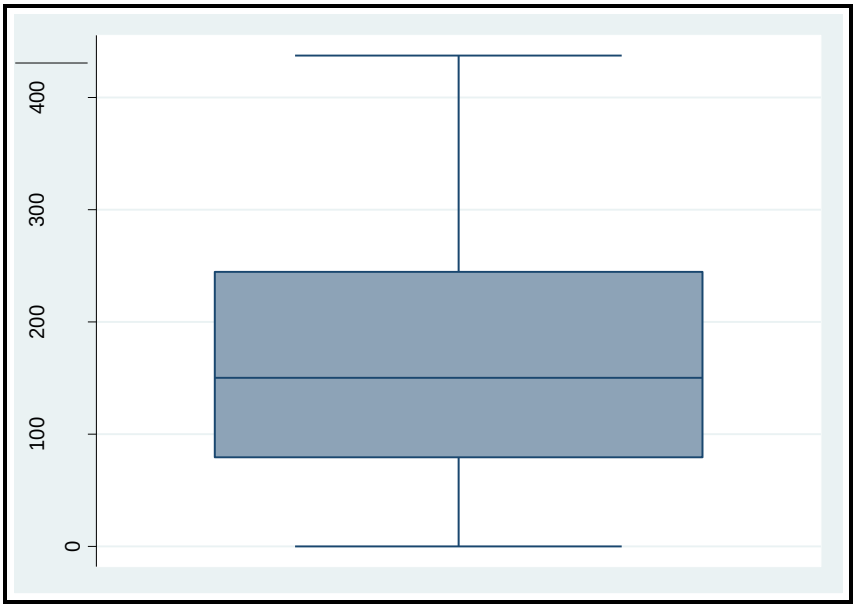

Figure 3.14 – Boxplot of near_dist variable data

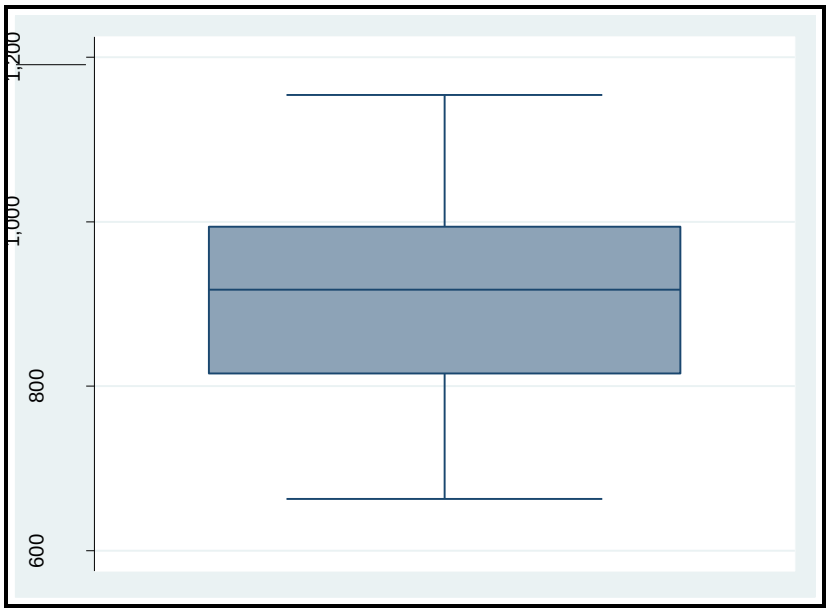

Figure 3.15 – Boxplot of elevation variable data

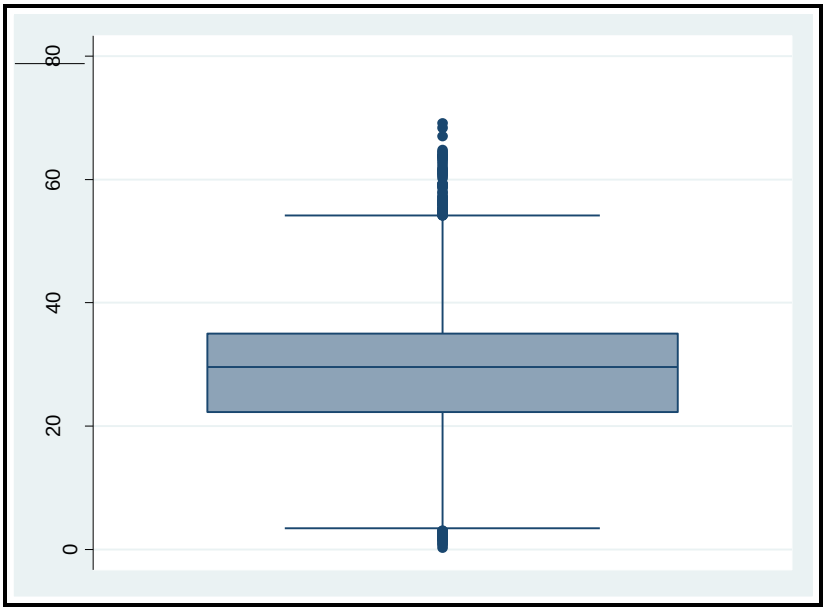

Figure 3.16 – Boxplot of slope variable data

Scatterplots were created for the continuous variables before executing the model for further visual analysis of the relationships between the variables and tree height (Figures 3.17-3.19).

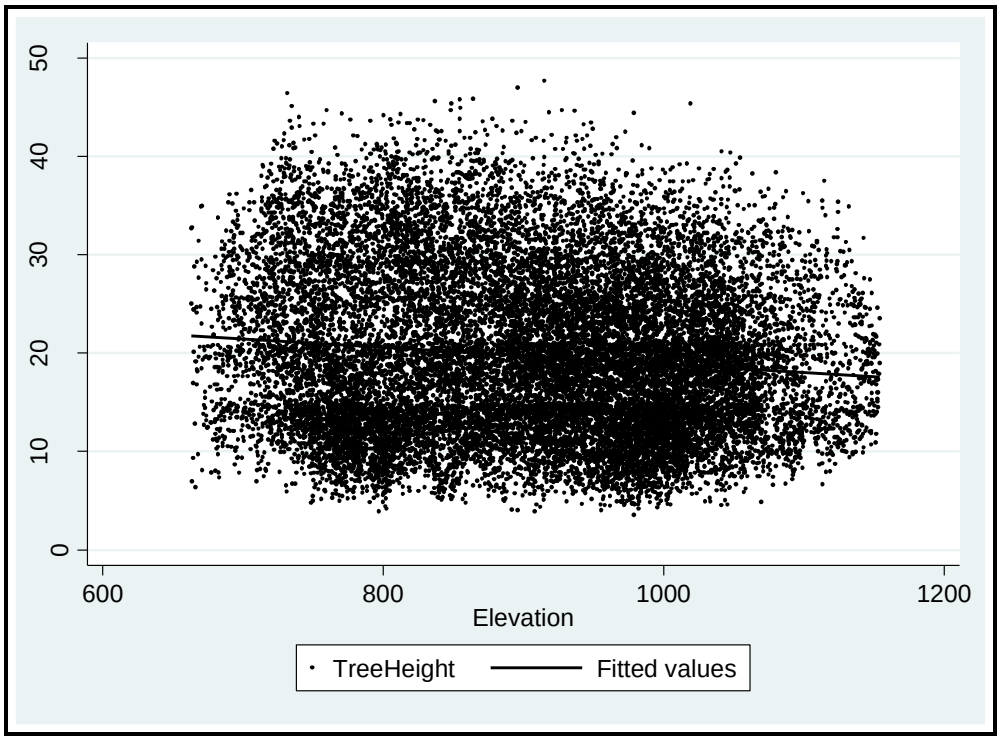

Figure 3.17 - Scatterplot of elevation (m) and treeheight (m)

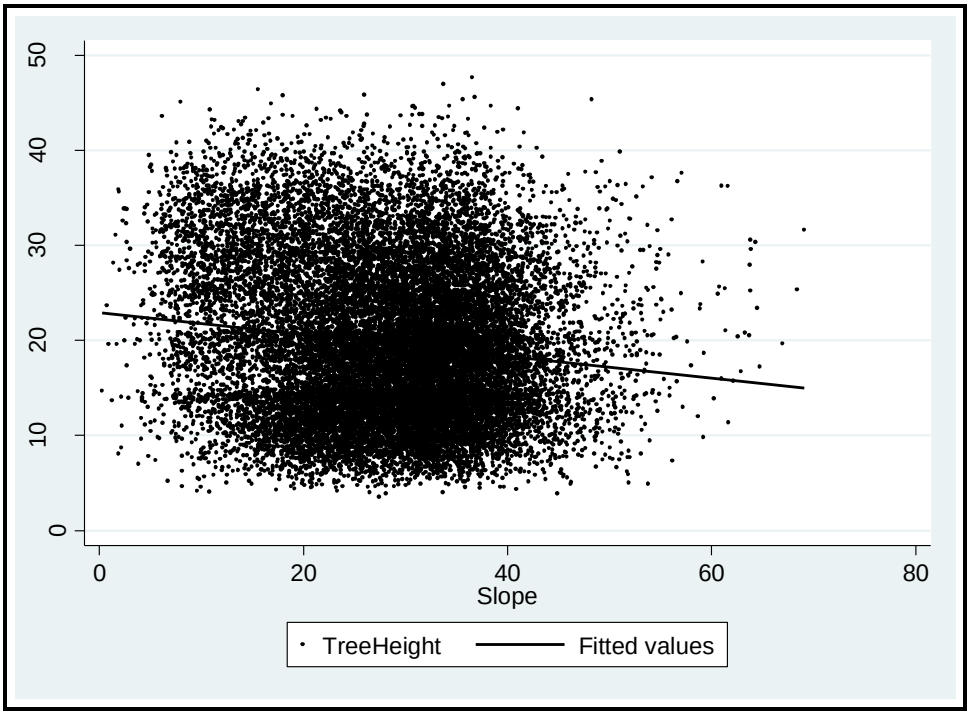

Figure 3.18 - Scatterplot of slope (%) and treeheight (m)

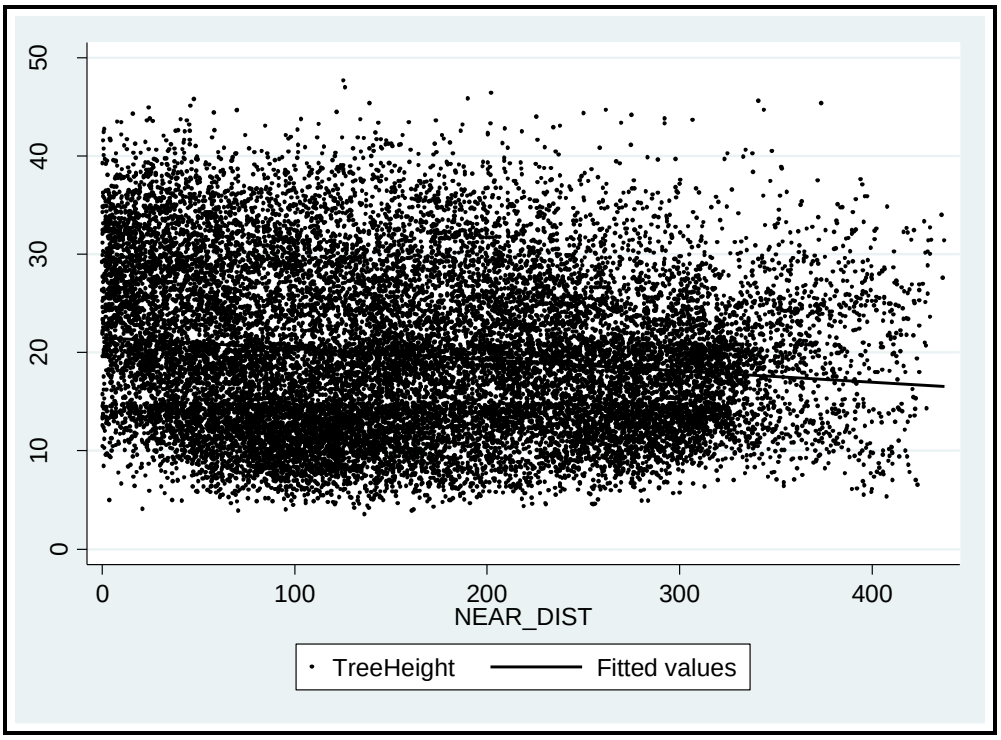

Figure 3.19 - Scatterplot of near_dist (m) and treeheight (m)

General comments can be made regarding these relationships. There is a visible decrease in tree heights with increasing elevation, increasing slope, as well as increased distance from water based on the fitted lines in each of these scatterplots. It also appears that increasing slope may limit the number of trees in areas with high slope percentages. These visual interpretations are crucial for basic understanding of the relationships between the continuous variables. A more rigorous statistical understanding can be obtained by examining the metrics of the model regression (Table 3.8).

In addition to visual interpretation of the continuous variables in the model, some general comments can be made regarding the categorical variables used in the regression. The bar graphs (Figures 3.20– 3.22) show some general trends regarding the mean height of trees for each category. For soils, it can be seen that the mean height of trees in areas with Spivey soils is

higher relative to the other two soil types. Also, the assumptions based on empirical evidence with respect to aspect seem to hold true as well with mean tree height greater in generally north facing slopes and lower in south facing slopes. An interpretation of the overstory vegetation graph shows less variability and the mean height in SA Mixed Hardwoods, Acidic category is somewhat misleading given that there are only two observations of the > 20000 for this community. By examining the scatterplot graphs of each categorical variable, we can see how the mean values of tree height change between variables. For the three soil variables, visual inspection implies that the mean value of tree height will increase in the presence of Spivey soils and decrease with Soco and Ditney soil complexes (Figures 3.23-3.25). Figures 3.26-3.33 show the effect of the aspect categories on the mean of tree height. As predicted by empirical evidence, the mean height of trees in the study area generally increases in the north and decreases in the south.

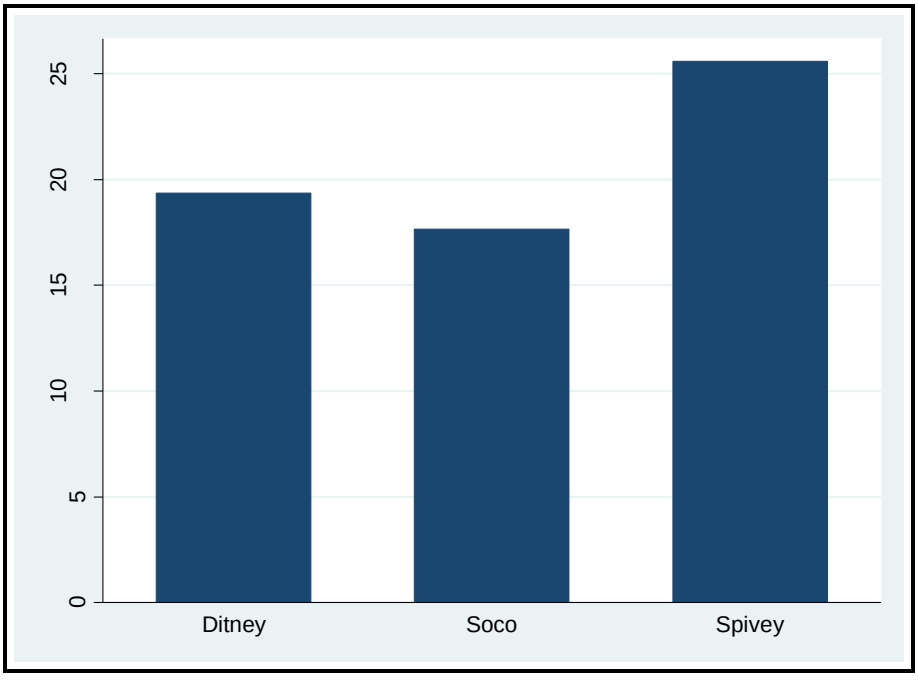

Figure 3.20 - Graph of soil types and mean tree height (m)

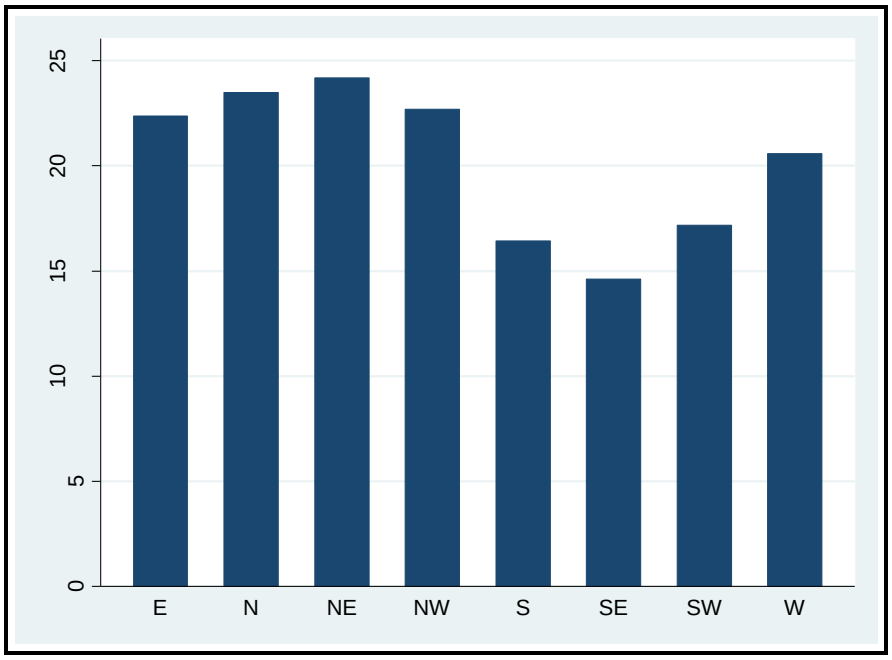

Figure 3.21 - Graph of aspect and mean tree height (m)

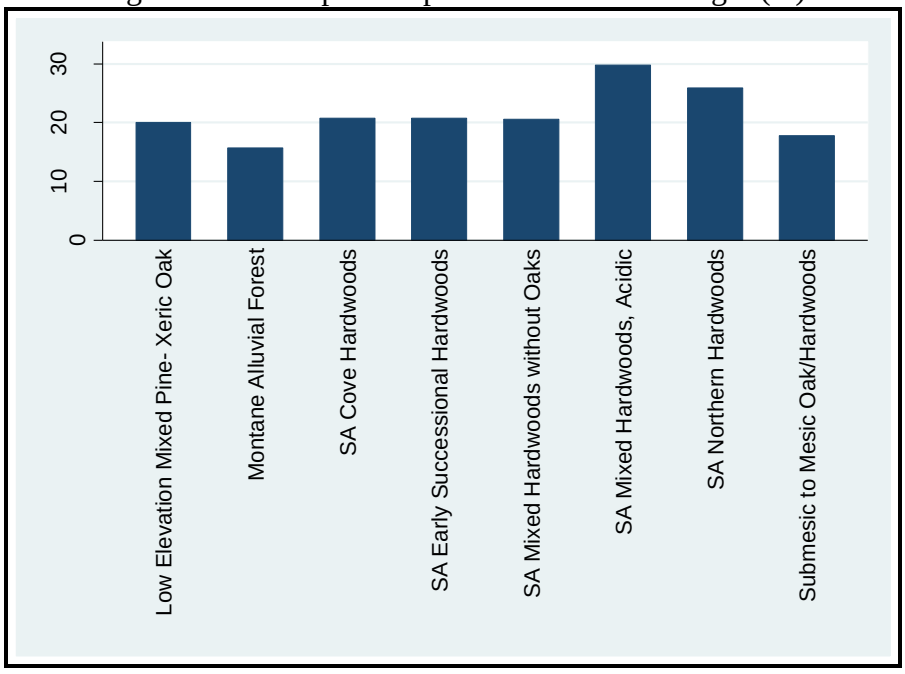

Figure 3.22 – Graph of tree category and mean tree height (m)

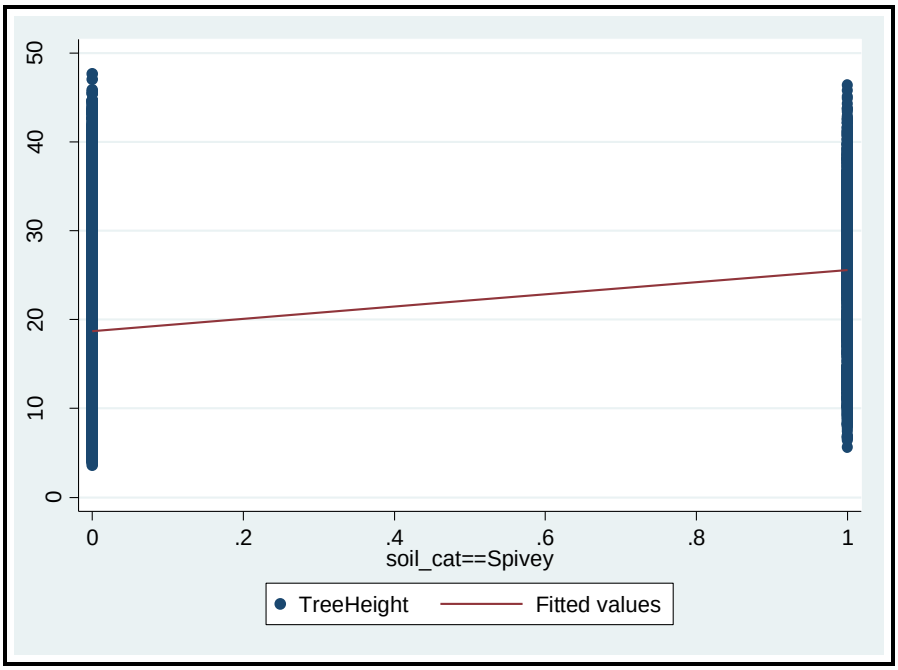

Figure 3.23 – Scatterplot of categorical variable s3 (Spivey soil) and tree height (m)

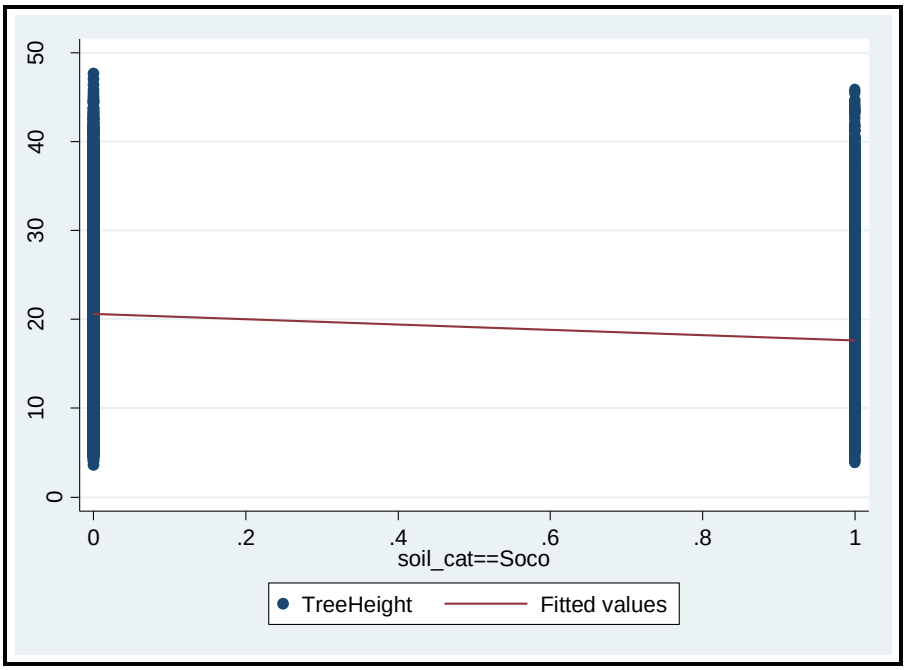

Figure 3.24 – Scatterplot of categorical variable s2 (Soco soil) and tree height (m)

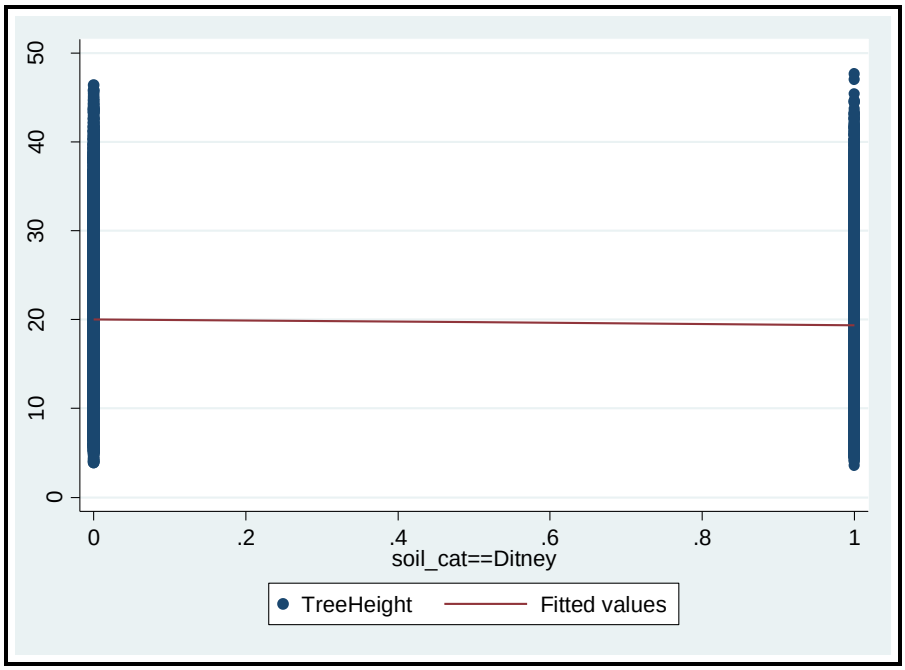

Figure 3.25 - Scatterplot of categorical variable s1 (Ditney soil) and tree height (m)

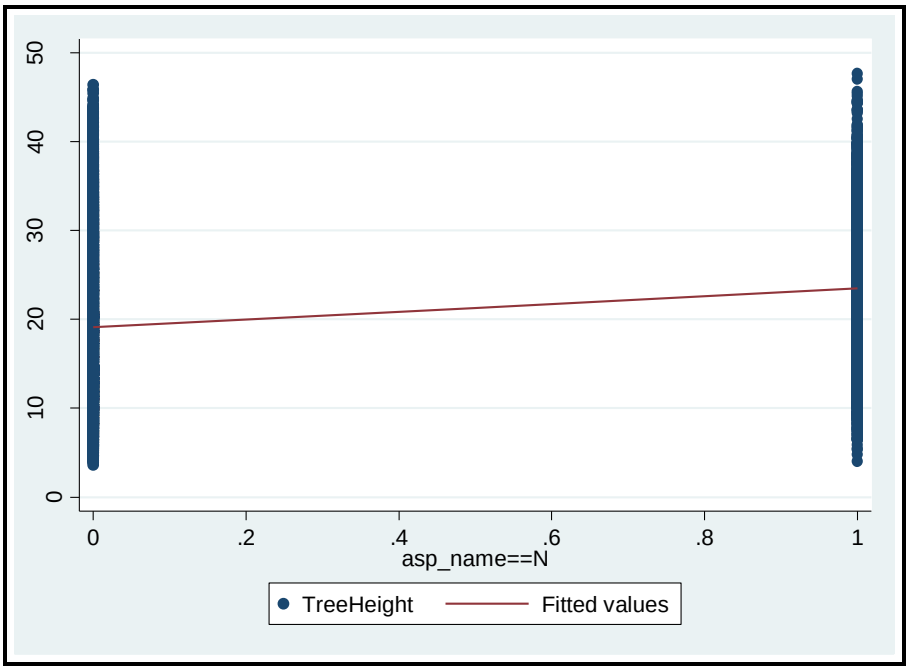

Figure 3.26 - Scatterplot of categorical variable a1 (North aspect) and tree height (m)

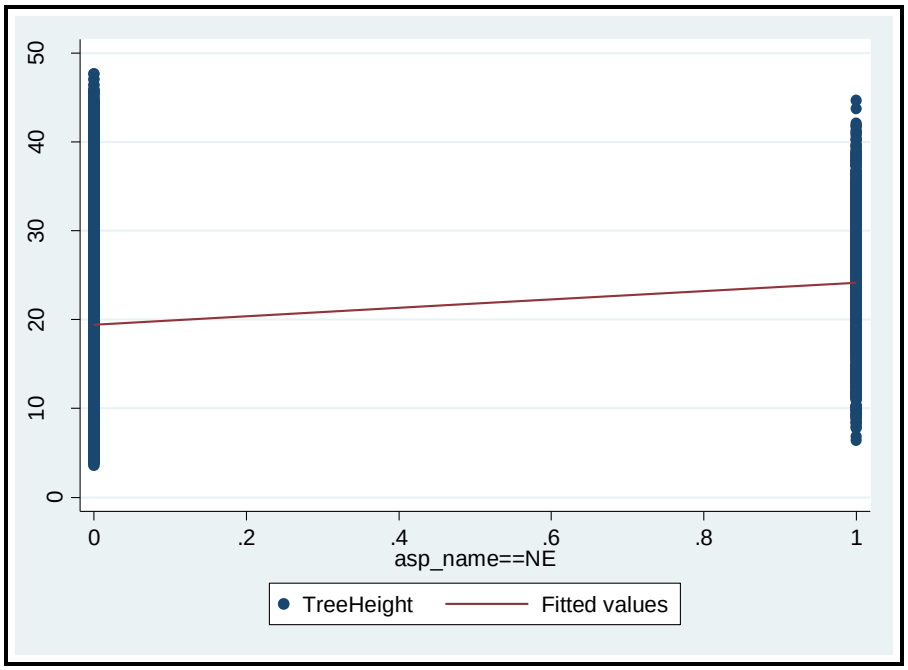

Figure 3.27 - Scatterplot of categorical variable a2 (Northeast aspect) and tree height (m)

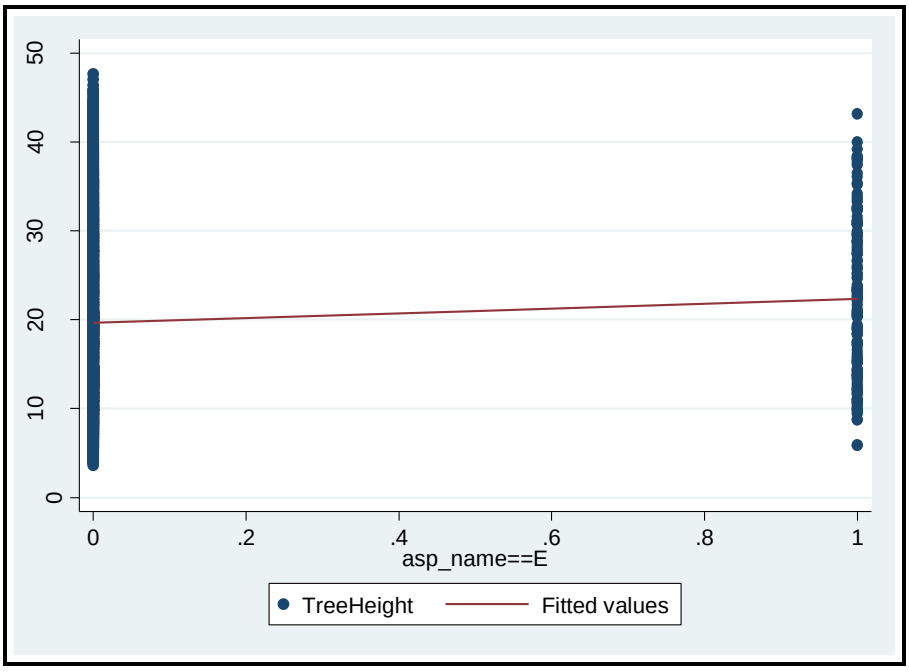

Figure 3.28 - Scatterplot of categorical variable a3 (East aspect) and tree height (m)

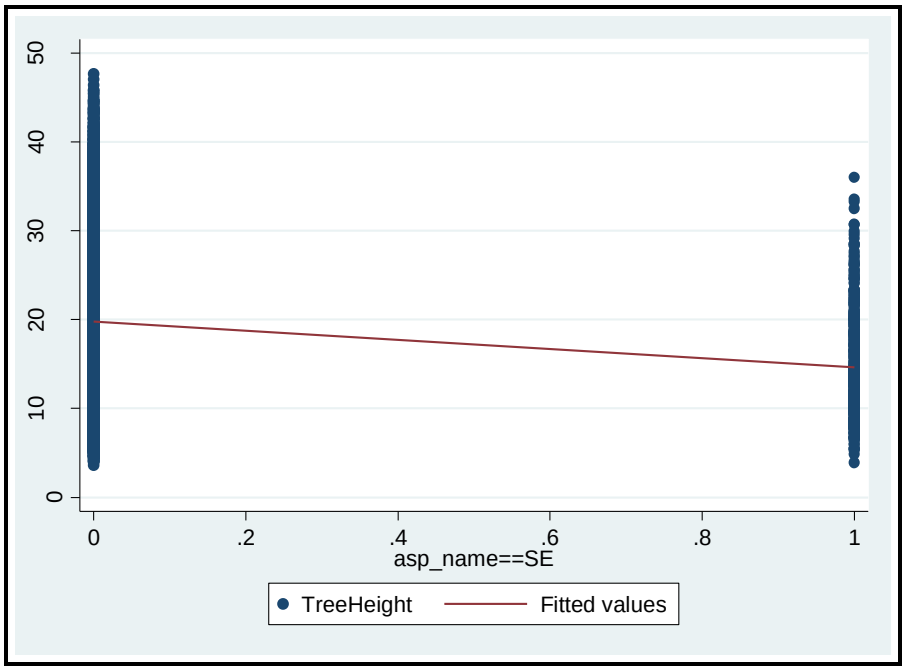

Figure 3.29 - Scatterplot of categorical variable a4 (Southeast aspect) and tree height (m)

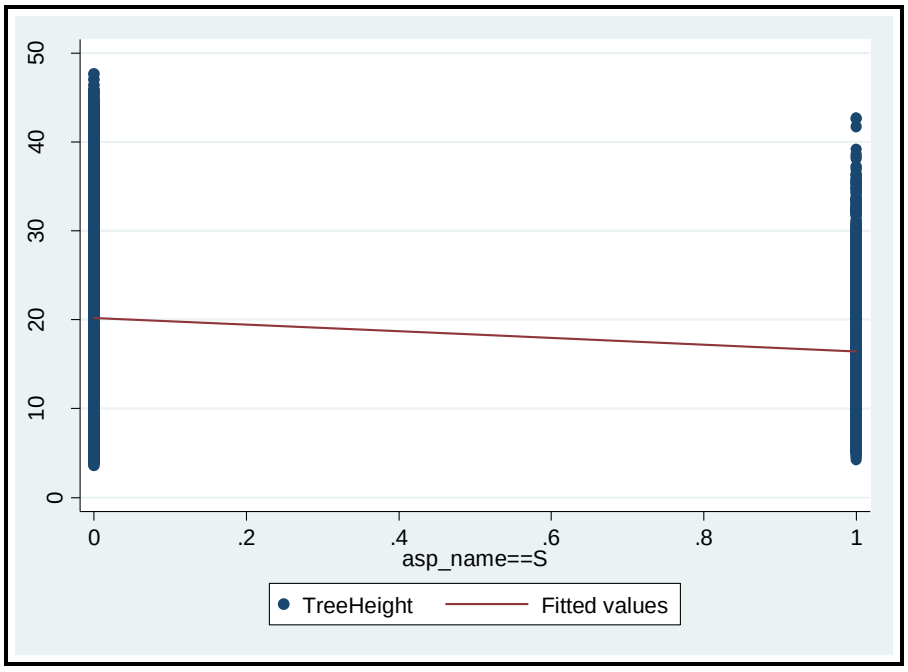

Figure 3.30 - Scatterplot of categorical variable a5 (South aspect) and tree height (m)

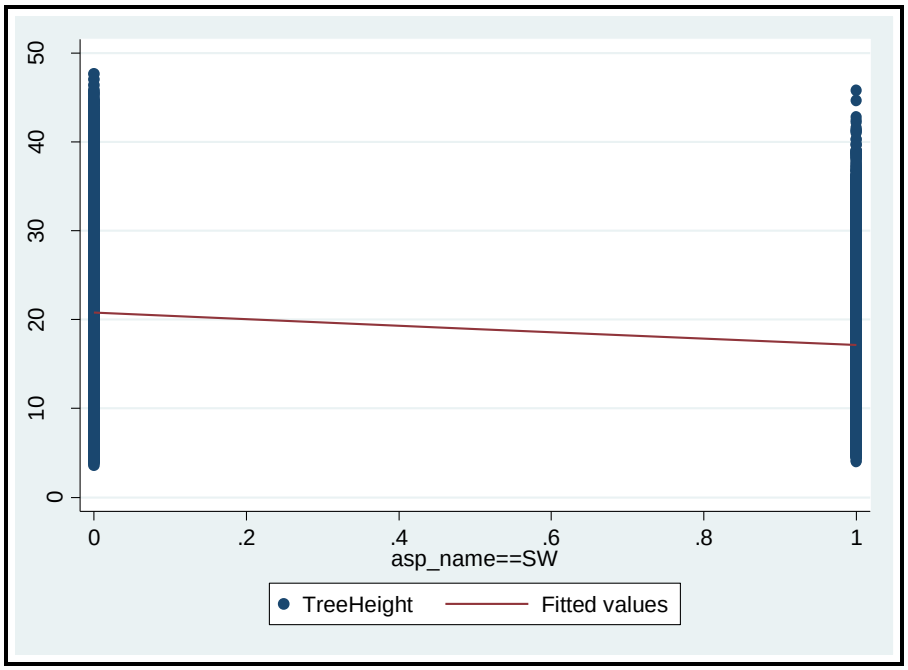

Figure 3.31 - Scatterplot of categorical variable a6 (Southwest aspect) and tree height (m)

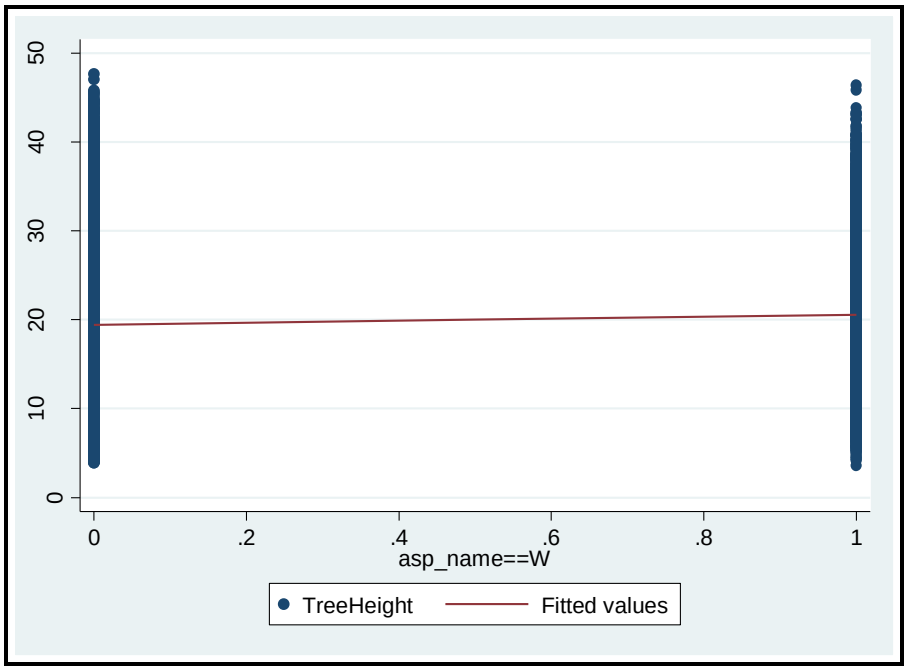

Figure 3.32 - Scatterplot of categorical variable a7 (West aspect) and tree height (m)

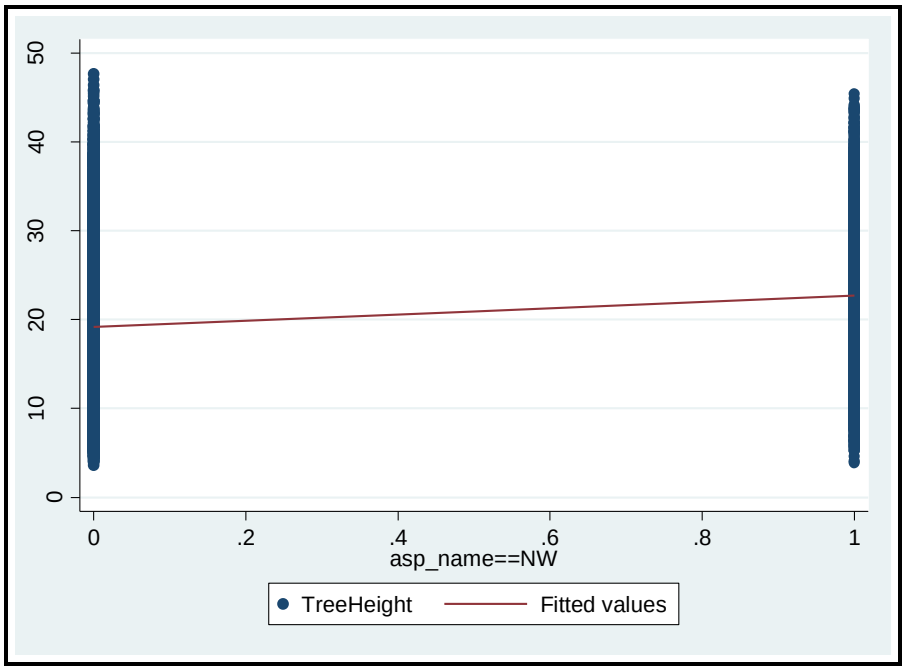

Figure 3.33 - Scatterplot of categorical variable a8 (Northwest aspect) and tree height (m)

Model regression and diagnostics

The initial model (Model 1) was executed using an OLS estimator because of its simplicity as a linear unbiased estimator with the formula defined as:

$$treeheight = b_0 + b_1*elevation + b_2*near\_dist + b_3*slope + b_4*s1 + b_5*s2 + b_6*a1 + b_7*a2 + b_8*a4 + b_9*a5 + b_{10}*a6 + b_{11}*a7 + b_{12}*a8 + b_{13}*n1 + b_{14}*n2 + b_{15}*n3 + b_{16}*n4 + b_{17}*n5 + b_{18}*n7 + b_{19}*n8$$

Table 3.8 shows the output results of the initial model regression. Review of the output shows that the *F-test* is statistically significant for the overall model (Prob > F = 0.0000). The R^2 value of 0.2390 means that approximately 24 percent of the variance of tree height in the study area is accounted for by Model 1. Because of the large number of observations (20708) relative to the

number of predictors (19), the *adjusted-R^2* (0.2383) is not significantly different from the R^2. The *RMSE* = ± 7.0991 m. Based on a *p-value* of 0.0000, Model 1 shows that the collection of independent variables reliably predict the dependent variable of tree height. Closer examination of the Model 1 results shows even more useful information. By examining the effect of each variable on the efficacy of the model, it can be seen that based on the *t-test*, none of the categorical variables concerning overstory vegetation are statistically significant at the *alpha = 0.05* level. This is interesting because, intuitively, one would assume that the type of tree community to which an individual tree belongs, with its individual genetic characteristics, would have a profound effect on the height of the organism. This does not appear to be the case in this particular model for this study area. Also surprising is the insignificance of slope in Model 1. This contradicted the assumptions made at the onset of the study that increasing slope would be influential in determining tree heights.

The remainders of the variables tell their own stories. By examining the raw coefficients, it can be seen that an increase in elevation has a negative effect on the predicted tree heights, but only slightly. This reinforces the pattern seen in the graphic representation of the data from earlier. For each meter increase in elevation, tree heights predicted in this model would decrease by 8 mm net of the effects of all of the other variables. The same negative effect on tree heights is provided by the distance to water. For each meter from a stream feature, tree heights decrease by 15 mm. The same relative relationship is apparent in the *beta* coefficients as well, with a 1σ increase in elevation creating a -0.1139112σ change in tree heights and a 1σ increase in the distance from water creating a -0.1938349σ change in height.

At least for the continuous variables in this model, the distance to water seems to be the predictor that contributes the most to the outcome. All of the remaining dummy variables except

a8 (NW aspect) are significant according to their *t-tests*, but the raw coefficients are of most importance here. The presence in this model of soil types Ditney and Soco shows a negative effect on the mean of tree heights compared to trees in Spivey soils. This was also represented in the graphical interpretation from earlier. We also can see that generally north facing slopes (a2, a3, a4) create a positive effect on the mean of tree heights with their individual presence in comparison to eastern slopes. The sign on the raw coefficients for the south facing slopes (a5, a6, a7) show a negative influence on the overall predicted mean tree height in this model. The b_0 value indicates that in Spivey soils, with eastern aspect in a SA Mixed Hardwood, Acidic community with low slope, low elevation, and near a stream we could expect to see tree heights around 40 m.

Diagnostics for Model 1 showed that there was reason to be concerned about the validity of the model for predicting tree heights. The *variance inflation factor (VIF*) analysis showed problems with the overstory vegetation variables as well as many of the aspect variables which suggest multicollinearity in the data (Table 3.9). This makes sense given that most of the communities of trees are hardwoods of one species or another and that slope aspect is not naturally categorized, but in reality more continuous. There also seemed to be correlation between the variables elevation and s1 (0.82) in Model 1 as shown in Table 3.10. The initial concerns regarding the normality of the data were eased by the analysis of the distribution of the residuals for the model which showed a nearly normal distribution (Figure 3.42). The scatterplot of fitted values vs. the residuals provided evidence that the model may contain heteroscedasticity and some influential observations (Figure 3.43).

Table 3.8 - Stata output of regression metrics for Model 1

Source	SS	df	MS		
Model	327499.746	19	17236.8288	Number of obs =	20708
Residual	1042606.06	20688	50.3966582	F(19, 20688) =	342.02
				Prob > F =	0.0000
Total	1370105.81	20707	66.1663114	R-squared =	0.2390
				Adj R-squared =	0.2383
				Root MSE =	7.0991

treeheight	Coef.	Std. Err.	t	P>|t|	Beta
slope	-.00099	.006185	-0.16	0.873	-.001138
elevation	-.0084774	.0008598	-9.86	0.000	-.1139112
near_dist	-.0155648	.0006473	-24.04	0.000	-.1938349
s1	-1.957992	.2471761	-7.92	0.000	-.119806
s2	-4.642676	.1929783	-24.06	0.000	-.2652735
a2	2.95577	.6263579	4.72	0.000	.1211549
a3	2.377997	.6427148	3.70	0.000	.0668991
a4	1.904571	.6253784	3.05	0.002	.0810565
a5	-3.447952	.6256608	-5.51	0.000	-.1480945
a6	-5.001658	.6928934	-7.22	0.000	-.0942032
a7	-2.860934	.6182169	-4.63	0.000	-.1622433
a8	.0606113	.6209349	0.10	0.922	.0029688
n1	-6.409651	5.028473	-1.27	0.202	-.1641169
n3	-4.583735	5.023344	-0.91	0.362	-.2531644
n2	-9.623848	5.045111	-1.91	0.056	-.1207541
n4	-6.478364	5.024034	-1.29	0.197	-.2710276
n5	-5.4207	5.024137	-1.08	0.281	-.2091607
n7	-4.416366	5.02934	-0.88	0.380	-.0921677
n8	-8.38914	5.022821	-1.67	0.095	-.5028744
_cons	39.72185	5.11124	7.77	0.000	.

Table 3.9 - Variance Inflation Factor analysis results for Model 1

Variable	VIF	1/VIF
n8	2464.52	0.000406
n3	2092.69	0.000478
n4	1201.03	0.000833
n5	1021.70	0.000979
n1	450.67	0.002219
n7	299.50	0.003339
n2	108.94	0.009179
a7	33.42	0.029926
a8	25.15	0.039765
a5	19.63	0.050935
a4	19.26	0.051925
a2	17.92	0.055804
a3	8.89	0.112510
s1	6.22	0.160805
a6	4.63	0.215979
elevation	3.63	0.275589
s2	3.31	0.302538
near_dist	1.77	0.565985
slope	1.37	0.727736
Mean VIF	409.70	

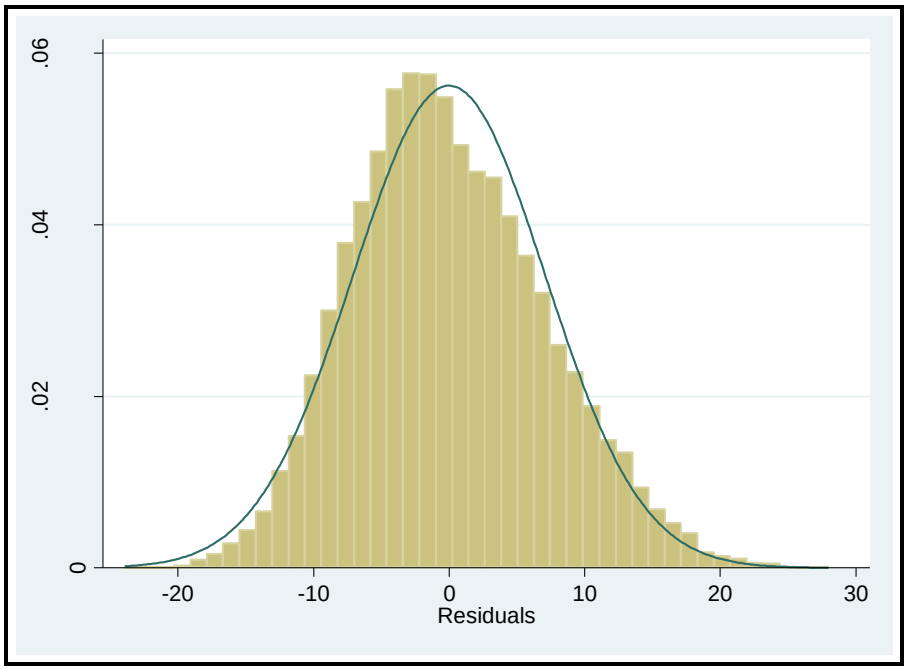

Figure 3.34 - Distribution of residuals for Model 1

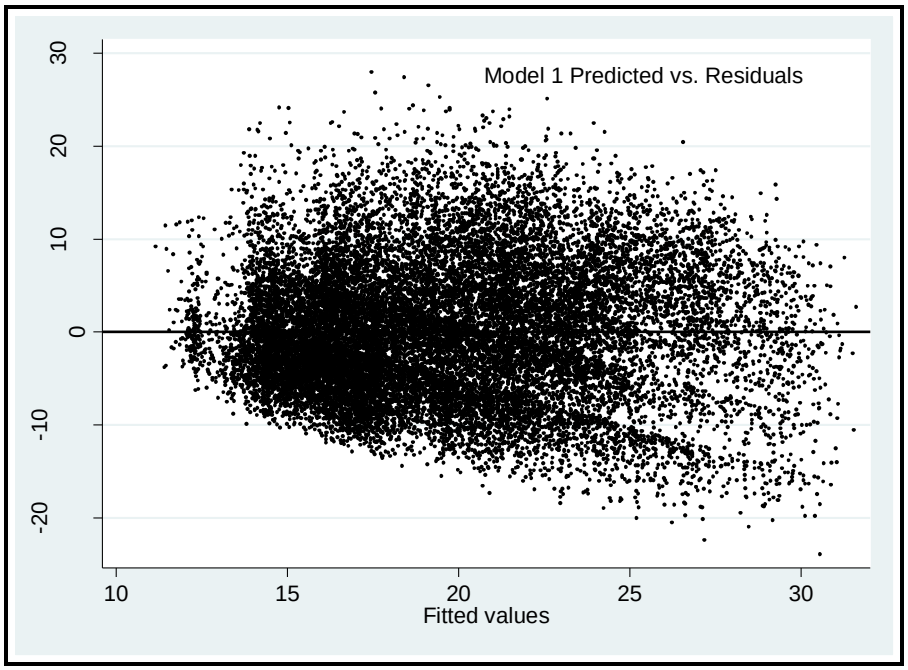

Figure 3.35 - Fitted vs. residual values for Model 1

Table 3.10 – Correlation matrix for continuous and categorical variables with high correlation between s1 and elevation highlighted

	treehe~t	elevat~n	slope	near_d~t	s1	s2	a2	a3	a4	a5	a6	a7	a8	n1	n3	n2	n4	n5	n7	n8
treeheight	1.0000																			
elevation	-0.1144	1.0000																		
slope	-0.1323	0.2377	1.0000																	
near_dist	-0.1416	0.4654	0.2664	1.0000																
s1	-0.0422	0.8205	0.1425	0.4035	1.0000															
s2	-0.1685	-0.5174	0.1994	-0.1360	-0.7475	1.0000														
a2	0.1792	0.0570	0.0434	0.0727	0.0457	-0.0491	1.0000													
a3	0.1345	0.0573	0.0188	-0.0407	0.0730	-0.1232	-0.0926	1.0000												
a4	0.1493	-0.1064	-0.0953	0.0868	-0.0870	-0.0060	-0.1537	-0.0974	1.0000											
a5	-0.1621	-0.0522	0.0183	-0.0187	0.0104	0.0487	-0.1556	-0.0987	-0.1638	1.0000										
a6	-0.0976	-0.1218	-0.0833	-0.0987	-0.0999	0.1165	-0.0600	-0.0380	-0.0631	-0.0639	1.0000									
a7	-0.2053	0.1355	0.0559	-0.0772	0.0844	0.0036	-0.2544	-0.1613	-0.2677	-0.2712	-0.1045	1.0000								
a8	0.0551	-0.0485	-0.0098	0.0376	-0.0733	0.0320	-0.1898	-0.1203	-0.1998	-0.2023	-0.0780	-0.3307	1.0000							
n1	0.0103	-0.1649	-0.0318	-0.0623	-0.1586	0.1227	-0.0229	-0.0225	-0.0228	-0.0099	0.0021	0.0357	0.0179	1.0000						
n3	0.0843	0.1944	0.1129	0.3519	0.1928	-0.1361	0.0084	-0.0122	0.0063	-0.0085	-0.0882	-0.0063	0.0500	-0.1362	1.0000					
n2	-0.0507	0.0982	-0.0332	0.1088	0.0918	-0.0700	0.0415	-0.0229	0.0364	-0.0393	-0.0162	0.0042	-0.0156	-0.0225	-0.0644	1.0000				
n4	0.0542	-0.0943	-0.0320	0.0518	-0.1196	0.0792	0.0955	-0.0127	0.1027	-0.1011	-0.0607	-0.1058	0.0742	-0.0857	-0.2453	-0.0405	1.0000			
n5	0.0399	0.2518	0.0938	0.1514	0.2313	-0.1715	0.0308	0.1848	-0.0144	-0.0394	-0.0343	-0.0475	-0.0347	-0.0770	-0.2204	-0.0364	-0.1386	1.0000		
n7	0.1341	-0.0065	-0.0978	-0.1269	0.0692	-0.1188	0.1098	0.0073	0.0800	-0.0696	-0.0219	-0.0443	-0.0448	-0.0382	-0.1092	-0.0180	-0.0687	-0.0618	1.0000	
n8	-0.1819	-0.2235	-0.0875	-0.4098	-0.2189	0.1843	-0.1316	-0.0869	-0.0942	0.1406	0.1559	0.1096	-0.0642	-0.1742	-0.4985	-0.0823	-0.3136	-0.2818	-0.1397	1.0000

Given the issues with insignificant variables, suspected heteroscedasticity, and suspected influential observations, a decision was made to create a second model – Model 2, to try to improve the performance of the predictor for tree heights in the False Gap study area. All of the overstory type variables were removed as well as slope and the variable for northwest aspect. The regression was run with the new parameters and the output can be seen in Table 3.11.

Table 3.11 - Stata output of regression metrics for Model 2

```
. regress treeheight elevation near_dist s1 s2 a2 a3 a4 a5 a6 a7, beta
```

Source	SS	df	MS		
Model	281846.684	10	28184.6684	Number of obs	= 20708
Residual	1088259.13	20697	52.580525	F(10, 20697)	= 536.03
				Prob > F	= 0.0000
Total	1370105.81	20707	66.1663114	R-squared	= 0.2057
				Adj R-squared	= 0.2053
				Root MSE	= 7.2512

treeheight	Coef.	Std. Err.	t	P>|t|	Beta
elevation	-.0085856	.0008671	-9.90	0.000	-.1153659
near_dist	-.0098157	.0005973	-16.43	0.000	-.1222386
s1	-1.900884	.2446983	-7.77	0.000	-.1163116
s2	-5.131421	.1782767	-28.78	0.000	-.2931994
a2	3.043134	.1803883	16.87	0.000	.1247359
a3	2.734529	.2438076	11.22	0.000	.0769293
a4	1.838883	.1762269	10.43	0.000	.0782609
a5	-4.000873	.17624	-22.70	0.000	-.1718432
a6	-6.147989	.3481603	-17.66	0.000	-.1157937
a7	-3.20735	.1468173	-21.85	0.000	-.1818886
_cons	32.63681	.6857501	47.59	0.000	.

The overall results of this model are similar to Model 1, but with some of the manageable factors corrected. Again the *F-test* is significant and the $p = 0.0000$ means that the overall model is statistically significant. This second model accounts for less variability than Model 1, with an $R^2 = 0.2057$ and *adjusted* $R^2 = 0.2053$ or around 21 percent of the variability compared to 24 percent in Model 1. The *RMSE* = ± *7.2512* m is slightly higher for this model, also.

All of the *t-tests* for variables in Model 2 are significant. As for Model 1, conclusions can be drawn by examining the raw and beta coefficients for the model. The variables elevation and near_dist display much the same relationship with tree heights as before as far as direction go,

but near_dist has less of an impact net the effect of the other variables now that slope, overstory, and northwest aspect have been removed from the regression (-0.1222386 vs. -0.1938349 *beta* coefficient). South facing slope variables still produce conditions that reduce the predicted mean heights of trees and north facing slopes have the opposite effect as hypothesized. The absence of Spivey soils, or conversely, the presence of Ditney or Soco soils in the study area will reduce the mean value of tree heights. The y-intercept in Model 2 is statistically more significant than in Model 1 (*t=47.59* vs. *t=7.77*). Therefore, more confidence can be placed in the predicted value of tree heights in conditions of low elevation, close to water, in Spivey soils and a west-northwest facing slope of 32.63681 m.

Diagnostics for this model revealed some improvement and some lingering problems with the data. Plots of the histogram of the residuals for Model 2 and pnorm and qnorm graphs show a distribution that is nearly normal, with some increased variance at the low end of the residuals (Figures 3.44-3.46).

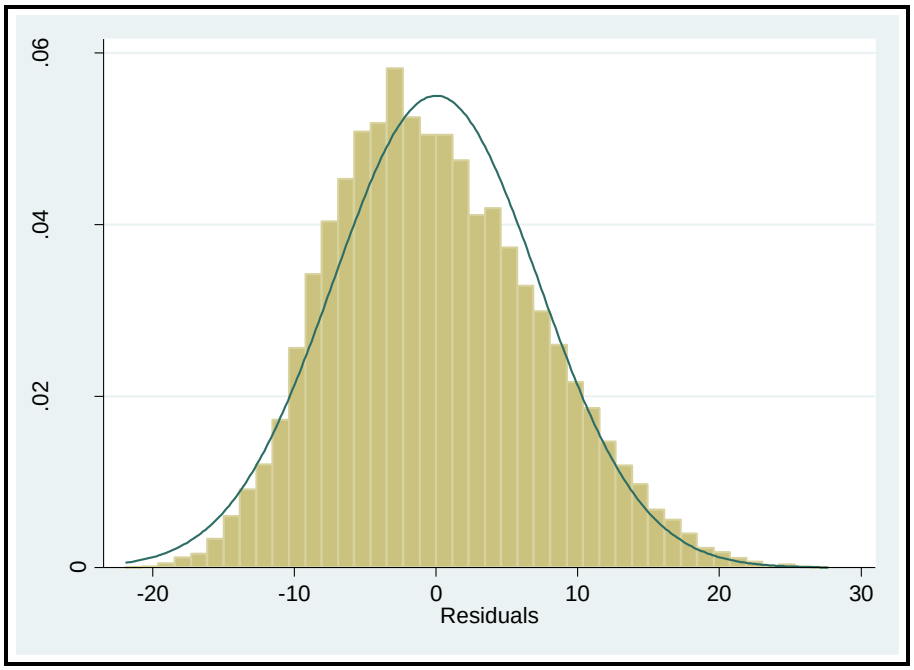

Figure 3.36 - Histogram of residuals of Model 2

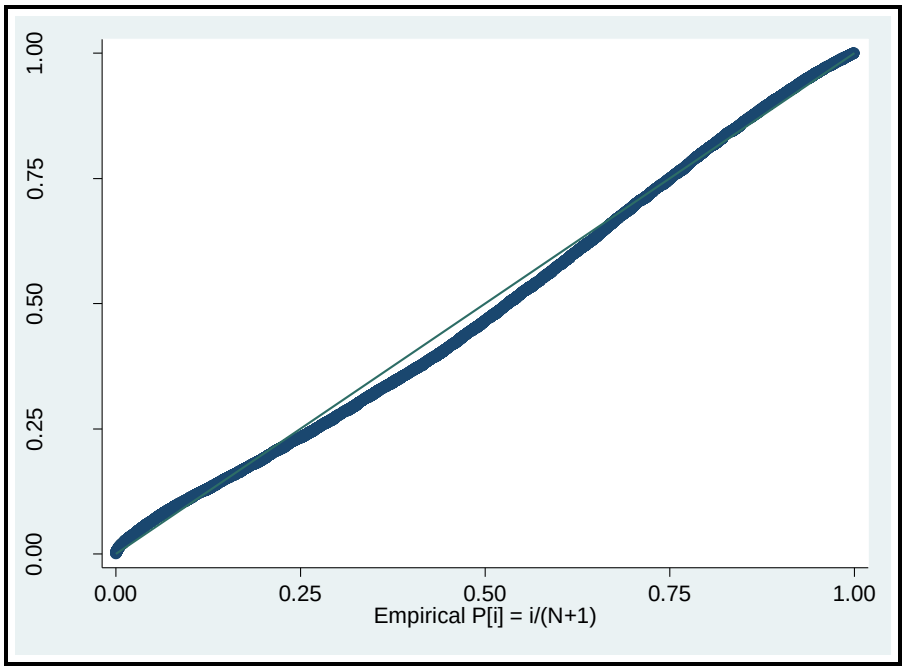

Figure 3.37 - pnorm plot of Model 2

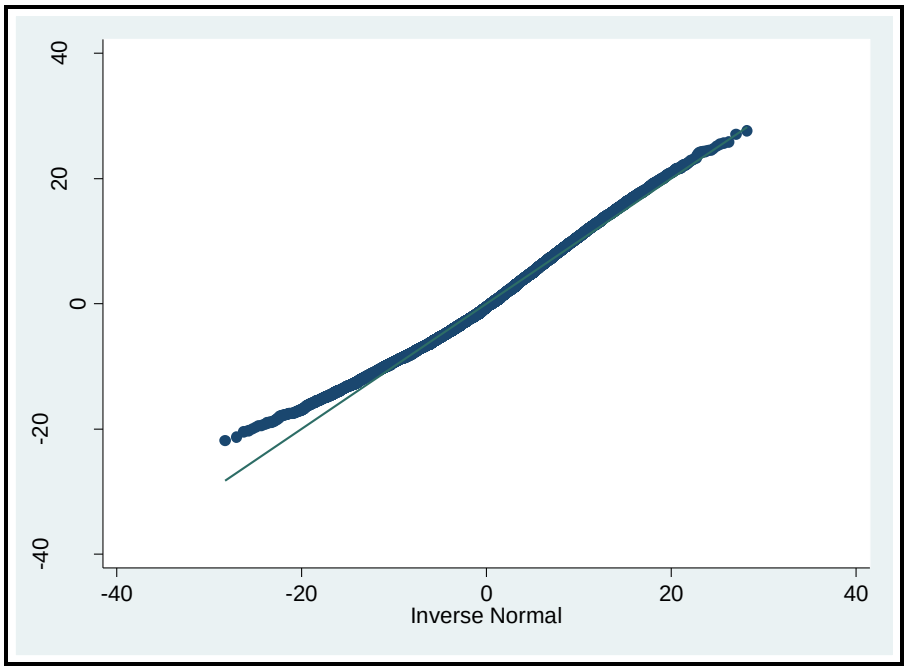

Figure 3.38 - qnorm plot of Model 2

The *VIF* test for multicollinearity of Model 2 was much improved over Model 1 as shown in Table 3.12. Only one value is possibly troublesome and that is for the s1 variable. Some

researchers set the threshold for concern with *VIF* at 5, while others use 10. The value of 5.84 is not well above the lower threshold, so it was retained in the model.

Table 3.12 - VIF values for Model 2

Variable	VIF	1/VIF
s1	5.84	0.171188
elevation	3.54	0.282710
s2	2.70	0.369854
a7	1.81	0.553602
a5	1.49	0.669741
a4	1.47	0.682251
near_dist	1.44	0.693548
a2	1.42	0.701961
a3	1.23	0.815752
a6	1.12	0.892501
Mean VIF	2.21	

Issues concerning correlation between the s1 variable and the elevation variable remain (0.8 value) and this relationship is logical considering the known location of the Ditney soil complex at higher, steeper, rocky elevations.

A rigorous examination of the model was performed to look for influential observations. Scatterplots of elevation and near_dist vs. treeheight were created to identify any possible outliers (Figures 3.47 and 3.48). Three observations were noted with tree heights above 46 m for further investigation. Next, an examination of students' residuals was completed using the residuals of Model 2 and shown in Figure 3.49. Studentized residuals can help to identify outliers. The threshold was defined as any value > 3. A leverage analysis was also completed with the threshold of 0.00106239 ((2k+2)/n, where k = number of predictors, 10) and n = number of observations, 20708). Cook's D was also performed with a threshold set at 0.00019316 (4/n). Cook's D measures overall influence by combining information about residuals and leverage. The final influential observation test was DFITS. This also combines information regarding residuals and leverage and the threshold was set as 0.04395 (2*sqrt(k/n). All of the data were

combined and 230 of the observations met three out of four of the tests. Only one observation of 20708 met all four of the tests. It was removed, but replaced after the model showed no significant improvement in performance. It should be noted that the original three possible outliers identified by the graphic interpretation did not fail any of the tests for influential observations. As a final test for influential observation, the DFBETA was examined with no discernible issues. The decision was made to leave the data set as it was based on the knowledge of the variability of natural phenomenon in general and the ability for some individual organisms to be outstanding without being outside of the realm of possibility.

Finally, the diagnostics for heteroscedasticity indicate that the data are not homoscedastic based on the graphic interpretation of the data as well as the significance calculated in White's Test and the Breusch-Pagan Test (Table 3.13). We must reject the null hypothesis of homoscedasticity in this case.

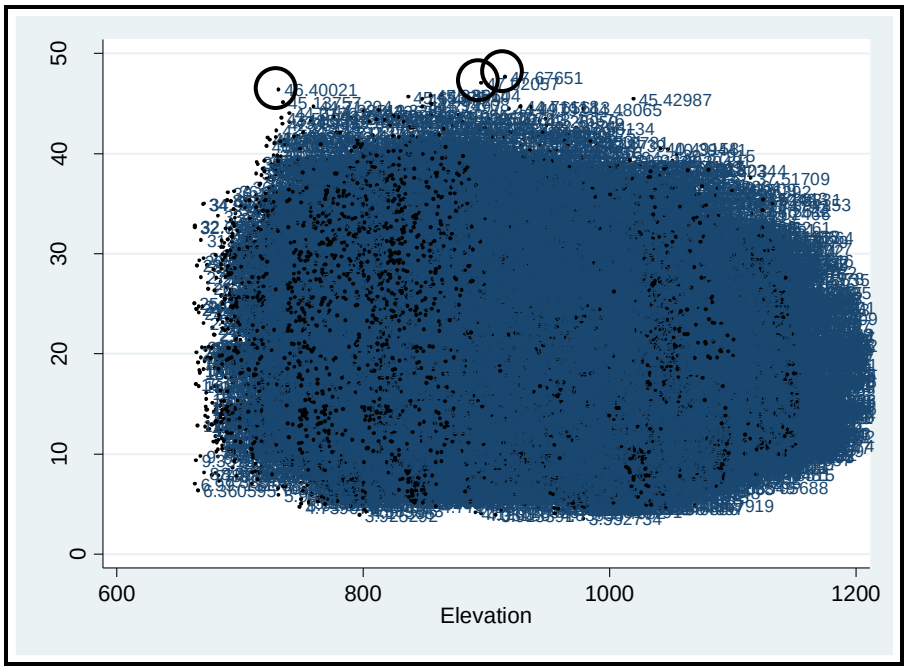

Figure 3.39 – elevation (m) vs. treeheight (m) with potential outliers circled

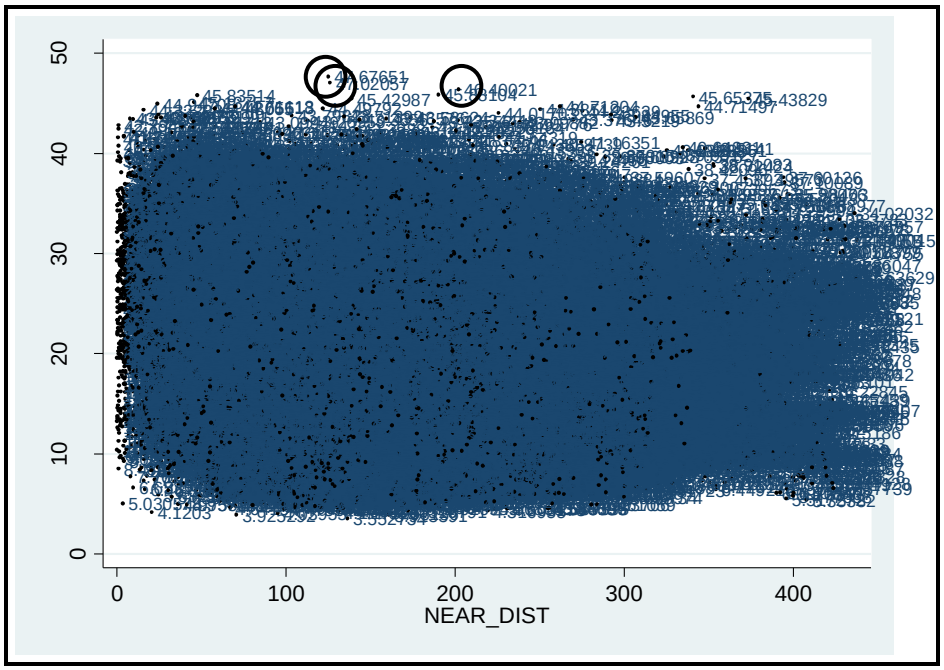

Figure 3.40 - near_dist (m) vs. treeheight (m) with potential outliers circled

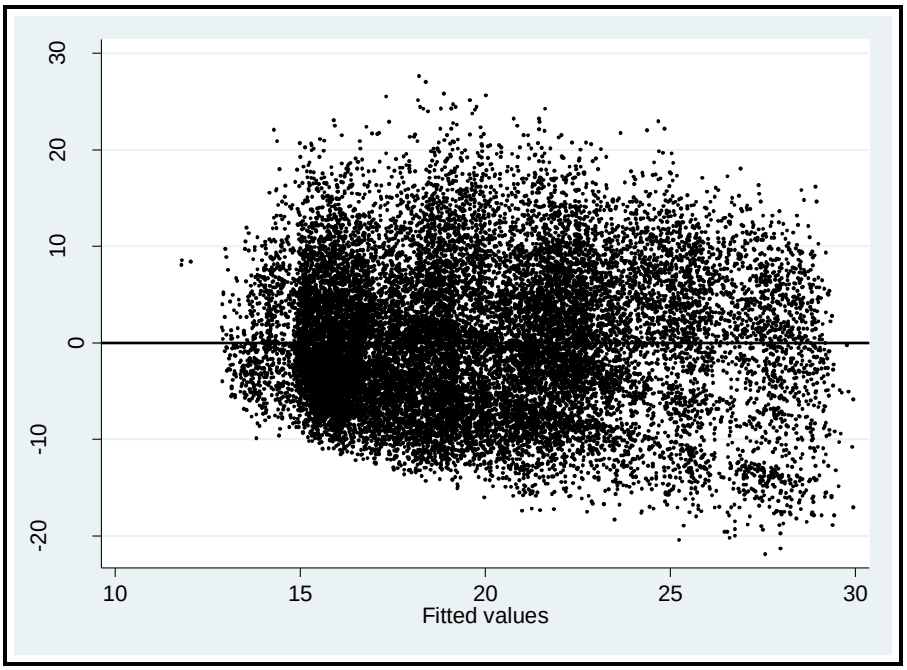

Figure 3.41 - Predicted treeheight values vs. residuals for Model 2

Table 3.13 - Results of White's Test and Breusch-Pagan test for heteroskedasticity

```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
  Heteroskedasticity |    1523.67     41    0.0000
            Skewness |    1017.32     10    0.0000
            Kurtosis |      46.27      1    0.0000
---------------------+-----------------------------
               Total |    2587.27     52    0.0000
---------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of treeheight

         chi2(1)      =   515.78
         Prob > chi2  =   0.0000
```

Conclusions and Recommendations

The use of LiDAR derived data in this study allowed for a rigorous assessment to be made of the environmental variables that contribute to tree heights in the GRSM. The use of feature extraction software provided over twenty thousand observations that were used to create predictive models that would have been costly and time consuming to collect by traditional fieldwork. This study emphasized the problems that are inherent in creating a predictive model using such a large sample set created by the LiDAR data. The predictive Model 2 was able to account for 21 percent of the variability in the heights in the False Gap area with a number of significant variables including elevation, distance to streams, soil types, and aspect as expected. At least for the continuous variables in this model, the distance to water seems to be the predictor that contributes the most to the outcome.

Unexpected was the poor performance of the slope variable in the model, but it is proposed that there be further examination of the effects of interactions of slope with elevation

and soil types, as well as interactions between soil types and elevation based on the evidence of collinearity in Model 2. Future work should include an investigation regarding the historical disturbance in forested areas in the Appalachians as well as climatic effects on tree heights. It is believed that special attention should be given to the Lidar Analyst tool used to create the initial database of tree heights. No substantive work was done in this study to ascertain the accuracy of the automatically derived database.

Also, it should be noted that the study was performed as an evaluation of the methodology of using OLS regression to model one of the most diverse forests in the world in a highly variable environment. This type of straightforward regression model was chosen because of its lack of bias, but it is possible that a more complex model such as a Geographically Weighted Regression model (GWR) or a Generalized Least Squares (GLS) model would perform better at predicting tree heights in this type of complex and varied environment.

Overall the study shows the challenges in modeling natural environments, which by their very nature, contain high levels of complexity and variability. It is surmised that because this area is dominated by hundreds of species of broadleaf deciduous trees, the interspecies variability in attainable tree heights make the task of modeling heights a challenging one. It is suggested that future work select communities/species of trees to model in order to remove variability and potentially improve the predictive model.

CHAPTER 4

CONCLUSIONS

Humans have always been fascinated by the natural world around them. This fascination and curiosity can be seen in those who choose to study forest structures and extraordinary trees in parks and forests around the world. Forestry experts, arborists, plant biologists, bio-geographers, and tree enthusiasts are all interested in searching the forests for exceptionally tall trees because of the information that this research can provide regarding old growth communities, favorable growth conditions, biomass stock, and carbon storage. These giant trees also serve as habitats for numerous other species of plants and animals that rely on the health of the tree for their own survival. In addition, the identification and measurement of these trees has implications in the issue of climate change because of their ability to store carbon and provide valuable oxygen into the Earth's atmosphere.

In Chapter 2, Determining the Location of the Ten Tallest Trees in the Tennessee Portion of the Great Smoky Mountains National Park Using LiDAR Data, ten trees were detected in a complex and rugged forested area of the Great Smoky Mountains National Park (GRSM) using LiDAR data with heights between 55 and 59 m. The two tallest trees measured are potentially the tallest trees ever recorded in the eastern U.S., pending field verification. All ten trees detected in the LiDAR data are taller than the tallest tree recorded to data in the Tennessee portion of the GRSM. The site information for these remarkable trees will be used by park managers, ecologists, and arborists to promote park visitation, as well as to provide the basis for future research of tall trees and their environmental impacts and ecological habitats.

In Chapter 3, Ordinary Least Squares Analysis of a LiDAR-Derived Tree Height Database, a database of tree heights was created using feature extraction software for a 225 hectare area within the GRSM. Over twenty thousand observations were obtained using this process, representative of the large volume of researchable data that can be analyzed using the LiDAR data format. Categorical and continuous variables representing elevation, terrain aspect, terrain slope, overstory vegetation community, distance to water, and soil type at each observed point were used in an Ordinary Least Squares (OLS) analysis to determine the strength of these variables in relation to the measured tree heights. The two models that were created were able to predict 21 – 24 % of the variability in tree heights in the study area. It is proposed that the poor performance of the models was the result of the difficulty in predicting highly variable natural environments. Future studies should include variables that account for climatic conditions at the tree sites, historical data that consider past disturbance in the region, and field research to determine the validity of the dataset created by the automated extraction software. This work also showed that LiDAR data can be valuable as the basis for a database of tree heights. It would be a herculean effort to document the heights of over 22,000 trees using manual field methods and now this process can be automated through feature extraction algorithms currently being used in the field of GIS. Although the fit of the OLS model in this study explained only a portion of the variability in tree heights in the False Gap area, the results were statistically significant and bear further examination.

Both chapters show that using LiDAR to extract accurate elevation and height information in the Great Smoky Mountains National Park allows researchers to obtain information about extraordinary trees and creates interest for arborists, ecologists, park managers, as well as the general public. Upon field verification of the tree sites studied here, it

may be found that these are among the tallest trees of their species ever recorded in the eastern United States. This would be valuable information that could lead to studies of the history of these trees. Further work is needed to examine the history of these sites and determine whether these trees are old-growth or successional.

Finally, the future of LiDAR data applications in forestry research is bright. Higher density point clouds and the introduction of new technologies, such as terrestrial LiDAR and Flash LiDAR mean that those working in the LiDAR remote sensing discipline will have many challenges and opportunities to work with for years to come. Flash LiDAR technology uses an array of sensors to measure ranges over an entire field of view with a single pulse. This technology is being used in military, space, and emergency management situations and has only begun to be explored in the last few years. Terrestrial LiDAR uses ground-based vehicles or platforms instead of airborne platforms to create three-dimensional point clouds which can then be used to create realistic and highly accurate (sub centimeter) models of ground features, including forest structure. These new technologies, as well as the now traditional airborne LiDAR sensors, will continue to create opportunities for researchers in many different fields. The future use of high resolution geospatial data to detect and quantify forest structure will only be limited by the ability for computer systems and processors to manage the large volumes of data that these sensors create.

REFERENCES

American Forests. (2013). Tree Facts. *americanforests.org*. Retrieved February 25, 2013, from http://www.americanforests.org/discover-forests/tree-facts/

American Society for Photogrammetry and Remote Sensing Board. (2008). LAS Specification Version 1.2. ASPRS.

Andersen, H. E., Reutebuch, S. E., & McGaughey, R. J. (2006). A rigorous assessment of tree height measurements obtained using airborne lidar and conventional field methods. *Canadian Journal of Remote Sensing*, *32*(5), 355–366.

Bruckheimer, I. (2011). Great Smoky Mountains National Park: Environmental Suitability for Trees > 180 ft. in Height. Retrieved February 15, 2013, from http://www.ents-bbs.org/download/file.php?id=3745&sid=f980b8b8630a83a09ef250b8a57d7095&mode=view

Coomes, D. A., & Allen, R. B. (2007). Effects of size, competition and altitude on tree growth. *Journal of Ecology*, *95*(5), 1084–1097.

Center for Remote Sensing and Mapping Science-University of Georgia. (2011). Acquisition of Lidar for the Tennessee Portion of Great Smoky Mountains National Park and the Foothills Parkway ARRA LIDAR Task Order)-GSM Project Metadata; USGS Contract: G10AC0015; Proposed Number: NM-ARRA-0073.

U.S. National Park Service.Great Smoky Mountains National Park - Clingmans Dome. (2012). Retrieved February 13, 2012, from http://www.nps.gov/grsm/planyourvisit/clingmansdome.htm

Houk, R. (2000). *Great Smoky Mountains National Park: The Range of Life*. Gatlinburg, TN: Great Smoky Mountains Natural History Association.

Husch, B., Miller, C., & Beers, T. (1972). *Forest mensuration* (2nd ed.). New York: Ronald Press Company.

Jensen, J. L. R., Humes, K. S., Conner, T., Williams, C. J., & DeGroot, J. (2006). Estimation of biophysical characteristics for highly variable mixed-conifer stands using small-footprint lidar. *Canadian Journal of Forest Research*, *36*(5), 1129–1138. doi:10.1139/X06-007

Jensen, J. R. (2007). *Remote Sensing of the Environment An Earth Resource Perspective* (2nd Edition.). Upper Saddle River, NJ: Pearson Prentice Hall.

Jordan, T., & Madden, M. (2011). *Acquisition of LiDAR for the Tennessee Portion of Great Smoky Mountains National Park and the Foothills Parkway* (p. 31). Center for Remote Sensing and Mapping Science: The University of Georgia.

Kenyi, L. W., Dubayah, R., Hofton, M., & Schardt, M. (2009). Comparative analysis of SRTM-NED vegetation canopy height to LIDAR-derived vegetation canopy metrics. *International Journal of Remote Sensing*, *30*(11), 2797–2811.

Kozlowski, T. T. (1971). *Growth and development of trees. Cambial growth, root growth, and reproductive growth* (Vol. 2). New York: Academic Press.

Laurance, W. (2012). How the mighty are fallen. *New Scientist*, *213*(2849), 39–41.

Madden, M., Welch, R., Jordan, T., Jackson, P., Seavey, R., & Seavey, J. (2004, July). Digital Vegetation Maps for the Great Smoky Mountains National Park. Center for Remote Sensing and Mapping Science.

Magnussen, S., Næsset, E., & Gobakken, T. (2010). Reliability of LiDAR derived predictors of forest inventory attributes: A case study with Norway spruce. *Remote Sensing of Environment, 114*(4), 700–712. doi:10.1016/j.rse.2009.11.007

Maune, D. F. (2001). *Digital Elevation Model Technologies and Applications: The DEM User's Manual*. Bethesda: American Society for Photogrammetry and Remote Sensing.

McGaughey, R. J., Carson, W. W., Reutebuch, S. E., & Andersen, H. E. (2004). Direct measurement of individual tree characteristics from LIDAR data. In *Proceedings of the 2004 Annual ASPRS Conference, May 23–28 2004*.

McMaster, R., & Usery, E. L. (Eds.). (2005). *A Research Agenda for Geographic Information Science*. Boca Raton: CRC Press.

Moore, H. L. A. (1988). *A Roadside Guide to the Geology of the Great Smoky Mountains National Park*. Knoxville, TN: University of Tennessee Press.

Naesset, E. (1997). Determination of mean tree height of forest stands using airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *52*, 49–56.

National Park Service. (1981). *Great Smoky Mountains National Park - North Carolina and Tennessee*. National Park Service.

Nilsson, M. (1996). Estimation of tree heights and stand volume using an airborne LiDAR system. *Remote Sensing of Environment*, *56*, 1–7.

Overwatch Systems, Ltd. (2010). LIDAR Analyst 5.0 for ArcGIS Reference Manual. Overwatch Systems, Ltd.

Petit, G., Anfodillo, T., Carraro, V., Grani, F., & Carrer, M. (2011). Hydraulic constraints limit height growth in trees at high altitude. *New Phytologist*, *189*(1), 241–252.

Popescu, S. C., Wynne, R. H., & Nelson, R. F. (2002). Estimating plot-level tree heights with lidar: local filtering with a canopy-height based variable window size. *Computers and Electronics in Agriculture*, *37*(1-3), 71–95.

Renslow, M. (Ed.). (2012). *Manual of Airborne Topographic Lidar*. Bethesda, MD: American Society for Photogrammetry and Remote Sensing.

Rucker, C. (2004). GSMNP tall trees. Retrieved February 29, 2012, from http://www.nativetreesociety.org/fieldtrips/gsmnp/gsmnp_tall_trees.htm

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. (2011, April 12). Web Soil Survey - Citation. *Web Soil Survey*. Retrieved March 13, 2013, from http://websoilsurvey.nrcs.usda.gov/app/Help/Citation.htm

Suárez, J. C., Ontiveros, C., Smith, S., & Snape, S. (2005). Use of airborne LiDAR and aerial photography in the estimation of individual tree heights in forestry. *Computers & Geosciences*, *31*(2), 253–262.

U.S. Forest Service, D. (2005). How to Measure a Big Tree. *How to Measure a Big Tree*. Retrieved from http://www.fs.fed.us/r6/uma/nr/silv/Tree%20Measurements.pdf

United States Department of Agriculture, National Resources Conservation Service. (2009). *Soil survey of Great Smoky Mountains National Park, Tennessee and North Carolina*.

Ussyshkin, V., & Theriault, L. (2011). Airborne Lidar: Advances in Discrete Return Technology for 3D Vegetation Mapping. *Remote Sensing*, *3*(3), 416–434. doi:10.3390/rs3030416

Walker, S. L. (1991). *Great Smoky Mountains: The Splendor of the Southern Appalachians*. Charlottesville, VA: Elan Publishing.

Welch, R., Madden, M., & Jordan, T. (2002). Photogrammetric and GIS techniques for the development of vegetation databases of mountainous areas: Great Smoky Mountains National Park. *ISPRS Journal of Photogrammetry and Remote Sensing*, *57*(1–2), 53–68.

Whittaker, R. (1956). Vegetation of the Great Smoky Mountains. *Ecological Monographs*, *26*, 1–80.

Zimble, D. A., Evans, D. L., Carlson, G. C., Parker, R. C., Grado, S. C., & Gerard, P. D. (2003). Characterizing vertical forest structure using small-footprint airborne LiDAR. *Remote Sensing of Environment*, *87*(2–3), 171–182.