

**Investigating trends and processes that impact tree species
diversity of southern forests at the landscape level**

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Andrew J. Hartsell
December 2017**

DEDICATION

I dedicate this research to my family. My parents, Ben and Mary Ann Hartsell, who taught me to love and respect the environment and developed a desire to continue to continually grow and improve as a person. My wife, Cynthia Barnes Hartsell, brought out the best in me while living with the distraction caused by this work. Finally my children, Ben, Dylan, and Sarah, who give me hope for the future and provide inspiration to maintain our planets natural resources.

ACKNOWLEDGEMENTS

I want to thank the people who helped mold my professional and scientific career. It all started when I enrolled as an undergraduate in the School of Forestry at Stephen F. Austin State University (SFASU) in the fall of 1985. I was not prepared for college life intellectually, emotionally or academically. The curriculum and professors at SFASU forced me to address these deficiencies. It was an awkward and painful process, but the process eventually yielded positive results. I was a confident and capable forester when I left SFASU. I want to particularly thank professors Michael Fountain, Monty Whiting, Kenneth Waterston, and James Kroll for their impact on my life. Additionally, this portion of my life's journey would not have been complete without the help of my classmates and friends. During this time I learned more about life from my two best friends, Ski and Derrick, than any other non-academic source. Other companions of note were Blake, Gordon, James, Jay, Jeff, Mark, Scotty, and Steve. Gordon would later earn his PhD and become a professor himself. Thanks to all of you for helping me grow from a confused teenager to a productive professional forester.

I began working for the U.S. Forest Service, Forest Inventory and Analysis (FIA) research unit soon after graduation. I was a plot data collector for FIA. This was a tough job. It required being on travel status 365 days a year. All work was performed outside and in some of the most remote areas of the south, in all weather conditions. Many of my classmates from SFASU were part of the FIA field crew at this time. The person who would have the most significant impact on me was the field supervisor, D. Vince Few. Vince required a great deal from us, but he provided the necessary training and insight on how to perform our tasks. He was a tough but fair taskmaster. I will carry the lessons and mindset I learned from throughout my life.

I moved to the FIA office during the fall of 1992. John Vissage was my supervisor and it was from him that I really learned about computers, math, and big data. John suggested and encouraged me to pursue my masters at Mississippi State University (MSU). John left to pursue his career elsewhere so Pat Miller and Joanne Lenehan successively took his place. Both contributed to and encouraged both my professional and academic development.

My original field of study for my masters was silviculture, and John Hodges was the best forest ecologist I have ever met. He was an even better person. John was my first major professor. Dr. Hodges and Frank Miller, who introduced me to GIS, revealed the underpinnings of forest dynamics to me. However, both of these gentlemen retired before I could finish my degree. Steve Bullard became my major professor and I changed my focus to forest economics. Steve, along with Keith Belli and Don Hodges, moved my thinking from

ecological to economic. I earned my MSF in Forest Economics during the spring of 1999. This would not have been possible without those that I mentioned above.

I moved to Knoxville in 2003 and soon began thinking about a PhD. This process gained traction when I realized one of my professors from MSU, Don Hodges, was a professor at the University of Tennessee (UT). I began a decade long adventure trying to earn a PhD with Don as my major professor. We had to overcome a few major roadblocks, but we managed to do it. Thanks Don for your insight and patience. I am truly thankful.

I would be remiss if I did not mention Tom Brandeis. Tom is my current supervisor at FIA and a member of my UT graduate committee. He brought a level of scientific rigor with him and applied it to all he supervised. I needed this structure to help me grow as a scientist. Additionally, he was an advocate for my academic advancement and he proved to be a beneficial “sounding board” when I encountered issues with work or school.

In closing, I want to say “Thank You” to all of those mentioned as well as many more who were not. Each of you had a vital role in helping me fulfill my professional and academic goals. I owe you all a great debt. I hope I can repay this debt by sharing the lessons I’ve learned from each of you to those I encounter throughout my life.

ABSTRACT

Global biodiversity is currently being threatened by multiple anthropogenic sources. The increase in genetically modified, artificially regenerated forests has raised concerns that this management scheme may be a source of species loss. This study investigates the impacts of urbanization, agricultural development, and southern yellow pine plantation establishment on tree species evenness, Shannon-Wiener index, and Simpson's index of forest communities within the Gulf Coastal Plain of the southeastern United States. The impacts of varying spatial scale on possible predictors was investigated by utilizing nested watersheds from the Watershed Boundary Dataset (WBD). Techniques used in the analysis included Getis-Ord G_i^* analysis, Local Moran's I , non-metric multidimensional scaling (NMS), multiple response permutation procedures, (MRPP) and spatial lag, spatial error, and geographically weighted regression techniques using ArcGIS, GeoDa, and PC-Ord software. The U.S. Forest Service Forest Inventory and Analysis inventory data were used as the source of all tree species data. The 100-meter resolution elevation raster data from the U.S. Geological Survey provided elevational data, while the 1992 200-meter national land cover data (NLCD) were used to derive land use patterns. The 1981-2010 parameter regressions on independent slopes model (PRISM) provided climatic data. All raster data were averaged to watershed basin. Conclusions from this study are 1) watersheds can be used for nested studies; 2) HUC04 and HUC06 watersheds are appropriate scales for FIA data and tree diversity studies; 3) forest area, productivity, and elevation are positively correlated to most diversity measures; 4) the Pine Belt has been repeatedly disturbed due to anthropogenic activity; 5) tree species diversity is impacted by these disturbances; 6) these impacts can be both positive and negative, as many disturbances increase species richness; 7) plantations increase the number of species, perhaps due to edge effects from management; and 8) increasing plantations has a negative impact on both Shannon's diversity index and Simpson's index as both decreases with plantation acreage.

TABLE OF CONTENTS

CHAPTER I INTRODUCTION	1
Biodiversity	1
Factors that influence diversity	3
Area and Scale	3
Energy and Water	4
Nutrition	5
Elevation	5
Latitude	5
Disturbance and Stability	6
Plantations	6
Research Topics	7
Appendix	9
CHAPTER II Defining the impacts of plantation forestry on biodiversity in Alabama and Mississippi	13
Abstract	14
Introduction	14
Hypothesis	15
Data, Diversity Indices and Spatial Unit	15
Tree and plot data	15
GIS Data Layers	16
Diversity Indices	16
Species Richness	16
Shannon-Wiener Index	17
Simpson's Index	17
Methods	18
Model Development	18
Ordinary Least Squares and Spatial Regression (OLS)	18
Dummy Variables	19
Spatial Lag Model (SLM)	20
Spatial Error Model (SEM)	21
Model Comparison and Goodness of Fit	21
Choosing the best regression model	23
Results	24
Plantation establishment	24
Exploratory data analysis	24
Spatial patterns & clustering	25
Spatial autocorrelation	26
Change in species richness	26
Change in Shannon's diversity index	26
Change in Simpson's diversity index	26
Discussion	27
Correlation analysis	27

Change in species richness	27
Change in Shannon's diversity Index.....	28
Change in Simpson's diversity index	28
Conclusions.....	29
References	30
Appendix	33
CHAPTER III The impacts of varying spatial scale in determinng predictors of tree species diversity using nested watersheds and Forest Inventory data in the southeastern United States.....	55
Abstract	56
Introduction	56
Study Objectives and Hypotheses.....	57
Data, Diversity Indices and Spatial Unit	58
GIS Data Layers	58
Diversity Indices.....	59
Species Richness.....	59
Shannon-Wiener Index	60
Shannon's Index	60
Methods	61
Hot-Spot Analysis	61
Regression Trees	62
Nonmetric-Multidimensional Scaling.....	62
Model Development.....	63
Results	64
Discussion.....	68
Conclusions.....	70
References	71
Appendix	75
CHAPTER IV The impacts of varying spatial scale upon predcitions of three measure of tree species diverstiy.....	110
Abstract.....	111
Introduction	111
Hypothesis	112
Data.....	112
Tree and Forest Data.....	112
GIS Data Layers	113
Methods	114
Results	115
Discussion.....	116
Conclusions.....	118
References	119
Appendix	122
CHAPTER V CONCLUSIONS	132
APPENDIX – DATA & METHODS	135

Forests of the Southern United States	136
Tree and plot data.....	137
GIS Data Layers	137
Diversity Indices.....	138
Species Richness.....	138
Shannon-Wiener Index	139
Simpson's Index.....	139
Multiple Comparison Procedures.....	140
Analysis of Variance (ANOVA).....	140
Multi-Response Permutation Procedures (MRPP)	141
Examples of MRPP	141
Multivariate Analysis	142
Classification Analysis.....	142
Ordination Analysis	143
Non-metric Multidimensional Scaling (NMS)	143
Examples of NMS	145
Classification and Regression Trees (CART).....	146
Examples of Regression Trees	149
Species Area Curves	150
Spatial Autocorrelation.....	151
Moran's I	152
Geary's C	153
Gettis-Ord General G	153
Hot Spot Analysis.....	154
Getis-Ord Gi* (Hotspot Analysis).....	155
Local Moran's I (Cluster Analysis).....	155
Examples of Local Moran's I (Cluster Analysis)	156
Model Development.....	157
Ordinary Least Squares and Spatial Regression (OLS).....	157
Dummy Variables	158
Spatial Lag Model (SLM).....	159
Spatial Error Model (SEM)	160
Model Comparison and Goodness of Fit.....	160
Choosing the best regression model.....	162
Examples of spatial regression.....	163
Geographically Weighted Regression (GWR).....	163
Examples of GWR	165
VITA.....	167

LIST OF TABLES

Table 1. Factors that influence biodiversity	10
Table 2. Area of forestland by state, survey period and stand origin, Alabama and Mississippi	10
Table 3. Number of live trees greater than or equal to one inch DBH in Alabama 1972 and Mississippi 1977 (top 15 species)	11
Table 4. Number of live trees greater than or equal to one inch DBH in Alabama and Mississippi 2012 (top 15 species)	12
Table 5. Variable name and description for all observations in dataset.	45
Table 6. Descriptive statistics for species richness, Simpson's index, and Shannon's index for Alabama 1972 and Mississippi 1977 and Alabama and Mississippi 2013.	46
Table 7. Correlation coefficients for the 1970s diversity indices	47
Table 8. Correlation coefficients for the 2013 diversity indices	49
Table 9. Correlation coefficients for the 1970s-2013 change in species richness, Simpson's index, and Shannon's index.	51
Table 10. Spatial lag model results for percent change in richness between the 1970s and 2013.	53
Table 11. Ordinary least squares model results for percent change in Shannon's diversity index between the two time periods.	53
Table 12. Spatial lag model results for percent change in Simpson's diversity index between the two time periods.	54
Table 13. Watershed sizes and the average, maximum, minimum, and total number of FIA plots in each HUC class	95
Table 14. Variable list and description for all watersheds	95
Table 15. MRPP results of HUC04 watersheds	96
Table 16. MRPP results of HUC06 watersheds	97
Table 17. Two dimensional NMS ordination solution of HUC04 watersheds	97
Table 18. Two dimensional NMS ordination solution of HUC06 watersheds	98
Table 19. Three dimensional NMS ordination solution of HUC08 watersheds....	99
Table 20. Three dimensional NMS ordination solution of HUC10 watersheds....	99
Table 21. Two dimensional NMS ordination solution of HUC08 watersheds categorized as being 'cold spots' for Simpson's diversity index	100
Table 22. Regression models for species richness, HUC04 watersheds	101
Table 23. Regression models for Shannon-Weiner diversity index, HUC04 watersheds	101
Table 24. Regression models for Simpson's diversity, HUC04 watersheds	102
Table 25. Regression models for species richness, HUC06 watersheds	102
Table 26. Regression models for Shannon-Weiner diversity index, HUC06 watersheds	103

Table 27. Regression models for Simpson's diversity, HUC06 watersheds.	103
Table 28. Regression models for species richness, HUC08 watersheds.....	104
Table 29. Regression models for Shannon-Weiner diversity index, HUC08 watersheds	105
Table 30. Regression models for Simpson's diversity, HUC08 watersheds	106
Table 31. Regression models for species richness, HUC10 watersheds.....	107
Table 32. Regression models for Shannon-Weiner diversity index, HUC10 watershed	108
Table 33. Regression models for Simpson's diversity, HUC10 watersheds.	109
Table 34. Watershed sizes and the average, maximum, minimum, and total number of FIA plots in each HUC class	126
Table 35. Variable list and description for all watersheds	126
Table 36. Effects of varying average all-live growth per acre and total forest area on OLS model on species richness for HUC04 watersheds.	127
Table 37. Effects of varying average percent agriculture and total forest area on OLS model on Shannon-Wiener diversity index for HUC04 watersheds. ..	128
Table 38. Effects of varying average percent agriculture and percent of watershed in plantations of OLS model on Simpson's diversity index for HUC04 watersheds.....	128
Table 39. Effects of varying average all-live hardwood growth/acre, total forest area (acres), and potential evapotranspiration (PET) of OLS model on tree species richness for HUC06 watersheds.	128
Table 40. Effects of varying percent of watershed in agriculture and total forest area (acres) of OLS model on tree Shannon-Wiener index for HUC06 watersheds.	129
Table 41. Effects of varying percent of forests in plantations and average hardwood tree growth/acre of OLS model on tree Simpson's index for HUC06 watersheds.....	129
Table 42. Geographically weighted regression model for species richness of HUC08 watersheds.....	129
Table 43. Effects of varying percent of watershed in agriculture, percent of watershed in forests and total forest area (acres) of GWR model on tree species richness for HUC08 watersheds.	130
Table 44. Geographically weighted regression model for Shannon-Wiener diversity index of HUC08 watersheds.	130
Table 45. Effects of varying percent of watershed in agriculture, percent of watershed in forests and percent of forests in plantations of GWR model on Shannon-Wiener diversity index for HUC08 watersheds.	130
Table 46. Geographically weighted regression model for Simpson's diversity index of HUC08 watersheds.	131
Table 47. Effects of varying percent of watershed in agriculture, percent of watershed in forests and percent of forests in plantations of GWR model on Simpson's diversity index for HUC08 watersheds.	131

LIST OF FIGURES

Figure 1. Forestland cover in the southeastern United States	9
Figure 2. Study area	33
Figure 3. Species richness (number of tree species) by county, 2013.....	34
Figure 4. Shannon-Weaver diversity index by county, 2013	35
Figure 5. Simpson's diversity index by county, 2013	36
Figure 6. Plantation acreage, in acres, by county, 2013	37
Figure 7. Forestland area, in acres, by county, 2013.....	38
Figure 8. Percent of land base in urban land uses by county, 2011	39
Figure 9. Change in species richness between 1970s and 2013 by county.....	40
Figure 10. Change in Shannon-Weaver diversity index between 1970s and 2013 by county	41
Figure 11. Change in Simpson's diversity index between 1970s and 2013 by county	42
Figure 12. Percent change in plantation acreage between 1970s and 2013 by county	43
Figure 13. Spatial autocorrelation report and Moran's I for change in species richness	44
Figure 14. HUC08 basins nested within HUC04 watersheds.....	75
Figure 15. Species - number of plots curves for all plots on forest industry and national forest lands with lines denoting the average size of HIC04, HOC06, and HUC08 watersheds.....	76
Figure 16. Species basal-area curves for plots that occur on either national forests or forest industry	77
Figure 17. Distribution of species richness, Shannon-Wiener Index, Simpson's Index, and planted forests (thousand acres) for HUC04, HUC08, HUC08, and HUC10 watersheds.....	78
Figure 18. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HUC04 watersheds	79
Figure 19. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HUC06 watersheds	80
Figure 20. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average	

temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HUC08 watersheds.....	81
Figure 21. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HUC10 watersheds	82
Figure 22. Regression tree for species richness, HUC04 watersheds	83
Figure 23. Regression tree for Shannon-Wiener diversity index, HUC04 watersheds	84
Figure 24. Regression tree for Simpson's diversity index, HUC04 watersheds ..	85
Figure 25. Regression tree for species richness, HUC06 watersheds	86
Figure 26. Regression tree for Shannon-Wiener diversity index, HUC06 watersheds	87
Figure 27. Regression tree for Simpson's diversity index, HUC06 watersheds ..	88
Figure 28. Regression tree for species richness, HUC08 watersheds	89
Figure 29. Regression tree for Shannon-Wiener diversity index, HUC08 watersheds	90
Figure 30. Regression tree for Simpson's diversity index, HUC08 watersheds ..	91
Figure 31. Regression tree for species richness, HUC10 watersheds	92
Figure 32. Regression tree for Shannon-Wiener diversity index, HUC10 watersheds	93
Figure 33. Regression tree for Simpson's diversity index, HUC10 watersheds ..	94
Figure 34. HUC08 basins nested within HUC04 watersheds	122
Figure 35. Map of residuals, intercept, percent agriculture, percent forest, total forest, and all-live growth coefficients from GWR model of species richness for HUC08 watersheds.	123
Figure 36. Map of residuals, intercept, percent agriculture, percent forest, total forest, and all-live softwood growth, and percent of forests in plantation coefficients from GWR model of Shannon-Wiener index for HUC08 watersheds.	124
Figure 37. Map of residuals, intercept, and percent of forests in plantations and percent of watershed in agriculture and forest coefficients from GWR model of Shannon-Wiener index for HUC08 watersheds.	125
Figure 38. Land uses of the southeastern United States, 2011 NLCD.....	136

CHAPTER I

INTRODUCTION

The Southern Forest Resource Assessment (SFRA) (Wear and Greis, 2002) was initiated in the late 1990s to investigate the history and status of, and possible threats to, the forest resources of the southern United States. Part of this process involved a series of public meetings where ideas and concerns of various groups could be heard. Many groups voiced a concern that southern pine plantations posed a threat to the diversity of southern forests (USDA, 2002). This concern can be distilled into researchable questions: What is diversity? What anthropogenic forces, particularly plantation forestry, impact diversity? How does diversity and covariates of diversity change with spatial scale? How can Forest Inventory and Analysis (FIA) data be used to answer these questions? The final SFRA report addressed many issues, but failed to address many of the issues pertaining to diversity stated above.

The South, as defined in this study, ranges across 13 states from Virginia to Texas. Forests dominate the landscape, as this land use constitutes 40 to more than 80 percent of the land in many counties, with a few exceptions in the western grasslands of Texas and Oklahoma, and in key agricultural areas in the Mississippi Valley, southern Florida, and parts of Kentucky and Tennessee. Forest types ranging from upland hardwoods, natural and artificial pine stands, lowland gum-cypress swamps, high elevation spruce-fir, and coastal mangrove/live-oak forests can be found within the southern region. Upland hardwoods dominate the Piedmont and northern Gulf Coastal Plain. Pine is the dominant type for most of the Gulf Coastal Plain, while lowland forest types dominate the Mississippi Alluvial Valley (Jin et al, 2013). Total forest area has been stable since the 1970s (Figure 1).

The forests of the South support a diverse biota, containing 1,027 native terrestrial vertebrate species; 178 amphibians, 504 birds, 158 mammals, and 187 reptiles. Recent increases in nonnative, invasive plants, insects, and diseases have affected the forest landscape as well (Wear & Gries, 2012). Emerald ash borer and laurel wilt have impacted many sites, causing a loss in host species. Additionally, nonnative tree species such as Chinese tallowtree (*Triadica sebifera*), chinaberry (*Melia azedarach*), and others have been spreading across the landscape replacing or competing with native species (Miller, 2003).

Biodiversity

Biodiversity, or biological diversity, can be defined as “the variety and variability among living organisms and the ecological complexes in which they occur” (OTA 1987). Humans perceive regions with a multitude of diverse species

to have more value than those that do not (Wilson 1993; Ehrlich & Wilson 1991). Possible reasons that species diversity is often valued by humans are larger numbers of plant species means a greater variety of crops and life; greater species diversity helps assure natural sustainability for all life forms; diverse ecosystems can better withstand and recover from a variety of disasters; and the planet's complex systems, ecological networks, and energy flows are dependent upon numerous organisms and interactions (Ehrlich & Wilson 1991, Gaston 1996).

A number of studies have linked species diversity to forest ecosystem function and productivity (Caspersen & Pacala 2001; Tilman et al. 1997). Species diversity is frequently used as a surrogate for forest stability, health, and productivity (McIntosh 1967; McNaughton 1977; Tilman 1996; Austin 2002). Species richness, or simply the number of species in an area, is currently the most popular measure of diversity (Stirling & Wisley 2001). Relative species abundance has been examined as well (Whittaker 1965; Hulbert 1971). Proportional indexes such as the Shannon-Wiener index (H') have been hypothesized to have a strong correlation as well (Hill 1973; Routledge 1980). Modern conservation planning is dependent upon the understanding a prediction of species distributions (Franklin 1993; Austin 1988; Guisan & Zimmerman 2000; Elith & Burgman 2002).

The main factors responsible for biodiversity loss include land use change, habitat change such as forest fragmentation and conversion, invasive alien species, overexploitation, climate change, and pollution (Hanski 2005). The impacts of intensive southern yellow pine plantation forestry on biodiversity are a contentious subject. Linder (2004) explains how this type of management could negatively impact biodiversity of southern forests. The June 1999 public review of the assessment questions for the Southern Forest Resource Assessment (Wear & Greis 2002) contains multiple examples of individuals and organizations expressing concern that this type of silviculture could negatively affect biodiversity. It is important that science ascertains the positive and negative impacts of this management regime, in order to facilitate informed public discourse and planning.

This study focuses on two types of biological diversity measures: richness and evenness. Richness is the number of different organisms or species living in an area (Wilson 1993; Ehrlich & Wilson 1991). It does not consider the proportion or distribution of each subspecies within the area of interest. All one must do is simply count the number of subspecies found in a community. It is the simplest and easiest diversity measurement to calculate and describe. Species richness is a key component in many the plant ecology community (Naeem et al 1996).

Relative species abundance within a community is often measured with an index ranging from 0 to 1 and equated to species evenness (Hulbert 1971; Alatalo 1981). Evenness refers to how the subspecies are distributed throughout the community (Routledge 1980; Magurran 1988). It, in effect, is a normalization of richness. For example, there are two communities composed of five different subspecies. The first one has counts of 20 for each subspecies for a total of 100 individual subspecies counted. The second one has one subspecies with a count of 96, and the other four have one observation each. The second community also has a total of 100 individual subspecies counted. The first community would be considered more even, as each subspecies accounts for the same percentage of the total, while the second one is dominated by a single species.

Factors that influence diversity

A review of literature reveals a multitude of factors that influence species distributions, richness, and diversity. Six factors were considered in the subsequent analyses and are described below. These factors and how they influence species richness and diversity are listed in Table 1. A review of the literature and discussion of each factor and the variables in the dataset that account for these factors follows.

Area and Scale

No universal scale exists for assessing species patterns and diversity studies (Levin 1992). Scale should be selected based on the study goals, the ecological system being investigated, and the data sets used in the analysis (Cushman & McGarigal 2002; Pearson & Dawson 2003). Climatic variables dominate species distributions for terrestrial systems at the global scale, while energy, water systems, nutrient availability, and parent material are more important at smaller, regional scales (Mackey & Lindermayer 2001). What is clear is that area control is vital if the examination and selection of variables is important to the study (Whittaker et al. 2001). That is, equal area grids should be used.

The relationship between scale and species richness and diversity has long been studied (Gleason 1922; Cain 1938; MacArthur & Wilson 1967). As the size of area in question increases, so does species richness and indices of diversity and evenness. This phenomenon is represented by species-area curves which depict this relationship (Cain 1938; MacArthur & Wilson 1967). These curves trace the accumulation of species with increasing sample size.

Confounding the analysis of scale is the distinction between geographic and environmental space. Geographic relationships are typically defined as two-

and sometimes three-dimensional models. Environmental systems are typically multi-dimensional in nature and contain a host of interacting relationships. Predictions are based on the location of the sample unit in the environmental space, but frequently analyzed in the geographic (Elith & Leatherwick 2009).

On larger scales, regional environmental conditions have stronger impacts on species richness (Currie 1991). Solar radiation, temperature, and precipitation in particular have been proven to have significant impacts on diversity (Brown 1981; Wright 1983). The impacts of the variables are described in the following paragraphs.

Energy and Water

Solar radiation and precipitation have a profound impact on tree species richness. Currie & Paquin (1987) reported that 76% of the variability in tree species richness can be explained by actual evapotranspiration (AET). Many studies utilize potential evapotranspiration (PET) instead of AET (Currie 1991; O'Brien 1993, 1998; Whittaker & Field, 2000; Whittaker et al. 2001).

A strong correlation between energy and richness exists at larger scales. Solar radiation, temperature, and precipitation have proven to be strong predictors of richness (Richerson and Lum 1980; Wright 1983; Currie 1991). Hutchinson (1959) introduced the idea that regional species richness was dependent upon energy. This concept was developed further by Connell & Orias (1964), Brown (1981), and Wright (1983). Wright studied animal richness on islands and concluded that richness on these land masses worldwide was related to the solar energy each island receives. A total energy hypothesis was developed, predicting that holding size, scale, and area constant, energy flux per unit area is the prime determinant of species richness. Both Wright and Currie discussed the species-energy hypothesis where the amount of available energy bounds the limits of species richness for the system. Again, measures of energy included temperature, AET, and PET. Rosenzweig (1968) stated that the best correlate for energy is AET.

It is not the question on the total amount of energy, but the availability of the energy to the population in question. O'Brien developed a "water-energy dynamics" theory and found that the best two-variable model for her studies in South African forests was a linear relationship with annual rainfall and PET. In general, richness is thought to increase as a linear relationship with rainfall, and a parabolic function with energy.

Nutrition

The species composition/soil nutrient relationship has been rigorously studied for some time. Prevailing wisdom would lead to an assumption that higher productivity leads to higher diversity. However, this is not always the case (Huston 1980; Waide, et al. 1999). A study of 46 Costa-Rican forest sites revealed that the highest species richness occurred on the poorest sites (Huston 1980). Other studies failed to find any statistical significance between nutrition and richness (Poore, 1968; Kwan & Whitmore 1970). Waide, et al (1999) conducted a comprehensive literature review of 200 relationships that indicated that 32% yielded a unimodal relationship between richness and productivity, 26% were positively linear, 12% negatively linear and the final 32% showed no correlation. Numerous other factors associated with richness such as plant growth, distribution, structural complexity, disturbance, and edaphic conditions were listed as potential reasons for these results. Huston (1979) concluded that species diversity would be low in areas of high nutrition and infrequent disturbances because rapid growth leads to competitive exclusion.

Elevation

The impact of elevation on species communities has been observed for a long time. Linnaeus (1781), von Humboldt and Bonpland (1799), and Holdridge (1947) observed the relationship between elevation and species communities. Linnaeus (1781) hypothesized that species diversity would be highest at high elevations because these areas would be free of floods which he concluded led to lower diversity. Von Humboldt and Bonpland (1799) noticed that climatic and site locations changed with elevation in the Andean's mountains, and therefore changed species composition. Holdridge (1947) developed "life zones" based on climatic and elevational attributes.

Elevational diversity gradient (EDG) is derived from these and other studies. EDG states that different climates, sites, and species occur at different elevations and species richness will increase as elevation increases. However, this increase will occur up to a certain point where elevation, climate, and site will start to have a negative impact on richness. This results in a diversity bulge in middle elevations.

Latitude

Latitude can have the same effect as elevation. Many studies reveal a decline in biodiversity at the poles. This phenomenon is often referred to as the latitudinal diversity gradient (Colwell & Hurtt 1994; Lyons & Willig 1997; Whittaker et al. 2001). The underlying theme is that species richness and diversity are lower at the poles and higher along the equator. The causes and impacts of

latitudinal gradient are dependent upon scale. Latitudinal gradient was found to be significant in global and regional datasets (Lyons & Willig 1997). Other variables often act as surrogates for latitude in studies confined to smaller areas. These surrogates may be energy and water supply, along with decreasing scale sizes where the gradient is the same across the study area.

Disturbance and Stability

The concept of competitive exclusion plays an important role in species richness. Often called Gause's Law of Competitive Exclusion, the basic concept is that two species competing for the exact same resources cannot stably coexist (Gause 1934). If one species has even a slight advantage over the other, the lesser of the two will eventually go extinct. The other option is for the disadvantaged species to evolve to occupy a different niche (Hardin 1960).

Current disturbance theory is centered on the intermediate disturbance hypothesis (IDH) originally presented by Connell (1978). IDH predicts that species diversity will be low in areas of low and/or infrequent disturbance, as competitive exclusion processes will have developed a dominant species. However, species diversity will also be low in areas of high and/or frequent disturbance, because only a few species are tolerant of such landscapes. These are generally pioneer species that first appear after disturbance. Species diversity will be highest only at levels of intermediate disturbance due to a mix of colonizers and competitors.

One important artifact of disturbances is that they would appear to be independent of climatic control variables, especially energy and water. Hurricanes, fire, anthropogenic forces, volcanoes, and glaciers have the potential to affect species richness independent of normal climatic forces (Whittaker et al 2001).

Plantations

Timber harvesting and management are specific forms of disturbance that play a significant role in southern forests. Timber has been harvested in the South since before the arrival of European colonists, who then intensified this practice. Much of this early harvesting can be associated with land clearing and the conversion of forests to farms (Williams 1989). Timber harvesting has continued to increase, doubling since 1950 (USDA Forest Service, 2012). Landowners have responded to this increased demand by establishing more planted pine stands. These forests are often regenerated by a single pine species that is often genetically modified to improve growth. Additionally, these plantations are managed on a shorter rotation than in previous years, ranging today from 15 to 35 years (Allen et al, 2005).

The majority of the forests in the South are controlled by private land owners. Currently 86 percent of the regions forests are owned by non-governmental agencies and are classified as private. This includes individuals and families as well as many corporate and industrial lands. While the percentage of owners in the private category has remained fairly stable, the types of landowners have recently changed. Many lands owned by timber industry have been sold to Timberland Investment Management Organizations (TIMOs) and Real Estate Investment Trusts (REITs). Parcelization has increased as well, resulting in a more fragmented landscape (Wear and Gries, 2002).

Table 2 illustrates the reason for concern, as it shows the increase in pine plantation acreage for two states in the study area, Alabama and Mississippi. These two states account for almost 20 percent of the South's standing softwood volume (Brandeis et al, 2012). Planted stands accounted for less than 10 percent of the total forest acreage for these states during the 1970s. Today, these forests account for over 31 percent of the forested area.

Species composition for the study area has changed over time. Table 3 details the most abundant species present in Alabama in 1972 and Mississippi in 1977. Table 4 reveals the same results for 2013. While the top two species remained the same, the percentage of loblolly pine almost doubled. In fact, the latest inventory numbers indicate that one in four trees in Alabama was a loblolly pine (Hartsell & Cooper, 2013). The data illustrate why many are concerned that plantation forestry poses a potential threat to species diversity in Southern forests.

Research Topics

This study is focused on three broad research areas centered on the relationships between southern tree species diversity and anthropogenic forces, particularly the impacts of southern pine plantation establishment. Three measures of tree species diversity were used: species richness, Shannon-Weaver index, and Simpson's index.

The first paper will ascertain if tree species diversity has changed over time, developing models that explain any changes. Current and historic FIA data for Alabama and Mississippi were used. The second topic is centered on aspects of spatial scale. This research used current FIA data in conjunction with nested watersheds as spatial units. Questions examined included determining the appropriate spatial scales for tree diversity studies and FIA data. The ability of watersheds to partition variance was explored also, as were the impacts of various scales on predictors of diversity. The last paper utilized the results of the two previous studies to develop a South -wide model of tree species diversity.

This model then was used to perform prediction scenarios based on changes in population, urbanization, agriculture, pine plantation establishment, and climate change.

Appendix

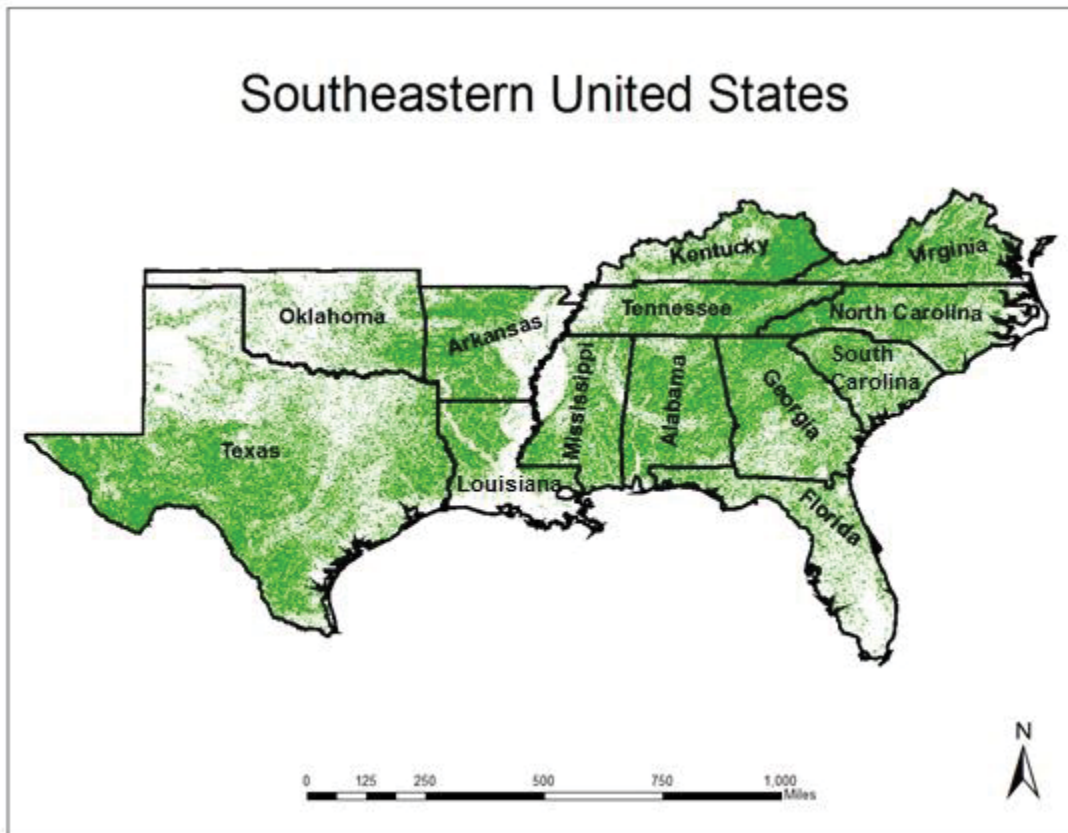


Figure 1. Forestland cover in the southeastern United States

Table 1. Factors that influence biodiversity

Factor	Impact
Area or scale	Richness and diversity increase with area
Energy & water	While these are usually discussed separately, areas with higher potential evapotranspiration or actual evapotranspiration have higher diversity indicators.
Nutrition	Higher nutrition results in higher diversity
Latitude & Elevation	Diversity is lower at poles and higher at equator
Disturbance & stability	Moderate disturbance retards competitive exclusion. More time permits more complete colonization and evolution of new species, however competitive exclusion may yield fewer species

Table 2. Area of forestland by state, survey period and stand origin, Alabama and Mississippi

State	Forest area 2013	Natural stands 2013	Planted stands 2013	Forest area 1970s	Natural stands 1970s	Planted stands 1970s
<i>----- million acres -----</i>						
Alabama	22.9	15.6	7.3	21.4	19.6	1.7
Mississippi	19.5	13.5	6	16.7	14.8	1.8
Total	42.4	29.1	13.3	38	34.5	3.6

Table 3. Number of live trees greater than or equal to one inch DBH in Alabama 1972 and Mississippi 1977 (top 15 species)

Common name	Scientific name		Number of trees	Percent of total
loblolly pine	<i>Pinus</i>	<i>taeda</i>	3,659.4	12.7%
sweetgum	<i>Liquidambar</i>	<i>styraciflua</i>	3,507.9	12.2%
shortleaf pine	<i>Pinus</i>	<i>echinata</i>	2,237.1	7.8%
Other or unknown live tree	<i>Tree</i>	<i>unknown</i>	1,844.1	6.4%
flowering dogwood	<i>Cornus</i>	<i>florida</i>	1,743.7	6.1%
hickory spp.	<i>Carya</i>	<i>spp.</i>	1,590.6	5.5%
red maple	<i>Acer</i>	<i>rubrum</i>	1,312.0	4.6%
blackgum	<i>Nyssa</i>	<i>sylvatica</i>	1,270.6	4.4%
water oak	<i>Quercus</i>	<i>nigra</i>	1,004.7	3.5%
southern red oak	<i>Quercus</i>	<i>falcata</i>	878.6	3.1%
post oak	<i>Quercus</i>	<i>stellata</i>	804.2	2.8%
white oak	<i>Quercus</i>	<i>alba</i>	795.8	2.8%
slash pine	<i>Pinus</i>	<i>elliottii</i>	563.3	2.0%
winged elm	<i>Ulmus</i>	<i>alata</i>	464.4	1.6%
Virginia pine	<i>Pinus</i>	<i>virginiana</i>	462.7	1.6%
All other species			6,572.3	22.9%
			28,711.3	

Table 4. Number of live trees greater than or equal to one inch DBH in Alabama and Mississippi 2012 (top 15 species)

Common name	Scientific name		Number of trees	Percent of total
loblolly pine	<i>Pinus</i>	<i>taeda</i>	6,487.4	21.5%
sweetgum	<i>Liquidambar</i>	<i>styraciflua</i>	4,319.1	14.3%
water oak	<i>Quercus</i>	<i>nigra</i>	2,234.5	7.4%
red maple	<i>Acer</i>	<i>rubrum</i>	1,978.2	6.5%
blackgum	<i>Nyssa</i>	<i>sylvatica</i>	994.7	3.3%
winged elm	<i>Ulmus</i>	<i>alata</i>	857.5	2.8%
yellow-poplar	<i>Liriodendron</i>	<i>tulipifera</i>	837.2	2.8%
flowering dogwood	<i>Cornus</i>	<i>florida</i>	730.6	2.4%
black cherry	<i>Prunus</i>	<i>serotina</i>	711.5	2.4%
American hornbeam, musclewood	<i>Carpinus</i>	<i>caroliniana</i>	682.7	2.3%
sweetbay	<i>Magnolia</i>	<i>virginiana</i>	606.4	2.0%
white oak	<i>Quercus</i>	<i>alba</i>	600.5	2.0%
mockernut hickory	<i>Carya</i>	<i>alba</i>	525.9	1.7%
southern red oak	<i>Quercus</i>	<i>falcata</i>	516.7	1.7%
green ash	<i>Fraxinus</i>	<i>pennsylvanica</i>	473.4	1.6%
All other species			7,683.1	26.8%
			30,184.0	

CHAPTER II
DEFINING THE IMPACTS OF PLANTATION FORESTRY ON
BIODIVERSITY IN ALABAMA AND MISSISSIPPI

Abstract

Global biodiversity has been affected substantially by anthropogenic sources, the main factors of which include land use change, habitat change due to as forest fragmentation and conversion, invasive species, overexploitation, global climate change, and pollution. A key component in the southern U.S. has been pine plantation forestry. This study assessed the primary causes of any changes in species composition in Alabama and Mississippi forests by first describing the recent historic trend (last 40 years) in species richness and diversity. Concurrently, the long-term impacts of plantation forestry on species diversity and richness were investigated. The results reveal that species richness and Shannon's diversity index increased over the time frame, while Simpson's diversity index declined. Southern pine plantation establishment was shown to be negatively correlated to Simpson's index as well as Shannon's index.

Introduction

Biodiversity, or biological diversity, is be defined as "the variety and variability among living organisms and the ecological complexes in which they occur" (OTA 1987). Humans perceive regions with a multitude of diverse species to have more value than those that do not (Wilson 1993; Ehrlich & Wilson 1991). Possible reasons that species diversity is valued by humans are that larger number of plant species means a greater variety of crops and life; greater species diversity helps assure natural sustainability for all life forms; diverse ecosystems can better withstand and recover from a variety of disasters; and the planet's complex systems, ecological networks, and energy flows depend on numerous organisms and interactions (Ehrlich & Wilson 1991, Gaston 1996).

A number of studies have linked species diversity to forest ecosystem function and productivity (Caspersen & Pacala 2001; Tilman et al. 1997). Species diversity is frequently used as a surrogate for forest stability, health, and productivity (McIntosh 1967; McNaughton 1977; Tilman 1996; Austin 2002). Species richness, or simply the number of species in an area, is currently the most popular measure of diversity (Stirling & Wisley 2001). Relative species abundance has been examined as well (Whittaker 1965; Hulbert 1971). Proportional indexes such as the Shannon-Wiener index (H') have been hypothesized to have a strong correlation as well (Hill 1973; Routledge 1980; Magurran 1988). Modern conservation planning is dependent upon the understanding a prediction of species distributions (Franklin 1993; Austin 2002, Austin & Cunningham 1981; Guisan & Zimmerman 2000; Elith & Burgman 2002).

The main factors responsible for biodiversity loss include land use change, habitat change such as forest fragmentation and conversion, invasive species overexploitation, climate change, and pollution (Hanski 2005). The impacts of

intensive southern yellow pine plantation forestry on biodiversity are a contentious subject. Linder (2004) explains how this type of management could negatively impact biodiversity of southern forests. The June 1999 public review of the assessment questions for the Southern Forest Resource Assessment (Wear and Gries 2002) contained multiple examples where individuals and organizations expressed concern that this type of silviculture could have negative impacts on biodiversity. Woolston (2013) details many of the tradeoffs involved in green energy and products including the possible loss of biodiversity. Woolston (2013) mentions how biodiversity loss due to plantation forestry is a key talking point for many environmental groups and non-governmental organizations (NGOs) such as the Dogwood Alliance and the Natural Resources Defense Council.

Hypothesis

Plantation management was hypothesized to affect the three measures of species diversity (species richness, Shannon-Weaver or Simpson's) considered in the study. Specific objectives evaluated in this analysis are:

1. Diversity indicators cluster in counties with large areas of pine plantations or increases in plantation area.
2. Diversity indicators cluster in areas with higher productivity rates, temperatures, potential evapotranspiration (PET), and precipitation.
3. Diversity is higher in counties with elevation ranges.
4. As plantations are established in fertile zones with higher temperatures, rainfall, and PET, collinearity between plantation establishment and environmental factors will increase.
5. Species diversity and Shannon-Weaver diversity index are positively correlated to plantation establishment, and Simpson's index, a measure of evenness, is negatively correlated to southern pine plantations.

Data, Diversity Indices and Spatial Unit

Tree and plot data

USDA Forest Service Forest Inventory and Analysis (FIA) data comprises the foundation of the dataset. This database stores inventory data from sample plots taken across the South. Information such as location, species, DBH, height, basal area, standing volume, and average annual growth, removals, and mortality are easily derived from this system. The 1998 Farm Bill requires that a systematic grid of plots is located across the county, and that a standard set of variables and measurements are performed on each plot. The FIA plot grid

assigns 1 plot to approximately every 6,000 acres. The Farm Bill further states that a systematic number of these plots be inventoried for each state every year (Bechtold & Patterson 2005). The southern states average about 6 years for each cycle. Thus, every 5-7 years, a complete set of plots has been re-inventoried for every state.

FIA estimates used include: average latitude and longitude of plots, percent forest/non-forest, acres of pine plantations, acres southern pine plantations expressed as a percentage of total forest area, total forest area, total land area, average site index, average all-live growth per acre, average all-live softwood growth per acre and average all-live hardwood growth per acre in each spatial unit. Additionally, species richness (or a count of species), Shannon's equitability (for evenness), and Simpsons diversity index were calculated for each spatial unit.

GIS Data Layers

Yearly average estimates of temperature and rainfall were derived from the Parameter Regressions on Independent Slopes Model (PRISM) raster data set published by the PRISM Climate Group at Oregon State University. 100-meter resolution elevation raster data were obtained from the U.S. Geological Survey. The average annual yearly potential evapotranspiration (PET) raster dataset from the Global Aridity and PET database was used to calculate average annual PET for each county. A GeoTIFF file of 2011, 30-meter national land cover data (NLCD) was used to calculate percent forest, urban, and agriculture for each county.

All raster data were aggregated to the HUC level using the zonal statistics tool in ArcGIS. Average estimates of temperature, rainfall, PET, and elevation were computed for each HUC. Additionally, the range of elevation was calculated as well. The percent of each county in the three main land uses, forest, urban, and agriculture were calculated in a similar fashion. Table 1 lists all FIA and GIS data used in this study.

Diversity Indices

This study investigated three measures of species diversity: species richness, Shannon-Wiener index, and Simpson's index. These indices were calculated with the "diversitycomp" command in the "BiodiversityR" package in R-Studio version 1.0.136. FIA tree and plot level data were used to create the data matrix of species counts by plot and watershed.

Species Richness

This study focused on two types of biological diversity measurements: richness and evenness. Richness is simply the number of different organisms or species living in an area (Wilson 1993; Ehrlich & Wilson 1991). It does not take into account proportion or distribution of each subspecies within the area of interest. All one must do is simply count the number of subspecies found in a community. It is the simplest and easiest diversity measurement to calculate and describe. Species richness is a key component in many the plant ecology community (Naeem et al 1996).

Relative species abundance within a community is often measured with an index ranging from 0 to 1 and equated species evenness (Hulbert 1971; Alatalo 1981). Evenness refers to how the subspecies are distributed throughout the whole community (Routledge 1980; Magurran 1988). It, in effect, is a normalization of richness. For example, there are two communities composed with five different subspecies. The first one has counts of 20 for each subspecies for a total of 100 individual subspecies counted. The second one has one subspecies with a count of 96, and the other four have one observation each. The second community also has a total of 100 individual subspecies counted. The first community would be considered more even, because each subspecies accounts for the same percentage of the total, while the second one is dominated by a single species.

Shannon-Wiener Index

Proportional abundance indices depend on both richness and relative abundance (Hill 1973). Shannon-Wiener (Shannon's for short) evenness and diversity index (H) can from information theory and measure the order and disorder within a population. This index is derived by calculating the proportion of species i relative to the total number of species (p_i), and then multiplying by the natural logarithm of this proportion ($\ln p_i$). The result is summed across species, and multiplied by -1:

Simpson's Index

Simpson's evenness and diversity index (D) was developed in 1949. Like Shannon's index, it also takes into account richness and diversity. Simpson's D can be found in three forms. The one used in this study is derived by taking the proportion of species i relative to the total number of species (p_i) and then squaring. The squared proportions for all the species are summed, and the reciprocal is taken. The formula for Simpson's index is:

An evenness or equitability index can be calculated for Simpson's D, just as one can be derived from Shannon's H. Simpson's equitability index (ED) is calculated by taking Simpson's index (D) and dividing it by the total number of

subspecies groups (S). Thus, the formula for ED is: $ED = D / S$, where S = total number of subspecies groups. Evenness or equitability takes a value between 0 and 1, with 1 being complete evenness. The variable names and descriptions are listed in table 5.

Methods

The first step was to analyze FIA data to determine the past and present status of the forest resources in the study area. This included forest area, species composition, land-use, and plantation establishment. This was performed by interrogating the data in an ORACLE database with SQL. Next, exploratory data analysis was performed using the SAS version 9.3. The procedure PROC means was applied to the diversity variables to get a general feel for the data. Correlation coefficients and matrices were computed to ascertain relationships and patterns.

The third step required searching for spatial patterns with ArcGIS10.3. A series of choropleth maps of both dependent and independent variables were created and analyzed. This line of interrogation was followed by the creation of Hot-Spot maps. The maps and the correlation analysis were used to gain an understanding of the dataset and underlying relationships. This knowledge is used in the next step, model development.

Model Development

Ordinary Least Squares and Spatial Regression (OLS)

Ordinary least squares (OLS) analysis is used to model relationships. The goal of OLS is to define the relationship between a dependent variable (y) and a set of explanatory variables (x). The relationship is generally linear in nature. The formula for OLS is: $Y = X\beta + e$. Ordinary least squares is a global model in that it produces one equation for the entire dataset

The OLS process accomplishes two goals. The first is to create a properly fitted model that fits predicted values of $\chi\beta$ onto observed values of the explanatory variable y. Discovering how the explanatory variables interact and contribute to the linear relationship of the dependent variable is the second goal (Matthews, 2006). This is accomplished by minimizing the sum of the squared prediction errors (least squares).

The best OLS model is often referred to as the best linear, unbiased estimator (BLUE). BLUE requires certain assumptions and inferences be made about the population and sample. These assumptions are that the observations are normally distributed, independent and unbiased. If the random error terms

have a mean of zero, then there is no systematic bias in the BLUE equation. Additionally, BLUE has random error terms that are uncorrelated and have a constant variance, i.e. they are homoscedastic (Matthews 2006; Ward & Gleitsch 2007).

Ecological processes and datasets often violate the BLUE assumptions. Often times the observations influence each other or are even related. The data is often skewed or bi-modal. One purpose of regression models is to interpret the significance and magnitude of the coefficients of prediction variables. It is important in analyzing spatial data to determine if the relationships revealed are random or linked together by underlying forces. Analysis conducted on spatial data may be incorrect if the variables show have significant spatial autocorrelation (SA) and steps were not made to correct for this SA. The best method to determine if SA is an issue is the Global Moran's I described previously. If Global Moran's I indicates that SA is an issue, then OLS may not be the best modeling process (Matthews 2006; Ward & Gleitsch 2007).

Spatial dependence can be considered either a nuisance or substance. If the goal is to derive proper statistical inference, estimation, hypothesis testing, and prediction, then the presence of SA is a nuisance. This is because the parameters estimated may be biased and incorrect. Including a spatially dependent variable in the regression error term can alleviate some of this nuisance. However, if the aim of the study is to discover if interactions and spatial spillover does occur, then the SA can be considered substantive. The solution for this is to incorporate the structure of the dependence into the model which will allow for correct estimation and interpretation (Anselin 1988, 1990, 1995; Matthews 2006; Ward & Gleditsch 2007; ESRI 2017). A third option for addressing SA is the use of dummy variables representing areas of known similarities in the OLS regression.

Dummy Variables

Scientists of all disciplines have recognized that there is often a large amount of heterogeneity within the study areas. Countries of similar economic wealth or education exist in similar regions. Productive and unproductive forests occur alongside each other. Regions of radioactivity or toxic waste are within a certain distance of a contaminant site. One way to address these issues is to include a dummy variable in the OLS process. These variables represent different regions across the geographic landscape that the researcher believes have similar underlying processes or relationships with each other. The number of dummy variables used is one less than the total number of regions.

Knowledge of which regions or areas are defined by a dummy variable is critical for model interpretation. For example, if a simple one variable model is created for six regions, the OLS would produce a beta for the intercept and the

variable in question. These two represent Bo and B1 for the unspecified region. All other coefficient estimates indicate the predicted difference of the value in question to the un-specified region.

Assume that a study is investigating the impacts of precipitation on plant growth. These are the only variables in the dataset. However, the scientist believes that there are three ecoregions in the study area and that they have an impact on plant growth. These three ecoregions are the southeastern mixed forest (SEMF), the outer gulf coastal plain (GCP), and eastern broadleaf forest. Dummy variables for southeastern mixed and gulf coastal plain are created. The resulting OLS model would have an intercept, and betas for growth, SEMF, and GCP. The intercept and growth coefficient are the model for broadleaved forest. The coefficient for SEMF and GCP represent the growth of the regions above or below the estimate for the broadleaf forest (Matthews 2006; Schmidhammer 2009).

Spatial Lag Model (SLM)

The spatial lag model (SLM) incorporates the influence of not only the independent variable, but the effect that neighboring attribute values by incorporating a “spatially lagged” variable of y on the right hand side of the regression equation (Anselin, 1988, 1990, 1995). The formula for the SLM model is: $Y = \rho W_y + X\beta + e$, where W_y is the weighted average of neighboring y 's, i.e. the spatially lagged variable and ρ is the spatial autoregressive coefficient. The SLM model primarily controls spatial autocorrelation in the dependent variable. The SLM treats spatial dependence as substance rather than a nuisance. If the SLM model is properly specified, there should be little or no spatial dependence remaining in the residuals (Anselin 1988, 1990, 1995).

The inclusion of the lagged y variable is appropriate in cases where the values of y interact and influence each other. Examples of possible dependence among observations include; the presence of a species is more likely found in a neighboring site due to seed fall or coppice; an individual home value is likely impacted by the value of neighboring houses; a country is more likely to have similar wealth and levels of education to its neighbors due to interactions between the two. The SLM incorporates the influence of unmeasured independent variables plus the effects of neighboring attribute values. This model is often referred to as the simultaneous spatial autoregression model (Anselin 1988; Ward & Gleditsch 2007; Matthews 2009).

OLS produces biased coefficients in cases where spatial lag is the correct model. Therefore, the reason for using a SLM model over OLS is to derive the proper inference on the coefficients of the other covariates in the model. This is because as one y is based on a neighboring y , which is then in turn based on its

own value of y , creating a feedback or dependency that is ignored by OLS. The feedback will increase as the size of the data frame increases, and the resulting OLS estimates will not converge on the true values as the number of observations increases. This dependency means that $\hat{\beta}$ no longer measures the treatment effect of x on y , but is probably biased downward. The lagged model coefficients assess the effect of $\hat{\beta}$ on the observation while controlling for spatial dependence among y . The spatially lagged variable represents that amount of y that can be attributed the value of y in neighboring observations (Anselin 1988; Ward & Gleditsch 2007; Matthews 2009).

Spatial Error Model (SEM)

The spatial error model (SEM) captures the influence of unmeasured independent variables upon each other. Spatial error models describe the extent to which the spatial clustering of a variable is not explained by the measures of independent variables but accounted for the clustering of errors. The formula for SEM is: $Y = X\beta + \lambda Wu + e$, where Wu is the spatial lag of the error terms and λ is the autoregressive coefficient and indicates the extent to which the spatial component of the errors is correlated with nearby observations (Anselin 1988, 1990, 1995; Ward & Gleditsch 2007).

If there is no SA between the errors for neighboring observations, then λ will be 0. If $\lambda \neq 0$ then there is SA between the error terms and the connected observations. There are consequences of failing to address SA between the errors and performing on OLS instead. The OLS estimates would actually be unbiased, but the standard errors of the coefficients would be wrong. This due to the fact that the OLS model depends on an estimate of variance than assumes independent observations. If this assumption is violated, then the OLS estimate of variance will tend to underestimate the actual variance. Additionally, the estimated coefficients may not be efficient or “close” to the true values (Anselin 1988, 1990, 1995; Ward & Gleditsch 2007; Matthews 2009).

Model Comparison and Goodness of Fit

A suite of regression modeling techniques is available for ecological and spatial analysis. These include OLS, OLS with dummy variables, SLM, and SEM. What are the advantages and disadvantages of each, and when should one be preferred of the others? If spatial dependence or autocorrelation are not present, then OLS is the preferred model. In instances where the researcher recognizes a few areas of clustering and can identify and discuss the forces for this clustering and is not interested in investigating these reasons, then OLS with dummy variables is a valid choice.

If SA is an issue, it must be addressed and OLS removed as a modeling option. This due to the fact that one of the key uses of modeling is for specification testing. The presence of SA will produce misspecified models. Of the two spatial regression models, SLM and SEM, the failure to account for spatially lagged dependence is the most serious, as it is similar to an omitting a key variable from the model. The end result would be biased and inconsistent estimates for all coefficients and lead to incorrect inference (Matthews 2006). However, both are serious issues and a procedure for choosing between the two is needed.

One method of choosing which spatial model to use depends upon the researcher's understanding of the area, topic of interest, and the data. If observations influence each other, the SLM model may be the best choice. This is a form of nuisance dependence and can be seen in studies involving home values or species migration. Spatial error model may be best in studies where the values of the dependent variable are not influenced by each other, but by connecting factors. This would be made evident by a correlation of the error terms.

Baller et al. (2001) illustrated how the presence or absence of SA and the types of SA impact the estimation of dependent variables. They go on to state that the lag model primarily controls SA in the dependent variable, while the error model controls SA in the both residuals. As a result, the error model controls autocorrelation in both the dependent and independent variables. This makes the SEM a more robust and generally better model (Ward and Gleditsch 2007 and Baller et al. 2008).

The criterion for deciding the best fit model for OLS is r-square (R^2) or adjusted R^2 . Sometimes referred to as the coefficient of determination, R^2 measures how close the data are to the regression line. It is a measure of how much variation in the dataset is explained by the model. However, neither of these two is appropriate in comparing different spatial regression models. The Akaike Information Criteria (AIC) is used to measure the model with the best fit. Unlike R^2 , where the larger the estimate the better the model, the smaller the AIC value the better the model. The formula for AIC is: $AIC = 2k + n [\ln(\text{Residual Sum of Squares})]$, where k is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept term. AIC can only be used to compare models with the same dependent variable (Akaike 1974).

Choosing the best regression model

Choosing the appropriate model is easy when either the reasons for clustering are known or spatial dependence is absent from the data. However, in ecological studies, these scenarios are rare. In most studies, the reasons for clustering are only partially known and captured by the suite of variables. Additionally, there is autoregression between both the dependent variables and error terms, so neither the SLM or SEM are obvious solutions. This study employed the model decision tree developed by Luc Anselin for use with the spatial analysis software package GeoDa that he developed (Anselin 2005).

The GeoDa decision process begins by performing an Ordinary Least Squares (OLS) analysis on the data. Spatial autocorrelation diagnostics based on Moran's I is included in this process. If this measure is insignificant, then the process can end and the results of the OLS procedure are used. However, if spatial dependence is found to be a problem, then a maximum likelihood regression is performed on either the spatially lagged variable or the spatial error terms. The decision on which model (SLM or SEM) depends on the results of OLS diagnostics (Anselin 2005).

The test diagnostics for determining the best spatial model are based on Monte Carlo simulation experiments that use Lagrange multipliers (LM) of spatial lag and error statistics. The spatial LM-lag and LM-error specifications are related, so that tests against one form of dependence will also have power against the other. The basic approach used in the GeoDa software developed by Anselin uses LM tests that are based on the residuals of OLS regression. Separate tests for LM-lag and LM-error are produced. If the LM-error diagnostics are significant, then a SEM procedure is performed. Likewise, if LM-lag diagnostics are significant, then a SLM procedure is carried out. If both LM-lag and LM-error statistics are significant and reject the null-hypothesis then a more Robust LM-lag and LM-error statistics are produced. Like the non-robust test procedure, the statistic that is significant or most significant will indicate that the matching model should be used in the analysis (Anselin 1998, 2005; Mathews 2007).

All spatial analysis in this study was performed using either ArcGIS version 10.3 or GeoDA version 1.8.14.8. ArcGIS was used for detecting global spatial autocorrelation (Moran's I), hot-spot analysis, and producing general maps. All OLS and spatial regression was performed using GeoDA and the model decision tree developed by Anselin in 2005 was followed. AIC was used to determine the best fit model.

Results

Plantation establishment

Table 1 illustrates the increase in pine plantation acreage for the study area. Planted stands accounted for less than 10 percent of the total forest acreage for the states during the 1970s. Today, plantations account for more than 31 percent of the forested area.

Species composition for the study area changed as well. Table 2 details the most abundant species present in Alabama in 1972 and Mississippi in 1977. Table 3 reveals the same results for 2013. While the top two species remain the same, the percentage of loblolly pine has almost doubled. In fact, the latest inventory numbers indicate that one of every four trees in Alabama was a loblolly pine in 2013. These data illustrate why many are concerned that plantation forestry poses a potential threat to species diversity in Southern forests.

Exploratory data analysis

Average species richness increased over the time period examined, while Simpson's index, a measure of evenness, decreased and Shannon's diversity index remained fairly stable. Species richness went from almost 41 to more than 44, indicating that on average, each county gained three additional tree species. This is almost a 10 percent gain in a 40-year time span (table 6).

Correlation coefficients between the three dependent variables and the rest of the dataset were produced in SAS (tables 7-9). The variables with the greatest positive correlation on species diversity for the 1970s surveys pertain to changes in forest and plantation area, potential evapotranspiration (PET) and measures of average annual growth. Higher PET often indicates more productive areas, and therefore greater diversity. Variables negatively correlated to the 1970 diversity measures were plantation area, percent of the county in urban development, and longitude. Interestingly, temperature was negatively correlated to Shannon-Weaver and Simpson's index and precipitation negatively correlated to all three 1970 independent variables. This may be due to plantation establishment being higher in these areas and an indicator of multicollinearity. Finally, the impacts of longitude suggest that richness was higher in the eastern portion of the study area.

Correlation analysis of the 2013 data revealed similar patterns as found in the 1970s dataset, with a few additional findings. Variables to plantation area, changes in plantation area and average annual softwood growth were positively related to species richness, and negatively correlated with the other two diversity measures (table 7).

Correlation analysis of percent change in species richness, Shannon-Weaver, and Simpson's reveals that plantation area in 2013 and changes in plantation area between the two time periods were negatively related to Shannon-Weaver and Simpson's. Temperature and precipitation were negatively related with change in species richness and positively related with change in the other two dependent variables. Loss of forest area leads to lower diversity measures. Finally, percent of the county in urban development was negatively correlated to the two indices (table 8).

Spatial patterns & clustering

ArcGIS version 10.1 was used to create a series of maps for the initial stages of exploratory data analysis (cite ESRI and version number here). Visual interpretation of these maps could lead to possible relationships existing in the dataset. Figures 3 through 5 reveal the spatial density patterns for species richness, Shannon's index, and Simpson's index, respectively. Interestingly, the northeast corner of the study contained high levels for all three indicators. The greatest concentration of high species richness was in the southcentral portion of the study area. Conversely, Simpson's index was lowest in this area. Shannon's index generally increased from southwest to northeast, with a few exceptions along the western edge. The 2013 plantation acreage (fig. 6) was highest in the same southcentral area where richness was highest and Simpson's was lowest. Total forestland area per county 2013 (fig. 7) exhibited similar patterns, but was not as strong as plantation acreage. Percent of county in urban land uses was the mirror image of plantation acreage, indicating that plantations were established away from urban development (fig 8).

Maps depicting the percent change from the 1970s to 2013 in species richness (fig. 10), Shannon's index (fig. 11), and Simpson's index (fig. 12) do not mirror the 2013 maps, but contain similar patterns where species richness was high and Simpson's low in southcentral Alabama and change in Shannon's index was highest in the northeast. The map illustrating percent change in plantation acreage from the 1970s to 2013 (fig. 13) reveals high clusters in southcentral Alabama and western Mississippi, and low clusters around Mobile and the northern portion of the study area.

From viewing the maps, possible positive correlations exist between change in plantation and increases in species diversity. Areas with high urban development appear to be correlated to Simpson's Index as well. Possible negative correlations appear with changes in plantation area and Simpson's index and urban land uses and species richness.

Spatial autocorrelation

ArcGIS was used to compute Moran's I for multiple variables in the dataset. In every instance, the results indicated that the data were clustered and produced similar results (fig. 14). A Moran's index of 0.45897 and a z-score of 7.48 reveal clustering within the dataset that is more pronounced than a random distribution. A high positive z-score indicates that the surrounding features have similar values, either high-high or low-low. Therefore, models developed from this dataset should be tested for spatial autocorrelation and any spatial dependence found should be addressed.

Change in species richness

The best change in species richness model produced was a spatial error model that accounted for 43 percent of the variation found within the dataset (table 10). The model contains three independent variables along with the spatial error variable lambda. These variables are range in elevation, precipitation, and percent of county land cover in agricultural uses. Range in elevation and percent agriculture were positively correlated to the dependent variable, change in species richness, over time. Precipitation was not, possibly due to correlation between the location of plantations and agricultural lands and precipitation.

Change in Shannon's diversity index

The diagnostics for the ordinary least squares model on change in Shannon's index model revealed that spatial dependence, heteroscedasticity, and multicollinearity were not significant and the results from the OLS model are valid (table 11). The model accounts for more than 43 percent of the variation in the dataset and includes four independent variables. Three of these variables, average all-live growth per acre, potential evapotranspiration (PET), and change in plantation acreage were negatively correlated to Shannon's index. The fourth variable, percent of the county in urban land uses in 2013, was positively correlated.

Change in Simpson's diversity index

The best Simpson's index change model contained a weighted variable from a spatial lag model with three independent variables and accounted for 70 percent of the variation found within the dataset (table 12). The variable difference in plantation acreage between the two time periods was used instead of percent change in plantation acreage because differences in plantation acreage alone had an R^2 of 0.41 in OLS and 0.67 in a spatial lag model. Additionally, while percent change in plantation area was always highly

significant in the OLS model, it became insignificant in all other spatial models. One reason is that change in plantation area was correlated to environmental variables such as precipitation and productivity, which also produced valid models. For example, a three-variable model of Simpson's index using latitude, PET, and average all-live growth per acre had R^2 s of 0.60 and 0.66 for OLS and spatial lag models, respectively. All three variables were negatively correlated to Simpson's diversity index. The reason for this is that pine plantations are located in fertile areas, managed to promote growth and are concentrated in the southern portion of the study area. This makes all three correlated to plantations and plantation establishment.

Discussion

Correlation analysis

Changes in species richness were positively correlated with variables related to elevation, precipitation, and longitude. Counties with higher elevations or range in elevation may contain more species, as certain species may only be found at certain elevations. Higher rainfall often led to a greater number of species. The impacts of longitude suggest that richness was higher in the eastern portion of the study area. Variables pertaining to plantation area or concentration were positively correlated to species richness. This may seem counter intuitive at first glance, but greater forest management can have the same effect of disturbance and creates an increase in forest edge, both of which could lead to a higher species diversity.

Changes in Shannon's and Simpson's indices were positively correlated to precipitation and percent of county in urban land uses. The reason for a positive correlation to urban land use are two-fold. First, urbanization often brings about the introduction of many exotic species. Second, as urban area increases, timber harvest decreases.

Negative correlations to Shannon's and Simpson's diversity include differences in forest and plantation area, pet, average softwood growth per acre, and plantation area in 2013. Plantations are often established in the most productive areas and also may produce more softwoods per acre than natural stands. Therefore, it is possible that there is some co-linearity existing between these variables.

Change in species richness

Two of the independent variables were related to the amount of forest area or percent forest area within each county. The simple explanation for this is

that the greater the forest area, the more likely one will encounter additional species. More species tended to exist in the eastern portion of the study area. Species diversity was also greater in counties with higher average hardwood growth per acre estimates. These counties may have more hardwood species which would tend to naturally have a greater diversity of species. Finally, counties with higher percentages of agriculture tended to contain more species. There are several possible explanations for this. Areas with agriculture may have more forest-nonforest edges and disturbance, which promote the establishment of different species. Secondly, introduced and exotic tree species may be introduced in areas with intensive agriculture.

Interestingly, the elevation variables did not have as great an impact as the correlation analysis indicated. This is possibly due to collinearity with other variables, hardwood growth and longitude in particular.

Change in Shannon's diversity Index

The negative correlation between change Shannon's index and change in plantation acreage was expected and discussed earlier. This suggests that Shannon's index declines in counties with increasing plantation acreage. Conversely, Shannon's index will increase in counties with increasing urban development. This seems logical, as southern pine plantations are typically dominated by one or two species, and urban development creates edges and disturbance in counties which would promote a variety of tree species. Additionally, urban development is often associated with the introduction of exotic tree species, which would have a positive effect on diversity indicators.

The reasons that PET and all-live growth were negatively correlated are unclear. One possible explanation would be that plantations are often established in more productive areas, that is areas with high PET and growth rates. This would often lead to spatial dependence within the model, however that is not the case in this instance.

Change in Simpson's diversity index

Both forest area in 2013 and percent of urban area in each county were positively related to Simpson's diversity index. This is to be expected, as one would expect all diversity indexes to increase with forest area and urban development is a form of disturbance and often includes the introduction of non-native species.

Change in forest area between the two time periods was negatively correlated to Simpson's index. As was hypothesized earlier, this is due to pine plantations being dominated by a few species and Simpson's index is a measure

of evenness. Thus, as plantation establishment increases in counties, species evenness will decrease.

Conclusions

Given that the issues of human development and climate change and their effects on forested ecosystems is both contentious and current, solid research into these areas is needed. This study suggests that both have an impact on tree species diversity. Plantations do have an impact on species diversity in terms of richness and evenness (Simpson's). Plantations seem to increase the number of species. Perhaps the edge effect from management increases the number of species. This is combined with the fact that many of these managed forests occur in rural areas where forest acreage is large, and area of forests is the greatest predictor of species richness.

Urban development and agriculture have an impact as well. Both have the potential to increase species richness and other indicators of diversity by promoting patches of disturbance and introducing new species. However, this study is based on the number of species only and does not track changes in species composition. Species that are associated with plantation establishment, urban development, and agricultural activities tend to be early successional and often non-native invasive (Laurance, 2010). These species are often established at the expense of native species and replace multiple species with a few species favored for management.

Increasing plantations has a negative impact on both Shannon's diversity index and Simpson's index. Simpson's is a measure of evenness, indicating that the species are being dominated by a few species, as shown in Table 4. So, while overall species numbers have yet to be impacted by plantation forestry, it has affected the distribution of species negatively.

Climate change has the potential to alter species composition as well. As the data indicate, warmer and wetter climates may actually increase species richness and create competition, increasing species diversity. However, these same climatic scenarios may also bring about increases in plantation forestry and agricultural establishment, which are negatively correlated to diversity. Additionally, dryer and or colder climates could lead to species loss and lower levels of tree species diversity.

Additional research into these associations is needed. One factor that needs resolving is determining the best spatial scale for the sample units. Are there ecological units such as watersheds or ecoregions that may be better suited for this type of analysis? What other sources of data can be incorporated into the models. And finally, can a model be developed that includes environmental and anthropogenic factors that allows for prediction scenarios?

References

- Alatalo, R. V. 1981. Problems in the measurement of evenness in ecology. *Oikos* 37:199–204.
- Anselin, L. 2005. Exploring Spatial Data with GeoDa: A workbook. Center for Spatially Integrated Social Science. Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana-Champaign. <http://geodacenter.asu.edu/system/files/geodaworkbook.pdf>. 244p.
- Austin M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Model.* 157:101–18.
- Austin, M.P., Cunningham, R.B., 1981. Observational analysis of environmental gradients. *Proc. Ecol. Soc. Aust.* 11: 109-119.
- Bechtold, W.A., Patterson, P.L. Editors. 2005. The enhanced Forest Inventory and Analysis Program – national sampling design and estimation procedures. Gen. Tech. rep. SRS 80. Ashville, NC: USDA Forest Service, Southern Research Station. 85p.
- Caspersen, J.P. Pacala, S.W. 2001. Successional diversity and forest ecosystem function. *Ecological Research* 116: 895-903.
- Ehrlich, Paul R. and Edward O. Wilson. 1991. Biodiversity studies: science and policy. *Science* 253.5021: 758p.
- Elith, J. and Burgman, M. A. 2002. Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. – In: Scott, J. M. et al (eds), Predicting species occurrences: issues of accuracy and scale. Island Press: 303–314.
- ESRI. 2017. Online help for ArcGIS desktop. <http://resources.arcgis.com/en/help/>.
- Franklin, J.T. 1993. Preserving biodiversity: species, ecosystems or landscapes? *Ecological Applications* 3(2); 202-205.
- Gaston, Kevin J. 1996. Biodiversity: A Biology of Numbers and Difference. Blackwell Science Ltd. 396p.
- Getis, A., and J. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24, 189-206.

- Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135, 147-186.
- Hanski, Ilkka. 2005. Landscape fragmentation, biodiversity loss and the societal response. EMBO reports VOL 6 - NO 5 – 2005.
- Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–432.
- Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52: 577–586.
- Laurance, W. F. 2010. Habitat destruction: Death by a thousand cuts. Conservation Biology for All. N. Sodhi & P. R. Ehrich (editors). Oxford University Press, Oxford. Pgs. 73-87.
- Linder, E.T. 2004. Chapter 25. Biodiversity and Southern Forests. In- Southern forest science: past, present, and future. Gen. Tech. Rep. SRS–75. Rausdher, H.M & Johnsen, K. eds. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 394 p
- Magurran, A.E. 1988. Ecological Diversity and Its Measurements. Chapman & Hall, London, 179 p.
- McIntosh, R. I. 1967. An index of diversity and the relation of certain concepts to diversity. *Ecology* 48:392–404.
- McNaughton, S. J. 1977. Diversity and stability of ecological communities: a comment on the role of empiricism in ecology. *American Naturalist* 111:515–525.
- Naeem, Shahid, et al. 1996. Biodiversity and Plant Productivity in a Model Assemblage of Plant Species. *Oikos* vol. 76, no. 2: 259–264. JSTOR, www.jstor.org/stable/3546198.
- OTA (US Congress Office of Technology Assessment). 1987. Technologies to Maintain Biological Diversity. US GPO, Washington, DC.
- Routledge, R. D. 1980. Bias in estimating the diversity of large, uncensused communities. *Ecology* 61:276–281.
- Stirling, G. and Wilsey, B. 2001. Empirical relationships between species richness, evenness, and proportional diversity. *The American Naturalist* 158:286-299.

- Tilman D., Knops J., Wedin D., Reich P., Ritchie M. & Sieman E. 1997. The influence of functional diversity and composition on ecosystem processes. *Science* 227: 1300–1302.
- Wear, David N.; Greis, John G., eds. 2002. Southern forest resource assessment. Gen. Tech. \ Rep. SRS-53. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 635 p.
- Whittaker, R. H. 1965. Dominance and diversity in land plant communities. *Science* 147: 250-260.
- Wilson, Edward O. 1993. The Diversity of Life. W. W. Norton & Company. 448p.
- Woolston, C. 2013. Biomass Boom: Threat or opportunity for Southern Forests. *Bioenergy Connection* v2.3.

Appendix

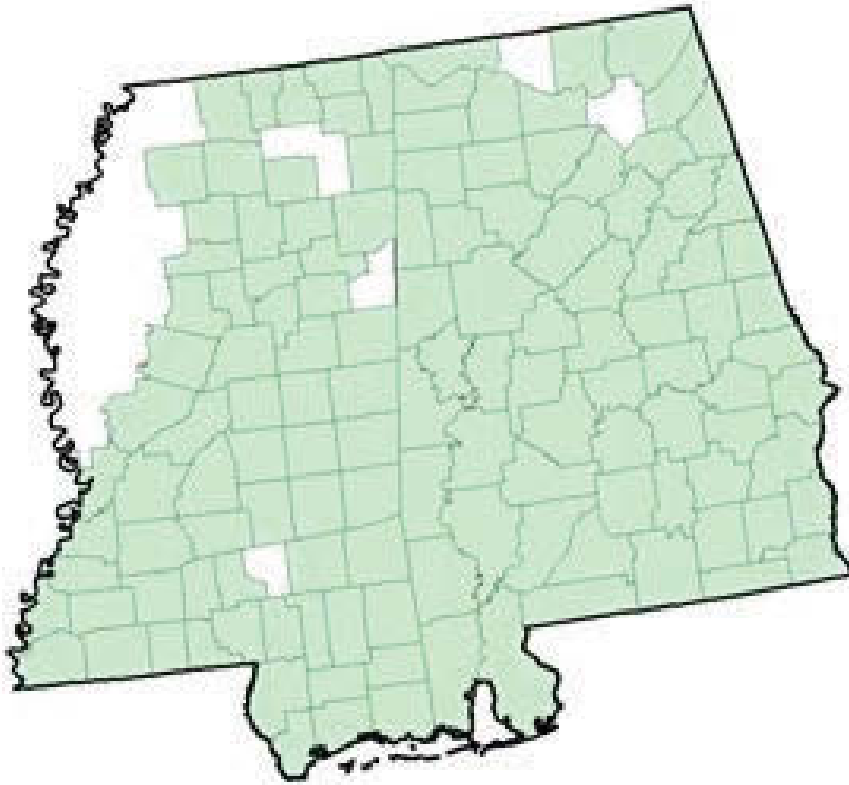


Figure 2. Study area

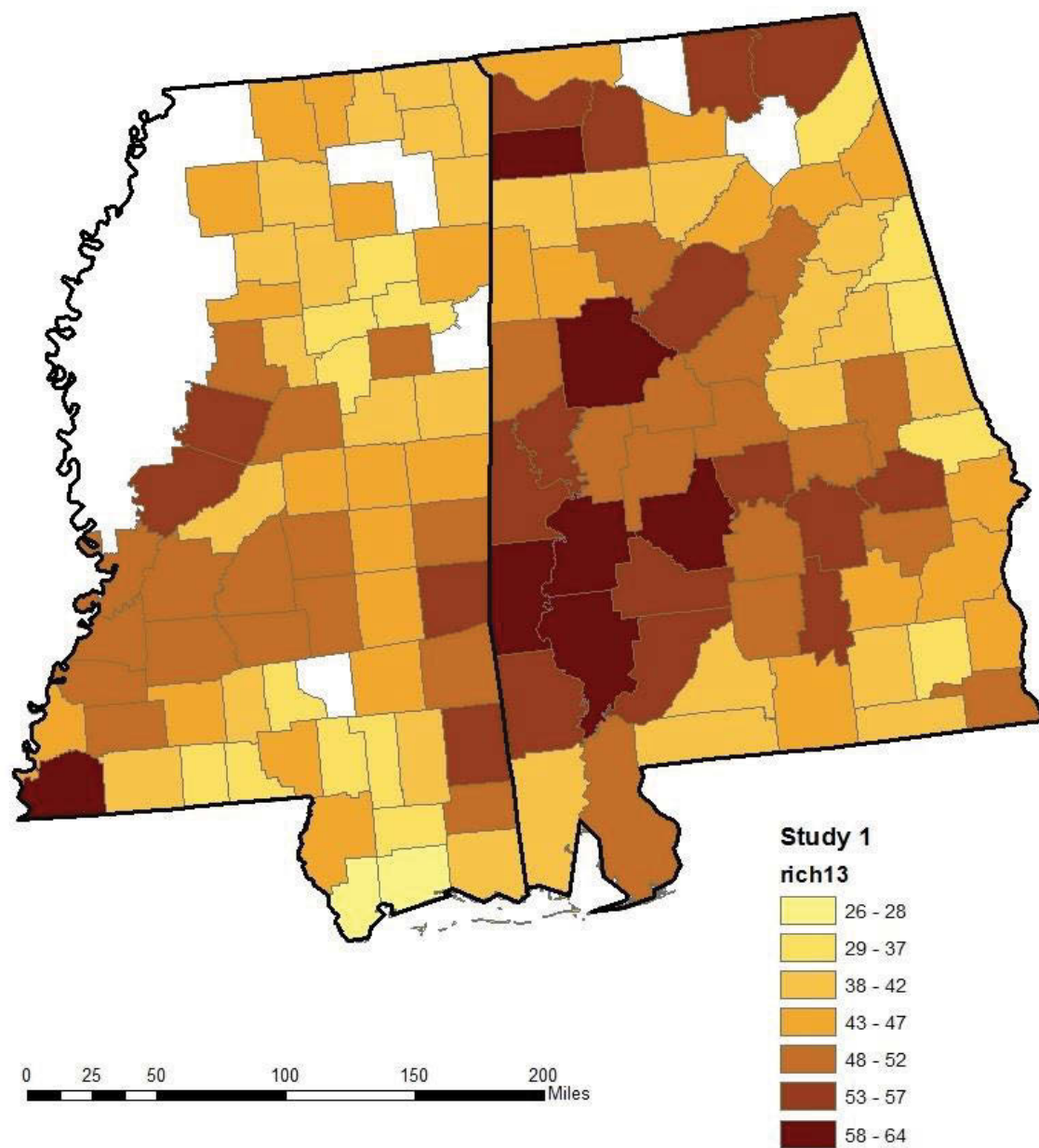


Figure 3. Species richness (number of tree species) by county, 2013

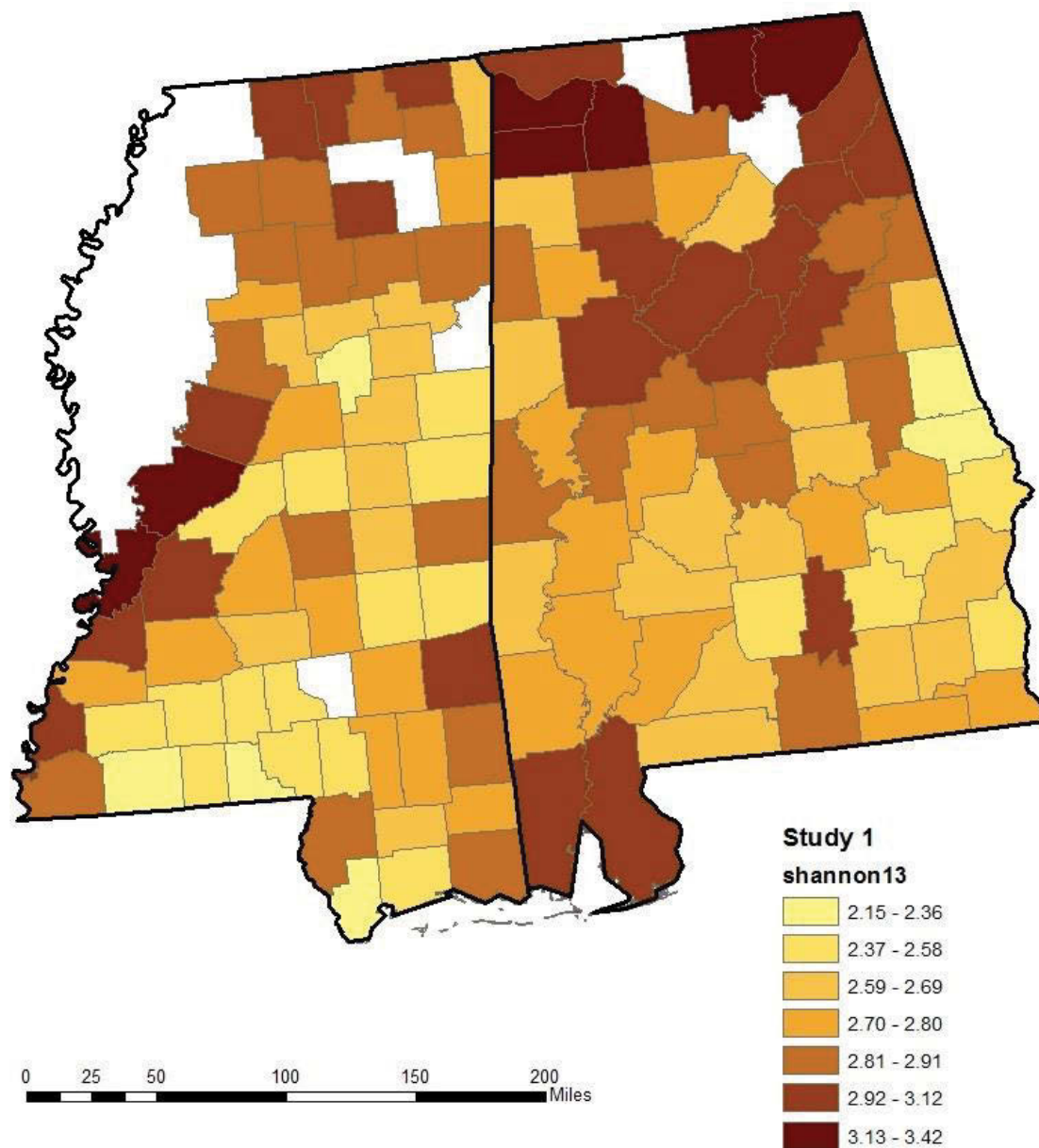


Figure 4. Shannon-Weaver diversity index by county, 2013

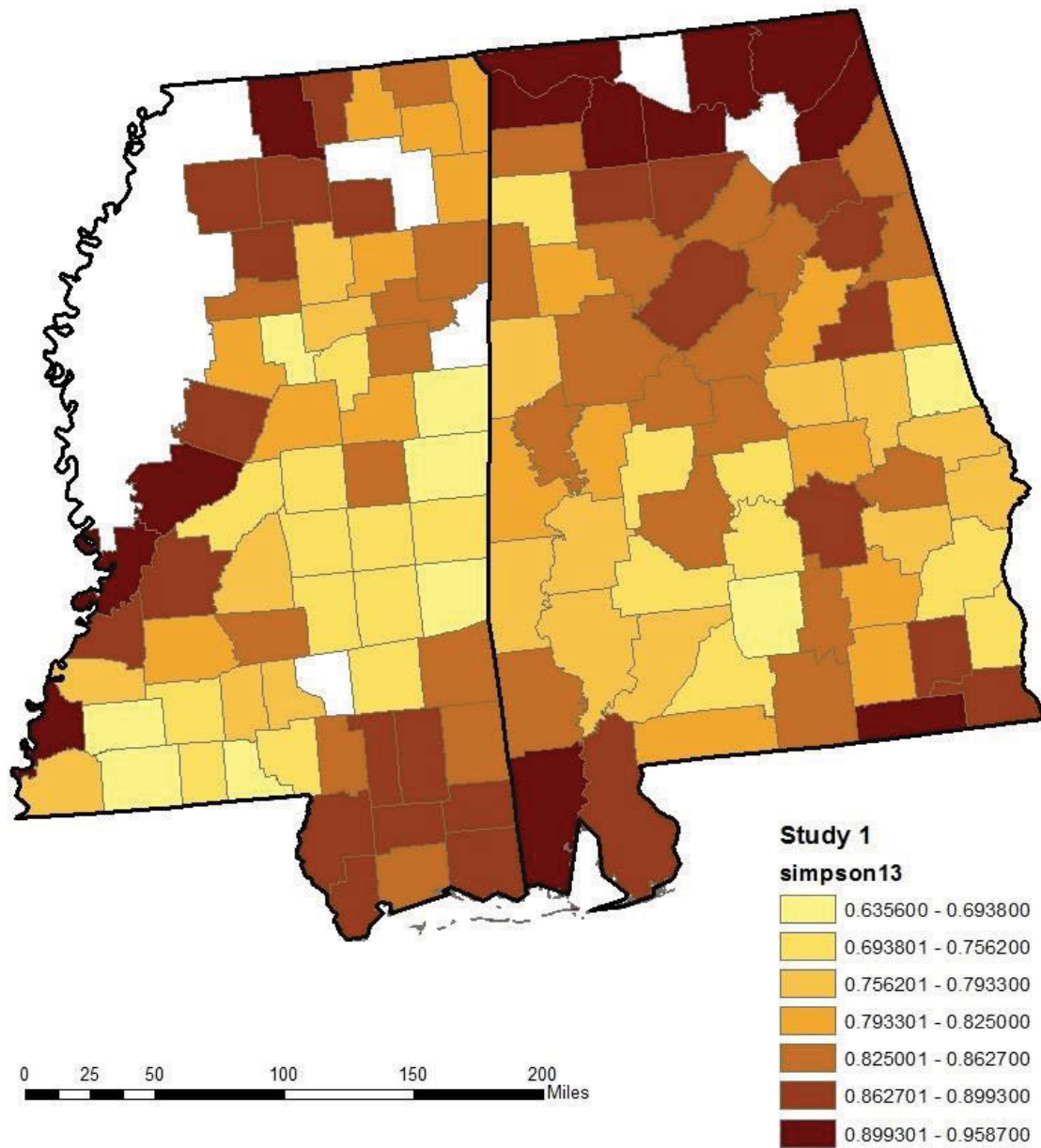


Figure 5. Simpson's diversity index by county, 2013

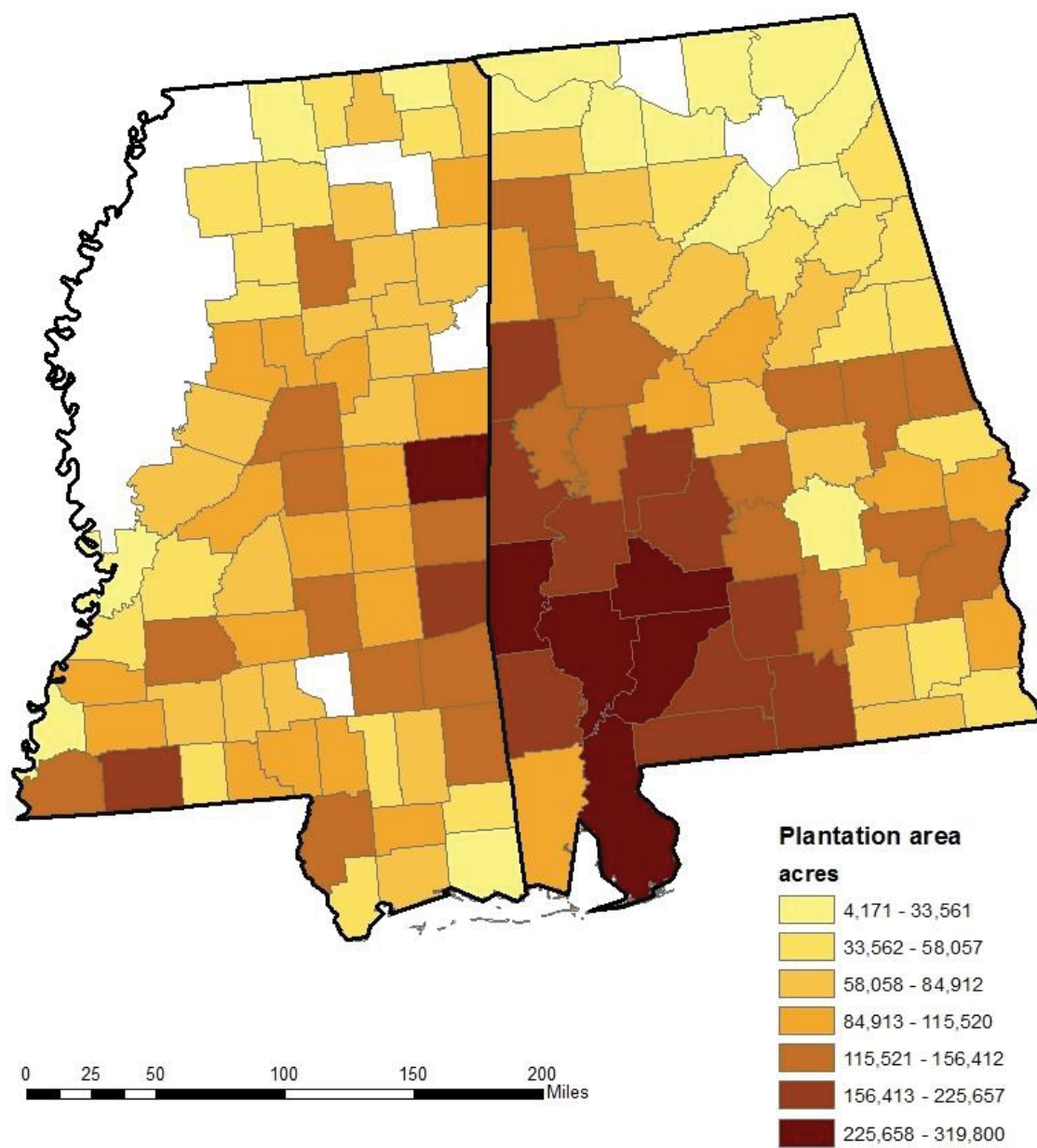


Figure 6. Plantation acreage, in acres, by county, 2013

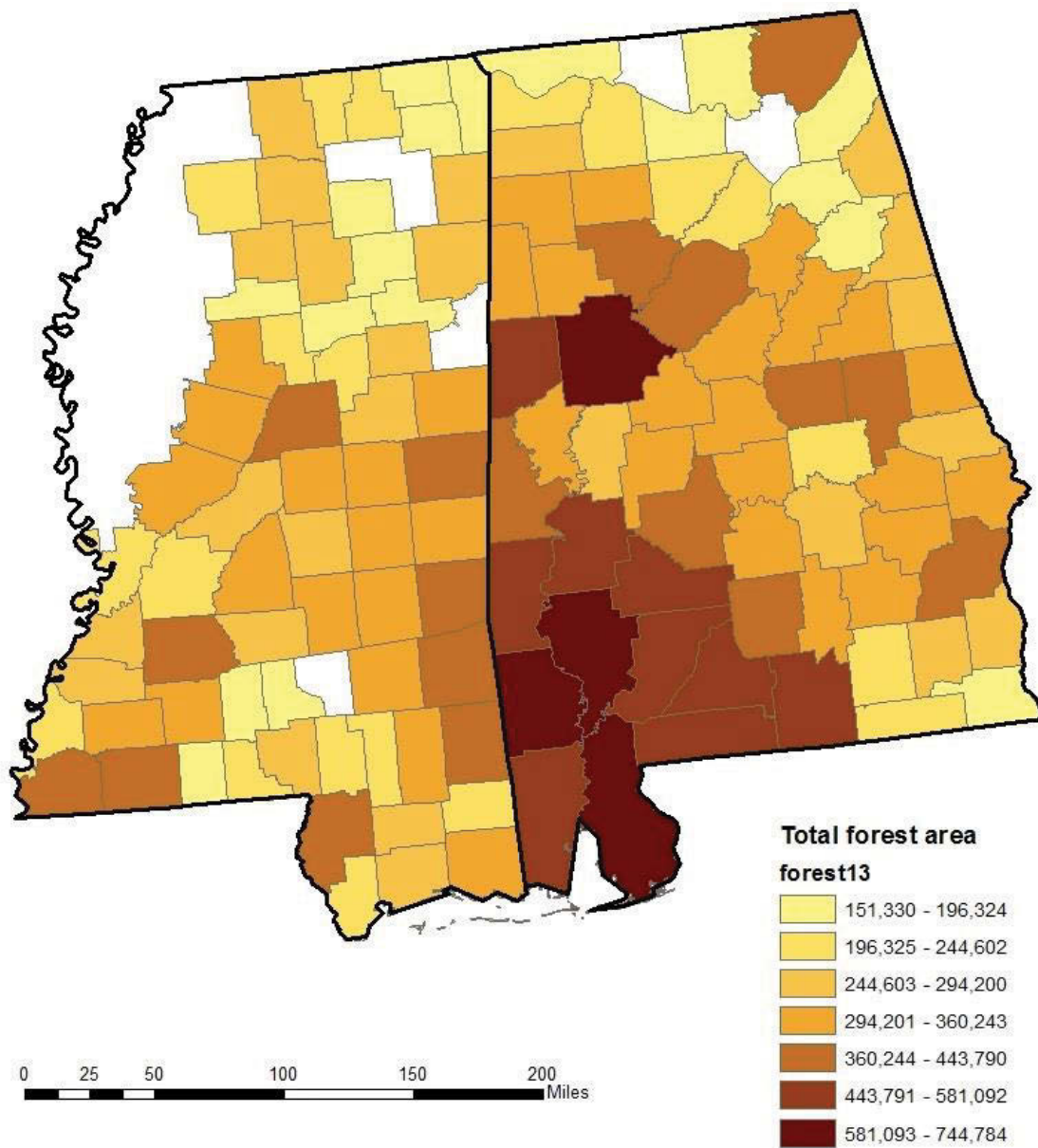


Figure 7. Forestland area, in acres, by county, 2013

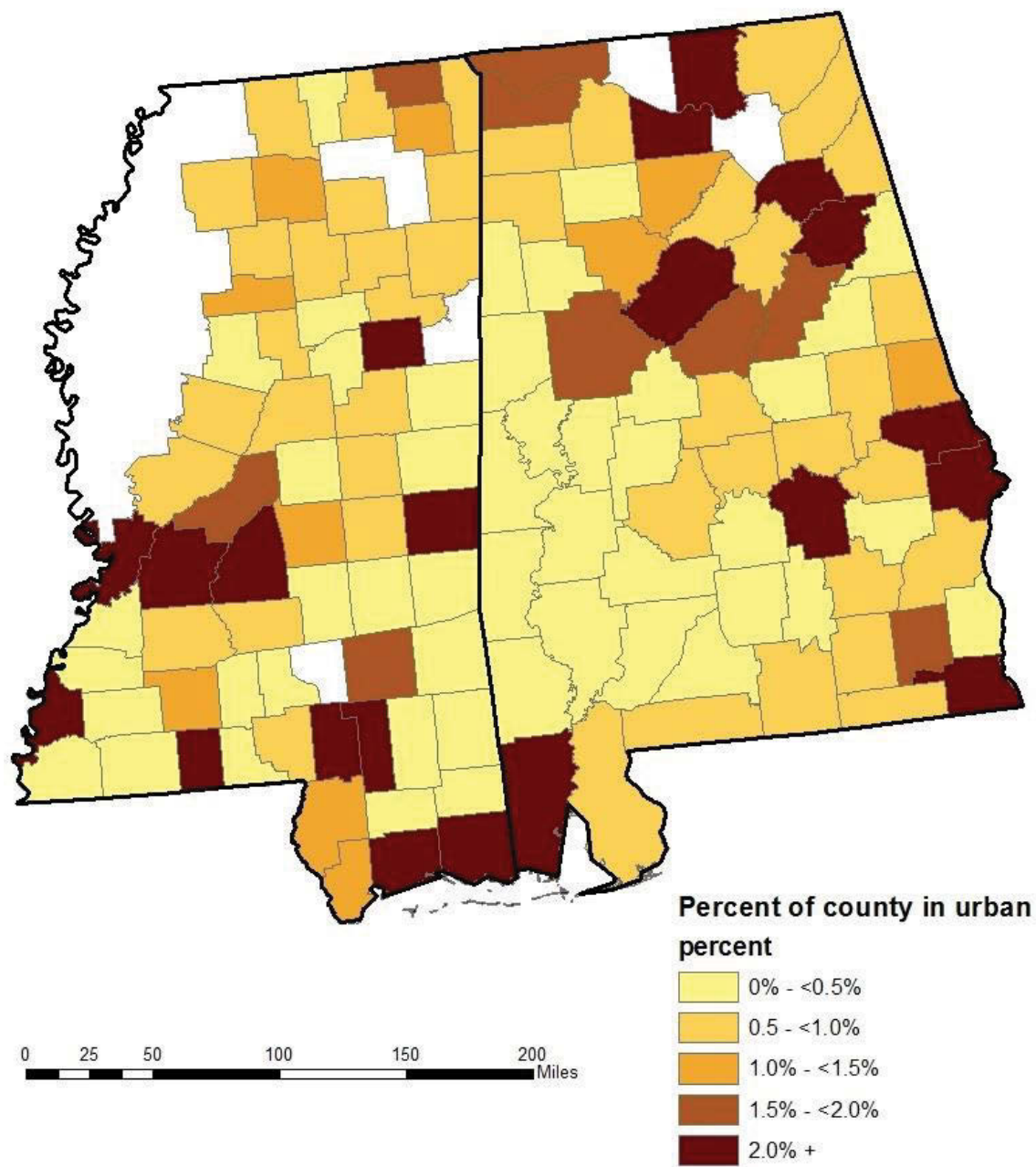


Figure 8. Percent of land base in urban land uses by county, 2011

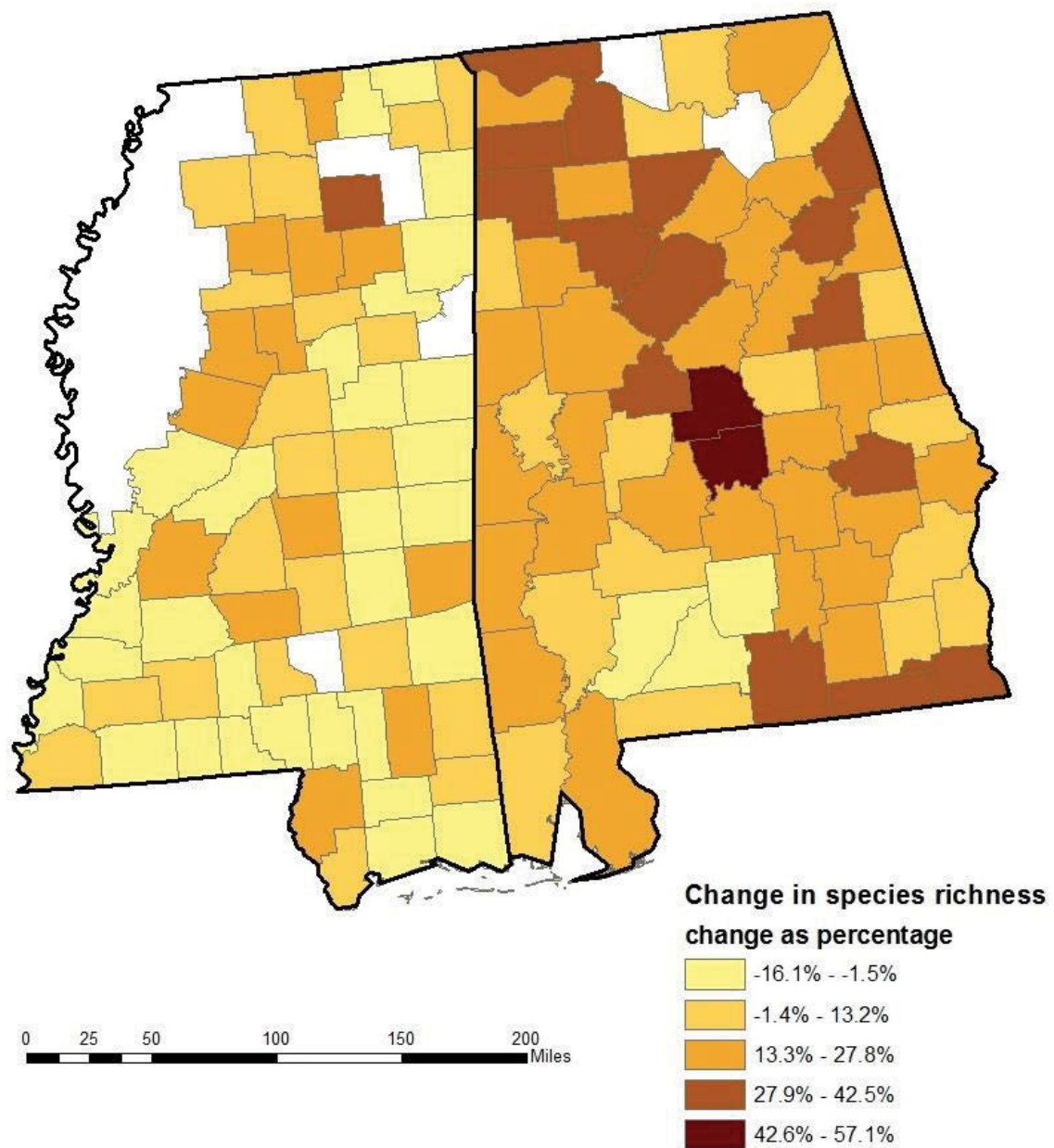


Figure 9. Change in species richness between 1970s and 2013 by county

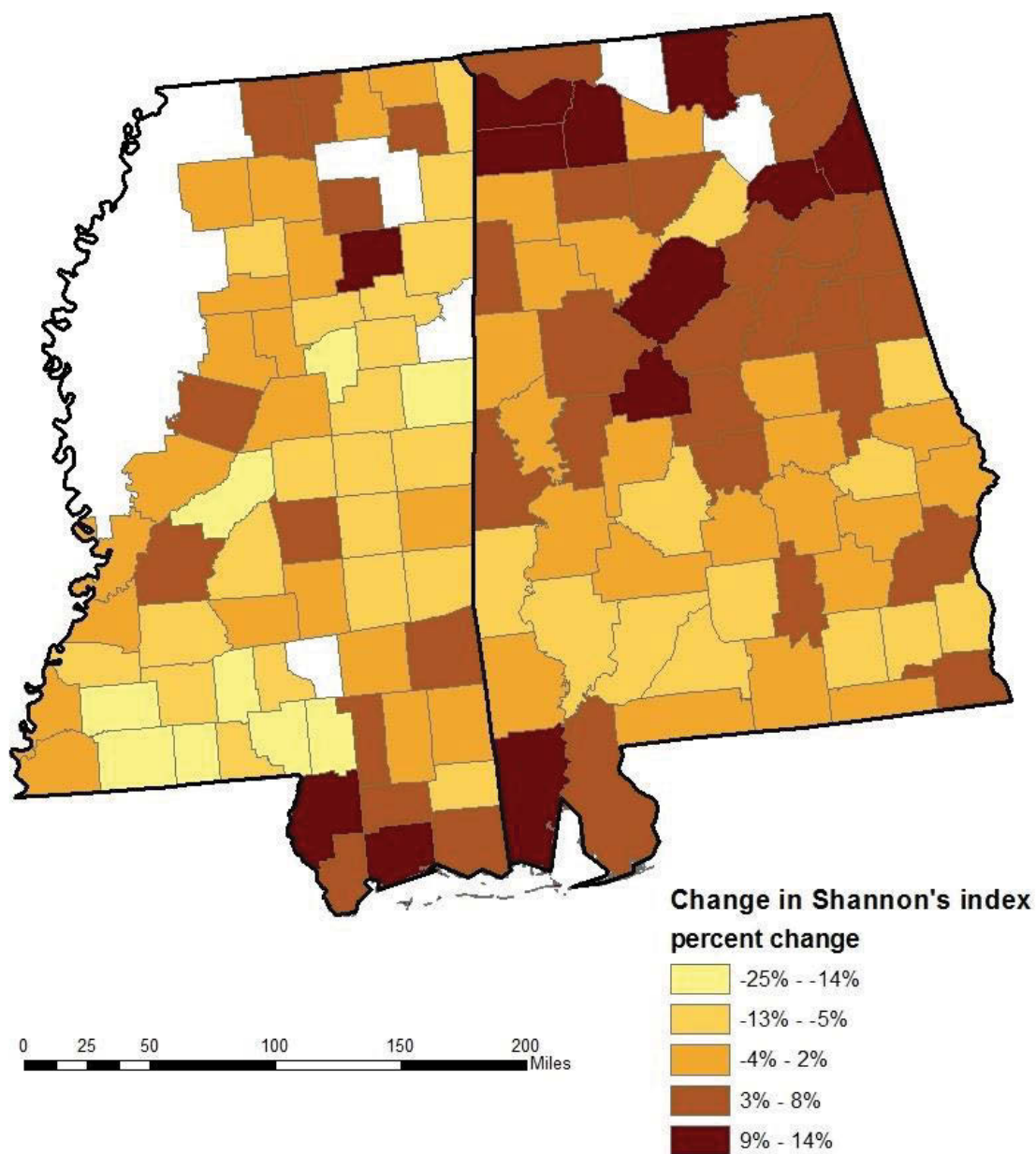


Figure 10. Change in Shannon-Weaver diversity index between 1970s and 2013 by county

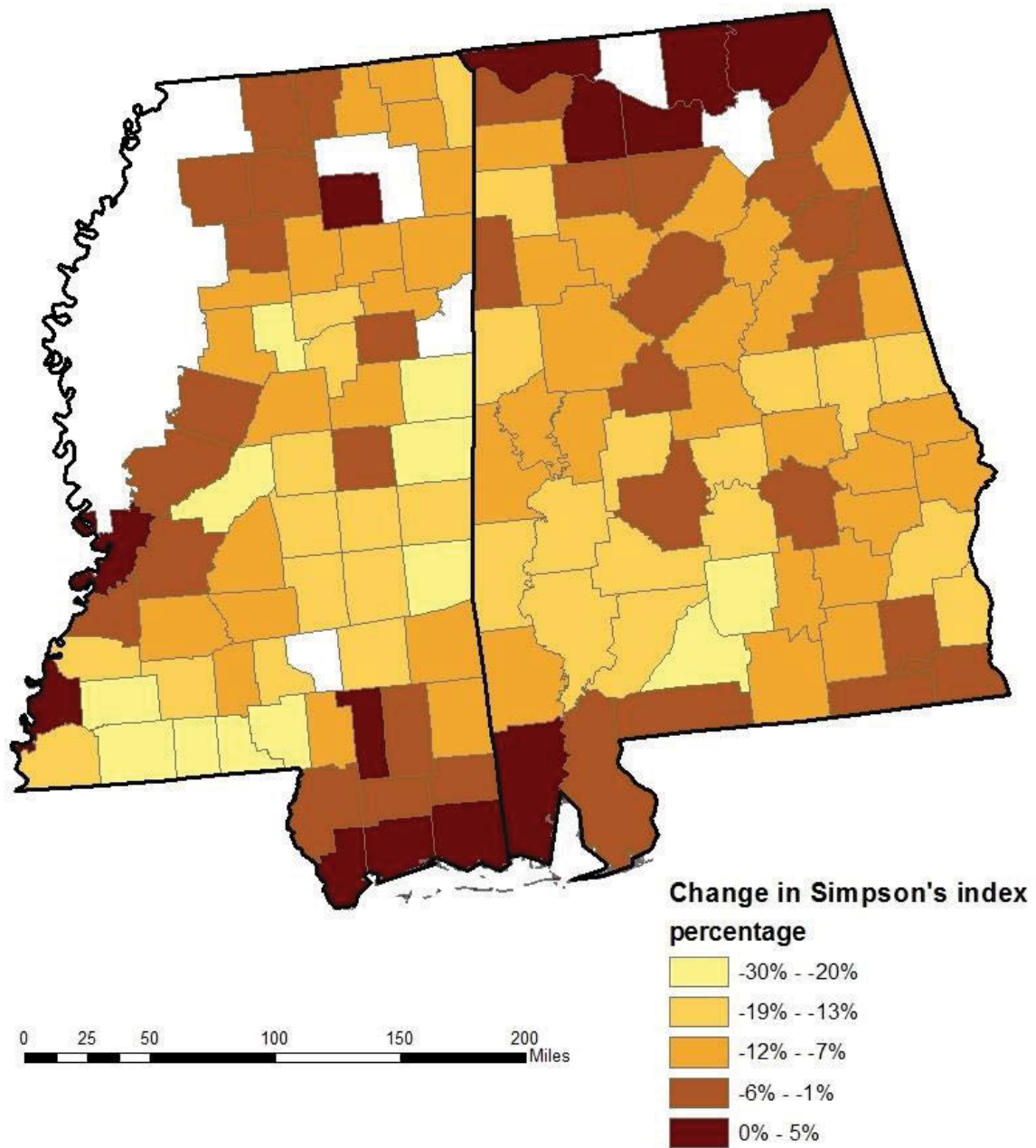


Figure 11. Change in Simpson's diversity index between 1970s and 2013 by county

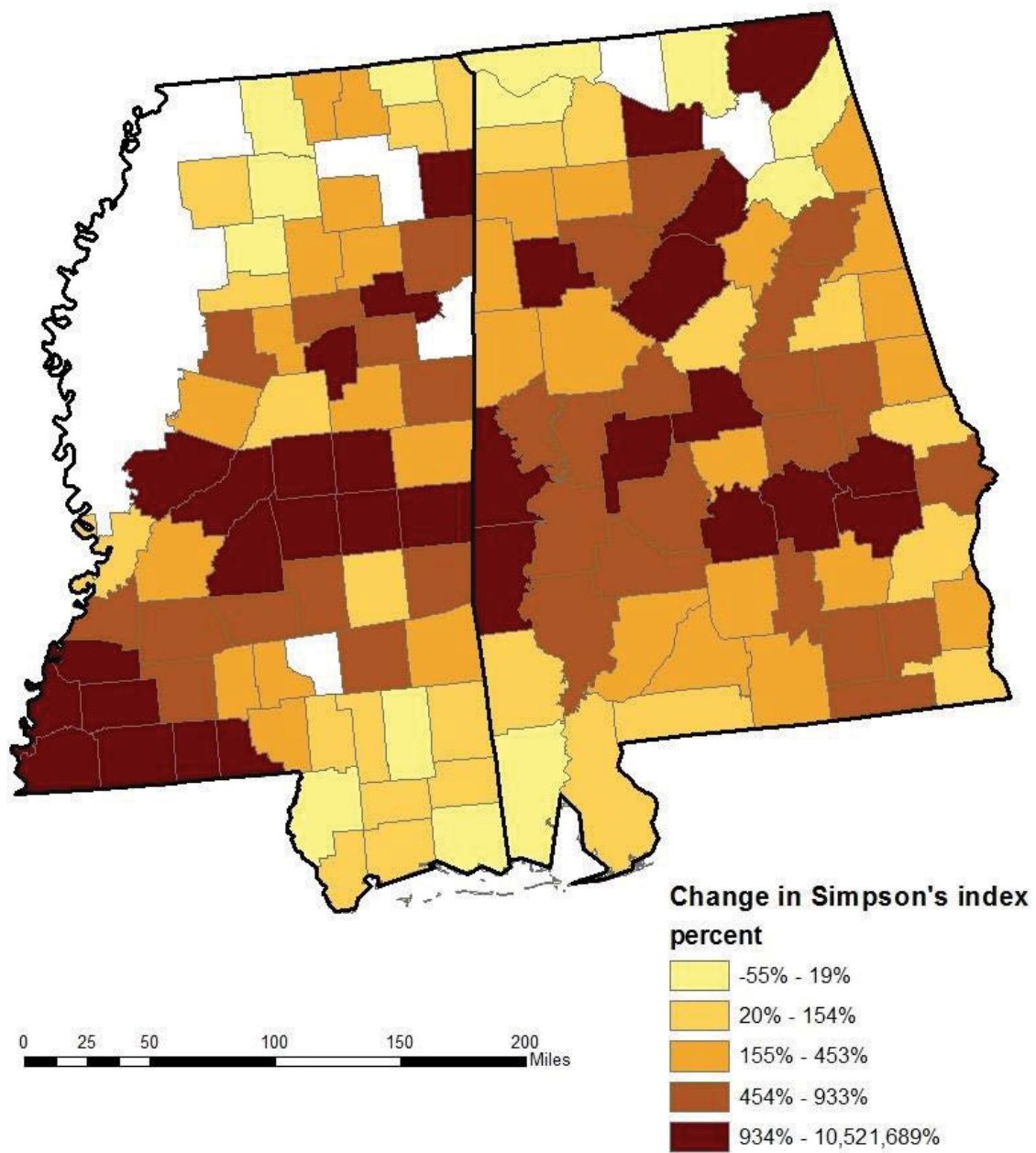


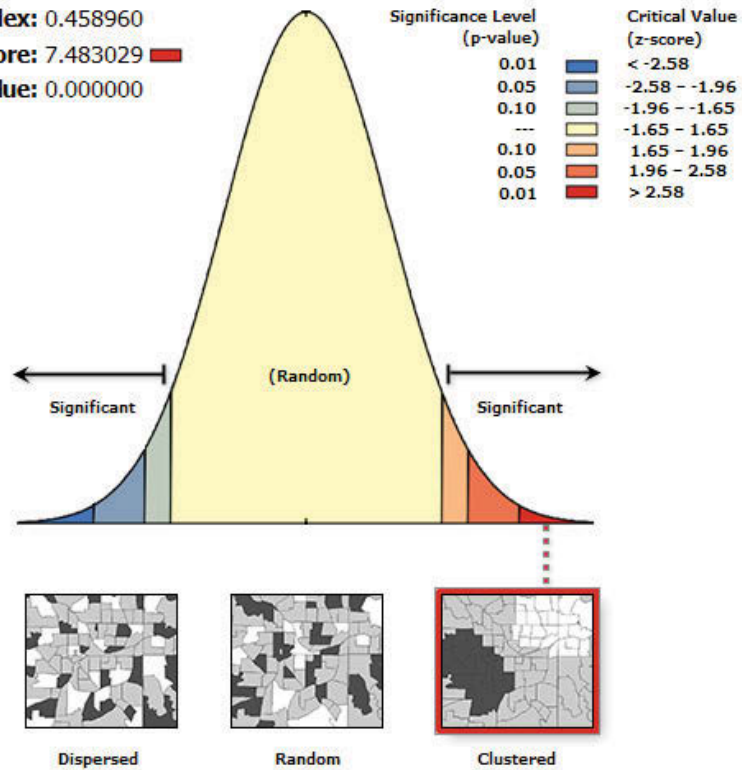
Figure 12. Percent change in plantation acreage between 1970s and 2013 by county

Spatial Autocorrelation Report

Moran's Index: 0.458960

z-score: 7.483029

p-value: 0.000000



Given the z-score of 7.48, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Global Moran's I Summary

Moran's Index:	0.458960
Expected Index:	-0.006410
Variance:	0.003868
z-score:	7.483029
p-value:	0.000000

Figure 13. Spatial autocorrelation report and Moran's I for change in species richness

Table 5. Variable name and description for all observations in dataset.

Variable	Description
STATE	State abbreviation (AL or MS)
COUNTY	County name
FIPS	State and county FIPS code
STATE_FIPS	State FIPS code
SQUARE_MIL	Area in square miles
forest70	Forestland area in 1970s (acres)
planted70	Area of planted forests in 1970s (acres)
pct_plant7	percent of forestland area in plantations in 1970s
rich70	1970s tree species richness
shannon70	1970s Shannon-Wiener index
simpson70	1970s Simpson's index
forest13	Forestland area in 2013 (acres)
planted13	Area of planted forests in 2013 (acres)
pct_plant1	percent of forestland area in plantations in 2013
rich13	2013 tree species richness
shannon13	2013 Shannon-Wiener index
simpson13	2013 Simpson's index
GR_ACRE	Average all-live tree growth (cubic feet/acre)
SWGR_ACRE	Average all-live tree growth of softwood tree species (cubic feet/acre)
HWGR_ACRE	Average all-live tree growth of hardwood tree species (cubic feet/acre)
pet	Average annual potential evapotranspiration (PET) in millimeters per year
ppt	Average annual precipitation in inches
min_elev	minimum elevation in county
max_elev	maximum elevation in county
range_elev	Range of elevations (maximum -minimum) in county
avg_elev	average elevation in county
celsius	average temperature (celsius)
pct_ag	percent of landbase in agricultural uses - 2011 NLCD
pct_urban	percent of landbase in urban uses - 2011 NLCD
pct_forest	percent of landbase in forests - 2011 NLCD
diff_fores	difference in forest area between 1970s and 2013 in acres (2013-1970s)
diff_plant	difference in plantation area between 1970s and 2013 in acres (2013-1970s)

Table 5 Continued. Variable name and description for all observations in dataset.

Variable	Description
chnng_fores	percent change in forest area between 1970s and 2013
chnng_plant	percent change in plantation area between 1970s and 2013
diff_rich	differecne in species richness
diff_shan	difference in Shannon-Wiener index
diff_simp	difference in Simpson's index
chnng_rich	percent change in species richness
chnng_simp	percent change in Simpson's index
chnng_shan	Percent change in Shannon-Wiener index
lat	average latitude of FIA plots in county
lon	average longitude of FIA plots in county

Table 6. Descriptive statistics for species richness, Simpson's index, and Shannon's index for Alabama 1972 and Mississippi 1977 and Alabama and Mississippi 2013.

Variable	N	Mean	Std Dev	Minimum	Maximum
Rich70	157	40.6	7.089	25	57
Simpson70	157	0.901	0.027	0.823	0.957
Shannon70	157	2.776	0.192	2.216	3.291
Richess13	157	44.3	7.775	26	64
Simpson13	157	0.833	0.072	0.636	0.959
Shannon13	157	2.783	0.222	2.153	3.424

Table 7. Correlation coefficients for the 1970s diversity indices

Variable	Pearson Correlation Coefficients, N = 157		
	Prob > r under H0: Rho=0		
	Richness 1970s	Shannon's 1970s	Simpson's 1970s
Forest area 1970s	0.0818	-0.1945	-0.126
	0.3083	0.0146	0.1158
Plantation area 1970s	-0.2143	** -0.3700	-0.1684
	0.007	<.0001	0.035
percent of forest in plantations 1970s	** -0.3090	** -0.3348	-0.1617
	<.0001	<.0001	0.0429
Forest area 2013	0.2222	-0.1047	-0.0815
	0.0052	0.1918	0.31
Planted area 2013	* 0.3334	0.0004	-0.1134
	<.0001	0.9996	0.157
Percent in plantations 2013	0.1833	0.0075	-0.1416
	0.0216	0.9255	0.0769
Average all-live growth per acre	0.2894	* 0.4340	0.2416
	0.0002	<.0001	0.0023
Average softwood growth per acre	0.2201	0.2679	0.0488
	0.0056	0.0007	0.5438
Average hardwood growth per acre	0.2072	* 0.4212	* 0.4130
	0.0092	<.0001	<.0001
pet	* 0.3041	0.0076	-0.0825
	0.0001	0.9251	0.3039
ppt	-0.2277	** -0.3230	-0.2643
	0.0041	<.0001	0.0008
minimum elevation	-0.1405	0.2248	0.2586
	0.0793	0.0046	0.0011
maximum elevation	-0.1449	0.1703	0.2515
	0.0701	0.033	0.0015
range in elevation	-0.1298	0.1201	0.2173
	0.105	0.134	0.0063
average elevation	-0.1335	0.2262	0.2707
	0.0956	0.0044	0.0006
Celsius	0.0043	** -0.3849	** -0.4121
	0.9568	<.0001	<.0001

Table 7 Continued. Correlation coefficients for the 1970s diversity indices

Variable	Pearson Correlation Coefficients, N = 157		
	Prob > r under H0: Rho=0		
	Richness 1970s	Shannon's 1970s	Simpson's 1970s
percent of county in agriculture	-0.0383 0.6335	0.2854 0.0003	*0.3553 <.0001
percent of county in urban	-0.2540 0.0013	-0.3069 <.0001	-0.1197 0.1353
percent of county in forest	0.0420 0.6013	0.0159 0.843	-0.0949 0.237
difference in forest area	*0.4286 <.0001	*0.3445 <.0001	0.1789 0.025
difference in plantation area	*0.4605 <.0001	0.1896 0.0174	-0.0331 0.6808
change in forest area (%)	0.2501 0.0016	0.2547 0.0013	0.2259 0.0045
change in plantation area (%)	0.0954 0.2345	0.1624 0.0421	0.0717 0.3717
average latitude	0.0537 0.5046	*0.4387 <.0001	*0.4404 <.0001
average longitude	-0.2826 0.0003	-0.2606 0.0010	-0.0344 0.6687
* statistically significant variable with a positive correlation > 0.33			
** statistically significant variable with a negative correlation < 0.33			

Table 8. Correlation coefficients for the 2013 diversity indices

Variable	Richness 1970s	Shannon's 1970s	Simpson's 1970s
Forest area 1970s	0.1696	0.2216	0.1935
	0.0338	0.0053	0.0152
Plantation area 1970s	-0.1852	0.1294	0.2804
	0.0202	0.1063	0.0004
percent of forest in plantations 1970s	** -0.3035	0.0067	0.1777
	0.0001	0.9339	0.026
Forest area 2013	0.3006	0.1497	0.0697
	0.0001	0.0613	0.3856
Planted area 2013	0.3773	-0.2525	-0.4700
	<.0001	0.0014	<.0001
Percent in plantations 2013	0.16800	** -0.5368	** -0.7508
	0.0355	<.0001	<.0001
Average all-live growth per acre	0.2859	-0.2356	-0.5561
	0.0003	0.003	<.0001
Average softwood growth per acre	0.1388	** -0.4495	** -0.7485
	0.0831	<.0001	<.0001
Average hardwood growth per acre	*0.3446	*0.3123	0.1811
	<.0001	<.0001	0.0232
pet	0.1792	-0.4389	-0.4038
	0.0247	<.0001	<.0001
ppt	** -0.4666	0.0676	*0.3530
	<.0001	0.3999	<.0001
minimum elevation	0.0486	0.1566	-0.0790
	0.5455	0.0501	0.3256
maximum elevation	0.1265	0.2687	0.0187
	0.1144	0.0007	0.8161
range in elevation	0.1535	0.2972	0.0695
	0.0549	0.0002	0.3869
average elevation	0.1129	0.1660	-0.0946
	0.1591	0.0377	0.2387
Celsius	-0.2395	-0.2698	0.0520
	0.0025	0.0006	0.5176
percent of county in agriculture	-0.0213	0.1362	0.1237
	0.7908	0.0890	0.1227

Table 8 Continued. Correlation coefficients for the 2013 diversity indices

Variable	Richness 1970s	Shannon's 1970s	Simpson's 1970s
percent of county in urban	-0.1996	*0.3027	*0.4757
	0.0122	0.0001	<.0001
percent of county in forest	0.1850	-0.1831	-0.4032
	0.0204	0.0217	<.0001
difference in forest area	*0.3737	-0.2943	-0.4534
	<.0001	0.0002	<.0001
difference in plantation area	*0.4918	-0.3319	** -0.6380
	<.0001	<.0001	<.0001
change in forest area (%)	0.1909	-0.2565	-0.3323
	0.0166	0.0012	<.0001
change in plantation area (%)	0.0535	-0.0995	-0.1331
	0.5059	0.2151	0.0964
average latitude	0.2700	0.2737	-0.0467
	0.0006	0.0005	0.5611
average longitude	0.10935	0.09774	0.0751
	0.1728	0.2233	0.3497
* statistically significant variable with a positive correlation > 0.33			
** statistically significant variable with a negative correlation < 0.33			

Table 9. Correlation coefficients for the 1970s-2013 change in species richness, Simpson's index, and Shannon's index.

Variable	Change in richness	Change in Simpson's	Change in Shannon's
Forest area 1970s	0.0909	0.2436	0.3634
	0.2577	0.0021	<.0001
Plantation area 1970s	0.0049	*0.3513	*0.4210
	0.9518	<.0001	<.0001
percent of forest in plantations 1970s	-0.0250	0.2444	0.2766
	0.7561	0.0020	0.0005
Forest area 2013	0.0823	0.0992	0.2242
	0.3057	0.2164	0.0048
Planted area 2013	0.045	** -0.4443	-0.2306
	0.5758	<.0001	0.0037
Percent in plantations 2013	-0.0247	** -0.7199	** -0.4947
	0.759	<.0001	<.0001
Average all-live growth per acre	0.0561	** -0.6713	** -0.5689
	0.4855	<.0001	<.0001
Average softwood growth per acre	-0.0605	** -0.7949	-0.622
	0.4513	<.0001	<.0001
Average hardwood growth per acre	0.2237	0.0254	-0.0703
	0.0049	0.7523	0.3818
pet	-0.1576	** -0.3849	-0.3910
	0.0486	<.0001	<.0001
ppt	** -0.3454	*0.4697	*0.3459
	<.0001	<.0001	<.0001
minimum elevation	0.2821	-0.1839	-0.0664
	0.0003	0.0211	0.4084
maximum elevation	*0.3916	-0.0792	0.0842
	<.0001	0.3241	0.2944
range in elevation	*0.4038	-0.0127	0.1559
	<.0001	0.8744	0.0512
average elevation	*0.3634	-0.2046	-0.0582
	<.0001	0.0101	0.4691
celsius	** -0.3612	0.2120	0.0904
	<.0001	0.0077	0.2602
percent of county in agriculture	0.0450	-0.0031	-0.1080
	0.5759	0.9697	0.1782

Table 9 Continued. Correlation coefficients for the 1970s-2013 change in species richness, Simpson's index, and Shannon's index.

Variable	Change in richness	Change in Simpson's	Change in Shannon's
percent of county in urban	0.0451	*0.5339	*0.5305
	0.5747	<.0001	<.0001
percent of county in forest	0.2076	** -0.3857	-0.1921
	0.0091	<.0001	0.016
difference in forest area	-0.0535	** -0.5339	** -0.5512
	0.5061	<.0001	<.0001
difference in plantation area	0.0448	-0.6473	** -0.4582
	0.5771	<.0001	<.0001
change in forest area (%)	-0.0643	-0.4224	** -0.4421
	0.4239	<.0001	<.0001
change in plantation area (%)	-0.0617	-0.1635	-0.2186
	0.4425	0.0408	0.006
average latitude	0.3236	-0.2161	-0.1275
	<.0001	0.0066	0.1115
average longitude	*0.5093	0.0846	0.2726
	<.0001	0.2922	0.0006
* statistically significant variable with a positive correlation > 0.33			
** statistically significant variable with a negative correlation < 0.33			

Table 10. Spatial lag model results for percent change in richness between the 1970s and 2013.

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	40.1761	18.8382	2.1327	0.0345
range_elev	0.0539	0.01369	3.9370	0.0001
ppt	-0.0267	0.01142	-2.3391	0.0206
pct_ag	0.0648	0.1041	0.6222	0.0347
n, DF	4, 153			
r-squared	0.3791			
AIC	1267.84			
Moran's I	0.00000			

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	14.2280	23.9743	0.5935	0.5529
range_elev	0.0609	0.0190	3.1988	0.0014
ppt	-0.0115	0.0143	-0.8068	0.0420
pct_ag	0.1582	0.1117	1.4166	0.0156
LAMBDA	0.4907	0.0978	5.0076	0.0000
n, DF	4, 153			
r-squared	0.4324			
AIC	1222.17			

Table 11. Ordinary least squares model results for percent change in Shannon's diversity index between the two time periods.

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	1.23	14.5557	6.2676	0.0000
GR_ACRE	-0.1132	0.0152	-7.4294	0.0000
pet	-0.0599	0.0104	-5.7819	0.0000
chng_plant	-0.0362	0.0000	-3.4826	0.0006
pct_urban	1.7523	0.2632	6.6567	0.0000
n, DF	5, 152			
r-squared	0.5881			
AIC	998.248			
Moran's I	0.92701			

Table 12. Spatial lag model results for percent change in Simpson's diversity index between the two time periods.

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	-4.6203	1.14973	-4.01865	0.00006
diff_plant	-0.0001	0.0001	-6.8388	0.0000
forest13	0.0001	0.0000	4.1356	0.0000
pct_urban	0.4458	0.2264	1.9691	0.0489
n, DF	4153.0000			
r-squared	0.6061			
AIC	951.937			
Moran's I	0			
Variable	Coefficient	Std.Error	z-value	Probability
W_chng_simp	0.5215	0.0731	7.1380	0.0000
CONSTANT	-4.6204	1.1497	-4.0186	0.0001
diff_plant	-0.0001	0.0001	-6.8388	0.0000
forest13	0.0001	0.0088	4.1356	0.0000
pct_urban	0.4458	0.2264	1.9691	0.0489
n, DF	5, 152			
r-squared	0.707524			
AIC	915.901			

CHAPTER III
THE IMPACTS OF VARYING SPATIAL SCALE IN DETERMINING
PREDICTORS OF TREE SPECIES DIVERSITY USING NESTED
WATERSHEDS AND FOREST INVENTORY DATA IN THE
SOUTHEASTERN UNITED STATES

Abstract

This chapter explores the relationship between tree species diversity and various climatic, environmental, and anthropogenic factors in the southeastern United States. Key among these is the impact of plantation forestry, agricultural establishment, and urban development on three measures of tree species diversity: species richness, Shannon-Wiener index, and Simpson's index. Forest Inventory and Analysis data was the source of tree species data and nested watersheds from the Watershed Boundary Dataset (WBD) were used as spatial boundaries. The measures of diversity were calculated for the differing watershed scales. Multivariate analysis and spatial analysis techniques were incorporated to understand how possible predictors and covariates of tree species diversity vary with spatial scale. Multivariate analysis techniques such as multi-response permutation procedures (MRPP), non-metric multidimensional scaling (NMS), and regression trees were used to assess relationships within the data. A species-area curves was created to determine appropriate spatial scales. Getis-Ord G_i^* hot-spot analysis was utilized to determine if diversity hot and cold spots clustered, as well as if clustering patterns changed in regards to changes in area. Finally, ordinary least squares, spatial lag, and spatial error models were developed to explore how scale dependencies exist and interact within these relationships. The results revealed that anthropogenic and local factors were more important in smaller scales, and climate and global factors increased in significance as spatial scale increased. Additionally, variables relating to plantation forestry were found to be positively related to species richness, and negatively related to diversity measures that included species evenness.

Introduction

Human-related activities are impacting the planet's landscapes. Between one-third and one-half of the earth's landscape has been altered by anthropogenic forces (Vitousek *et al.* 1997). These alterations of natural systems and structures can have an impact of maintaining ecosystem integrity (Forman and Gordon, 1981; O'Neill and Hunsaker, 1997; Dale *et al.* 2000). Many studies have established a link between species diversity and forest ecosystem function and productivity (Caspersen & Pacala 2001; Tilman *et al.* 1997). Species diversity is frequently used as a surrogate for forest stability, health, and productivity (McIntosh 1967; McNaughton 1977; Tilman 1996; Austin 2002).

Heywood (1995) states the need for conservation planning to maintain long-term biodiversity. A key component of conservation planning is the predications of tree species distributions across the landscape by incorporating inventory data with a suite of environmental, climatic and anthropogenic factors

(Franklin, 1995; Austin, 2002; Guisan and Zimmerman, 2000; Ellith and Burgman, 2002).

Most common indices of diversity combine the evaluation of two components of diversity: richness and evenness. Richness simply refers to the number of species in a given area. Evenness refers to the relative distribution of individuals within the species in an area, i.e. the number of individuals within in each species. The more equitable the distribution, the more even the estimate. These two components can be ranked differently within the same landscape. Many indices try to balance the two. Recognizing the differences in diversity measures and how they may be similar or different in the same area can reveal a great deal about the underlying forces that are affecting plant communities (Nagendra, 2002).

A key component to diversity studies is spatial scale. Patterns of species diversity differ as a function of spatial scale. The relationship between scale and species richness and diversity has long been studied (Gleason 1922; Cain 1938; MacArthur & Wilson 1967). As the size of area in question increases, so does species richness and indices of diversity and evenness. Additionally, the factors that influence diversity can differ with scale. Regional variables such solar radiation, temperature, precipitation, and elevation may prove to be more significant at larger scales (Currie 1991), while local factors such as disturbance may play a larger role in smaller areas. Determining what factors impact tree species diversity at various scales and understanding the underlying processes involved in these relationships is critical for long-term land management and sustainability.

Study Objectives and Hypotheses

The primary goal of the analysis is to determine the appropriate spatial scale for assessing tree diversity. Specific questions addressed include:

1. What is the appropriate scale to use FIA data for tree diversity studies?
2. How well do nested watersheds partition variance of three measures of tree species diversity?
3. Are there any patterns of similarity/dissimilarity within watersheds?
4. How do predictors of tree species diversity change with spatial scale?
5. Is there a scale at which local factors such as disturbance, nutrition, and anthropogenic forces give way to larger scale factors such as climate or region?

Data, Diversity Indices and Spatial Unit

The USDA Forest Service Forest Inventory and Analysis (FIA) data comprise the foundation of the dataset. This database stores inventory data from sample plots taken across the South. Information such as location, species, DBH, height, basal area, standing volume, and average annual growth, removals, and mortality are easily derived from this system. The 1998 Farm Bill requires that a systematic grid of plots is located across the county, and that a standard set of variables and measurements are performed on each plot. The FIA plot grid assigns one plot to approximately every 6,000 acres. The Farm Bill further states that a systematic number of these plots be inventoried for each state every year (Bechtold & Patterson 2005). The southern states average about 6 years for each cycle. Thus, every 5-7 years, a complete set of plots has been re-inventoried for every state.

FIA estimates used include: average latitude and longitude of plots, percent forest/non-forest, acres of pine plantations, acres southern pine plantations expressed as a percentage of total forest area, total forest area, total land area, average site index, average all-live growth per acre, average all-live softwood growth per acre and average all-live hardwood growth per acre in each spatial unit. Additionally, species richness (or a count of species), Shannon's equitability (for evenness), and Simpsons diversity index were calculated for each spatial unit.

GIS Data Layers

Yearly average estimates of temperature, rainfall and potential evapotranspiration (PET) were derived from map layers containing polygons of average annual estimates of the variable in the contiguous United States, for the climatological period 1990-2009. These Parameter Regressions on Independent Slopes Model (PRISM) derived raster data are published by the PRISM Climate Group at Oregon State University. 100-meter resolution elevation raster data were obtained from the U.S. Geological Survey. A GeoTIFF file of 2011, 30-meter national land cover data (NLCD) were used to calculate percent forest, urban, and agriculture for each county.

The Watershed Boundary Dataset (WBD) forms the areal extent for this study. The WBD was developed by the U.S. Geological Survey. The Dataset is composed of the entire United States which is divided and sub-divided into successively smaller regions or hydrologic units. These nested units are classified into four levels: regions, sub-regions, accounting units, and cataloging units, with regions being the largest and cataloging units the smallest. Each hydrologic unit is identified by a unique hydrologic unit code (HUC) consisting of two to eight digits based on the four levels of classification in the hydrologic unit

system. The smaller the HUC the larger the area. A two digit HUC can be decomposed into a four digit HUCs, which can be further sub-divided into six digit HUCs (U.S. Geological Survey. 2013). Fig 14 illustrates the nesting effect of the WBD. Table 13 reveals the number of HUCs, average, maximum and minimum number of FIA plots in each HUC size, and the total number of FIA plots in each HUC dataset.

Nested watersheds were selected over other possible boundaries such as ecoregions because they can be continuously sub-divided into successively smaller units. Ecoregions may prove to be better units of delineation, but they are not always the same size and cannot be broken down as the smallest basin in the WBD dataset. As one of the goals of this study was to investigate the impact varying spatial scale, WBD was deemed to best accomplish this goal.

All raster data were aggregated to the HUC level using the zonal statistics tool in ArcGIS. Average estimates of temperature, rainfall, PET, and elevation were computed for each HUC. Additionally, range of elevation was calculated as well. The percent of each county in the three main land uses, forest, urban, and agriculture were calculated in a similar fashion.

Diversity Indices

This study assessed three measures of species diversity: species richness, Shannon-Wiener index and Simpson's index. These indices were calculated with the "diversitycomp" command in the "BiodiversityR" package in R-Studio version 1.0.136. FIA tree and plot level data were used to create the data matrix of species counts by plot and watershed.

Species Richness

This study was focused on two types of biological diversity measurements: richness and evenness. Richness is simply the number of different organisms or species living in an area (Wilson 1993; Ehrlich & Wilson 1991). It does consider the proportion or distribution of each subspecies within the area of interest. All one must do is simply count the number of subspecies found in a community. It is the simplest and easiest diversity measurement to calculate and describe. Species richness is a key component in many the plant ecology community (Naeem et al 1996).

Relative species abundance within a community is often measured with an index ranging from 0 to 1 and equated species evenness (Hulbert 1971; Alatalo 1981). Evenness refers to how the subspecies are distributed throughout the whole community (Routledge 1980; Magurran 1988). It, in effect, is a normalization of richness. For example, there are two communities composed

with five different subspecies. The first one has counts of 20 for each subspecies for a total of 100 individual subspecies counted. The second one has one subspecies with a count of 96, and the other four have one observation each. The second community also has a total of 100 individual subspecies counted. The first community would be considered more even, as each subspecies accounts for the same percentage of the total, while the second one is dominated by a single species.

Shannon-Wiener Index

Proportional abundance indices depend on both richness and relative abundance (Hill 1973). Shannon-Wiener (Shannon's for short) evenness and diversity index (H) can from information theory and measure the order and disorder within a population. This index is derived by calculating the proportion of species i relative to the total number of species (p_i), and then multiplying by the natural logarithm of this proportion ($\ln p_i$). The result is summed across species, and multiplied by -1:

$$H = - \sum_{j=1}^s p_i \ln p_i$$

Shannon's Index

Simpson's evenness and diversity index (D) was developed in 1949. Like Shannon's index, it also takes into account richness and diversity. Simpson's D can be found in three forms. The one used in this study is derived by taking the proportion of species i relative to the total number of species (p_i) and then squaring. The squared proportions for all the species are summed, and the reciprocal is taken. The formula for Simpson's index is:

$$D = \frac{1}{\sum_{j=1}^s p_i^2}$$

An evenness or equitability index can be calculated for Simpson's D, just as one can be derived from Shannon's H. Simpson's equitability index (ED) is calculated by taking Simpson's index (D) and dividing it by the total number of subspecies groups (S). Thus, the formula for ED is: $ED = D / S$, where S = total number of subspecies groups. Evenness or equitability takes a value between 0 and 1, with 1 being complete evenness.

All raster data were aggregated to the HUC level using the zonal statistics tool in ArcGIS. Average estimates of temperature, rainfall, PET, and elevation were computed for each HUC, as was range of elevation. The percent of each

county in the three main land uses (forest, urban, and agriculture) were calculated in a similar fashion. Table 14 lists all FIA and GIS data used in this study. Species-area curves and species-basal area curves were developed from plots that occurred on either National Forests or forest industry lands using PC-Ord version 6.22.

Methods

Analysis of variance (ANOVA) and multiple response permutation procedures (MRPP) were conducted to test the hypothesis that smaller watersheds nested inside larger watersheds are more similar than if they belonged in other groups. In each instance, larger watersheds were chosen as the test unit, and species richness, Shannon's index, and Simpson's index were tested for each nested watershed to determine if nesting had an impact.

Hot-Spot Analysis

The next step involved Hot-Spot analysis with ArcGIS10.3. Getis and Ord developed a technique for analyzing high and low clustering within the dataset. The standardized G_i^* statistic is (Gettis & Ord 1992; ESRI 2017):

$$G_i^* = \frac{\sum_{j=1}^n W_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{\sum_{j=1}^n w_{i,j}^2 - \left(\sum_{j=1}^n w_{i,j} \right)^2}{n-1}}}$$

Where:

x_j is attributed to the feature j

$w_{i,j}$ is the spatial weight between the feature in i and j

n is equal to the total number of features

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2}$$

The G_i^* statistic is a z-score, so no additional computations were required. The Getis-Ord G_i^* Hotspot function identifies statistically significant clusters of high values (hot spots) and low values (cold spots). The resulting p-values and z-scores were used to determine if the null hypothesis if neighboring observations

not being similar could be rejected. In order to be statistically significant, the observation must be surrounded by others that feature similar high/low values. The larger the z-score, the more intense the clustering (ESRI 2017). The resulting maps were used to gain an understanding of the dataset and underlying relationships. This knowledge was used in subsequent steps.

Regression Trees

To further gain an understanding of how possible predictors of species diversity change in regards to spatial scale, regression trees for species richness, Shannon's index, and Simpson's index for all watershed sizes were developed using all possible explanatory variables. In prediction trees, data are subdivided or partitioned into smaller boxes. Creating these splits is analogous to variable selection in regression analysis. The split occurs on a single explanatory variable with the goal of maximizing homogeneity between the two groups (Loh 2002; Zimmerman et al 2016). This is referred to as binary partitioning because it is always a two-way split (Loh 2002; Moison 2008). The point of the split is called a node. A parent node is always split into two child nodes. Each of the resulting child nodes are then subjected to the same splitting criteria as the parent node. Thus, each child node becomes a parent (Loh 2002; Moison 2008). The term recursive is used to describe this continuous splitting. Recursive partitioning is the phrase used to describe the prediction tree process (De'Ath and Fabricius 1999; Moisen 2008; Schmidhammer 2009; and Loh 2011).

A typical prediction tree is displayed growing upside down. The root is the initial split each node or split is can be viewed as a leaf or branch. The tree will "grow" downward as more leafs are created. Ultimately an observation passes down through the tree and its series of decision splits until it resides in a terminal node composed of similar observations that form a homogenous group. The goal of prediction trees is to partition the response into the groups than maximize homogeneity (or minimize impurity) but also keep the tree size small (De'Ath & Fabricius 1999; Moisen 2008).

Nonmetric-Multidimensional Scaling

Non-metric multidimensional scaling (NMS) is a relatively new free ordination tool that is becoming more popular in species datasets. Peck (2010) suggests that it is even beginning to replace CCA as the "tool of choice for ecological analysis". The reasons for this are many. First NMS is truly a non-parametric tool. NMS handles data that is nonnormal or are arbitrary, unimodal, linear, discontinuous, or has questionable scales (McCune and Mefford 1999; Peck 2010). It is useful when communities exist on a long gradient. Paliy and Shankar (2016) indicate that NMS is one of the best analysis techniques to use when seeking similarities among objects is to be assessed using distance based

calculations. NMS is useful when the goal is to find pattern within a matrix of multiple responses. It does not require a specific sample design and can be used to explore the structure of almost any dataset (Peck 2010).

NMS is an iterative search and ranking process where the placement of n entities on k dimensions (axis) is accomplished in a way to minimize stress of the k -dimensional configuration. The final output is an NMS ordination score that have the same rank as the original data matrix. Once this has been achieved, NMS works backward to create a final ordination score. These final scores are then refined through an iterative process until their rank order matches the original rank as close as possible.

Model Development

Finally, a series of linear models was developed using the methodology developed by Anselin in 2005 with GeoDA version 1.8.14.8. Ordinary least squares (OLS) models were created for change in species diversity, change in Shannon-Wiener index, and change in Simpson's index for all watershed sizes. Diagnostics for spatial autocorrelation (SA) were performed. If the results proved inconclusive, then the OLS results would be kept. If SA proved to be an issue, then either a spatial lag model (SLM) or spatial error model (SEM) was then used.

The spatial lag model (SLM) incorporates the influence of not only the independent variable, but the effect that neighboring attribute values by incorporating a "spatially lagged" variable of y on the right hand side of the regression equation (Anselin, 1988, 1990, 1995). The formula for the SLM model is: $Y = \rho W_y + X\beta + e$, where W_y is the weighted average of neighboring y 's, i.e. the spatially lagged variable and ρ is the spatial autoregressive coefficient. The SLM model primarily controls spatial autocorrelation in the dependent variable. The SLM treats spatial dependence as substance rather than a nuisance. If the SLM model is properly specified, there should be little or no spatial dependence remaining in the residuals (Anselin 1988, 1990, 1995).

The inclusion of the lagged y variable is appropriate in cases where the values of y interact and influence each other. Examples of possible dependence among observations include; the presence of a species is more likely found in a neighboring site due to seed fall or coppice; an individual home value is likely impacted by the value of neighboring houses; a country is more likely to have similar wealth and levels of education to its neighbors due to interactions between the two. The SLM incorporates the influence of unmeasured independent variables plus the effects of neighboring attribute values. This model is often referred to as the simultaneous spatial autoregression model (Anselin 1988; Ward & Gleditsch 2007; Matthews 2009).

The spatial error model (SEM) captures the influence of unmeasured independent variables upon each other. Spatial error models describe the extent to which the spatial clustering of a variable is not explained by the measures of independent variables but accounted for the clustering of errors. The formula for SEM is: $Y = X\beta + \lambda Wu + e$, where Wu is the spatial lag of the error terms and λ is the autoregressive coefficient and indicates the extent to which the spatial component of the errors is correlated with nearby observations (Anselin 1988, 1990, 1995; Ward & Gleditsch 2007).

If there is no SA between the errors for neighboring observations, then λ will be 0. If $\lambda \neq 0$ then there is SA between the error terms and the connected observations. There are consequences of failing to address SA between the errors and performing on OLS instead. The OLS estimates would actually be unbiased, but the standard errors of the coefficients would be wrong. This due to the fact that the OLS model depends on an estimate of variance than assumes independent observations. If this assumption is violated, then the OLS estimate of variance will tend to underestimate the actual variance. Additionally, the estimated coefficients may not be efficient or “close” to the true values (Anselin 1988, 1990, 1995; Ward & Gleditsch 2007; Matthews 2009).

Results

Species/area curves were developed to ascertain the best spatial scale for this study. As FIA data were the basis of the diversity indexes, species/plot curves were used instead of species/area. Curves for two divergent ownership groups (National Forest and forest industry) were developed. Plots belonging to the two ownership groups were separated into two files. The species/area tool in PC-Ord version 6.22 was then used to create the curves. Interestingly, both curves have almost the same shape, with the only difference being National Forest plots have slightly more species (figure 15). The curves peak around the 100 to 150 plot range. As each FIA plot represents approximately 6,000 acres, this translates into roughly 600,000 to 900,000 acres. This is the same size as HUC08 watersheds. Therefore, it is safe to assume that HUC08 and larger watersheds can be used for analysis.

Species/basal area plots were developed at the same time (figure 16). The results for these reveal that while both ownership groups have similar species/area curves, their species distributions differed substantially. The majority of forest industry plots basal area was comprised of just one species, while the basal area of National Forest plots was spread among 2-7 species.

Nested watersheds provide a logical spatial unit to analyze species diversity. One would assume that species composition and structure would be similar within watersheds. But is this a valid assumption? Have anthropogenic

forces impacted and fragmented the landscape to a degree to which this is not the case? Do larger spatial scales wash-out homogeneity? The traditional tool for determining between group and among differences is ANOVA. ANOVA would not be appropriate, however, as this dataset is unbalanced, non-normal, and most likely non-independent. Therefore, multi-response permutation procedures (MRPP) were used to test for between and among group differences.

MRPP was performed with PC-Ord version 6.22. Table 15 reveals the MRPP results for HUC04 watersheds. MRPP was performed on each of the three watersheds nested within the HUC04 size as well as each of the three diversity measures used in this study. All diversity measures in all nested watersheds possessed highly significant p-values, indicating that the species compositions of each smaller watershed were more similar to the ones within the same HUC04 watershed than would be expected from a random distribution. Additionally, the A-statistic approached 0.3 for HUC08 Shannon's and Simpson's index. McCune and Grace (1999) indicate that 0.3 is considered high for ecological datasets.

The same MRPP procedures were performed on all watershed nested with HUC06. These results indicate that watersheds inside HUC06 were more similar to each other (table 16). All p-values are 0.0000 and the A-statistic for HUC08 Shannon's and Simpson's diversity indices were greater than 0.3. The MRPP analysis indicated that watersheds partitioned variance for the dataset. Therefore, the next step involved mapping the data to determine if any patterns exist that may be identified visually.

Figure 17 displays the three diversity measures and planted acres for the four watershed sizes. The basic pattern is the same for all spatial scales. Species richness was highest in the central Gulf Coastal Plain, while Simpson's and Shannon's diversity was highest in the northern portions of the study area that coincide with the Appalachian Mountains. Plantations occur least in the Mississippi Delta, southern Florida, and the northern portions of the study area, particularly in the southern Appalachian Mountains. Thus, there visually appears to be positive correlation between plantations and species richness and a negative correlation between plantations and the other two diversity measures.

Hot-spot (Gettis-Ord G_i^*) analysis on all four watershed datasets was performed with ArcMap 10.3.1. Analysis of the diversity indexes, measures of plantation establishment, three classifications of land use by percentage of watershed, elevation, climatic variables, and forest productivity were included in the analysis. Like the choropleth map in figure 17, all hot-spot maps reveal similar patterns for all watershed sizes. Species richness, plantation measures, and forest growth per acre measures were highest in the Gulf Coastal Plain. Shannon's, Simpson's, elevation were hot in the same areas and hot where

plantations and softwood growth are cold. Agriculture and urban land uses were hot in areas where all diversity measures are cold (figures 18-21).

Regression trees were fitted to further investigate the relationship between the variables in the dataset and the diversity measures for each watershed size (figures 25-36). As a general rule, the trees were small and simple at larger spatial scales, and grow in size and complexity as spatial scale decreases. This indicates that major environmental and geologic forces play a greater role in larger scale studies.

The regression trees for species richness (figures 22, 25, 28 and 31) indicate that percent of watershed in forest, followed by measures of productivity were the first nodes for the two larger watersheds (HUC04 and HUC06). As the spatial scale decreases, elevation variables became more important. The most complex tree, HUC08, includes leaves for percent urban, precipitation, and longitude as well.

Shannon-Wiener regression trees all have elevation as the first node (figures 23, 26, 29 and 32). The HUC04 tree has only one additional variable, percent of the watershed in urban land-use. This is the only HUC04 tree with a variable that is not environmental, geological, or environmental. Productivity, land-use, plantation establishment and environmental variables play roles in smaller scale trees.

Simpson's regression trees for the three largest watersheds have variables pertaining to elevation as the most important variable (Figures 24, 27 and 30). Measures of productivity is almost always in the second level of leaves. Latitude, land-use, and temperature appear in later break points. The regression tree for HUC10 Simpson's index is one of the deeper trees and has temperature as the first node (figure 33). Second level variables included percent forest and range in elevation. Other variables include productivity, range in elevation, percent in plantations, and precipitation.

Non-metric multidimensional scaling (NMS) was performed to further search for relationships within the dataset. NMS was accomplished using PC-Ord version 6.22 with autopilot set for long and thorough and a Sorensen (Bray-Curtis) distance measure. Ordinations of all watershed sizes was successful (tables 17-20). Larger watersheds, HUC06 and HUC06, had two-dimensional solutions, while the smaller watersheds had three-dimensional solutions. Axis 1 for the HUC04 NMS ordination had variables relating to plantation with high negative scores and dummy variables accounting for the Mississippi delta and watersheds neighboring the gulf coast ranked highest. HUC06 NMS axis 1 scores are similar to HUC04 with the addition of softwood growth now being the lowest score. The first axis for both HUC08 and HUC10 three-dimensional

solution is similar but reversed. HUC08 has geographic dummy variables with the lowest scores and elevation variables the highest. Conversely, HUC10 has elevation variables the lowest and geographic flags the highest.

NMS analysis was then performed on the watersheds classified as either hot or cold in the previously discussed Hot-Spot analysis. For example, all watersheds classified as being hot for species richness were grouped in a dataset and run through the NMS procedure. This was performed on all watershed sizes and all three diversity measures. Unfortunately, all but one failed to produce a successful ordination. The cold spots for Simpson's diversity index for the HUC08 watersheds did produce a successful ordination (table 21). Comparing the first axis of this ordination to the one derived by the entire dataset reveals that both give the lowest scores to the geographic flags. However, the NMS ordination for Simpson's index for the entire dataset (table 19) provided the highest scores to the variables related to elevation. This seems logical, as the hot spot analysis for HUC08 Simpson's index (figure 23) revealed a hot spot in the southern Appalachian mountain region where elevation variables are prominent. Conversely, cold spots appear away from the mountainous regions and in the same watersheds that have plantations.

Models for all three diversity indicators for each watershed size were developed (tables 22-33). As a general rule, larger watersheds produced simpler models that contained fewer independent variables, higher R^2 s, less multicollinearity, and almost no spatial autocorrelation (SAC). This changed as watershed area decreased. Smaller watersheds contained more variables, and had significant multicollinearity, heteroskedasticity, and SAC issues. Many of the smaller watershed's models had lower R^2 and AIC scores as well.

The species richness model for HUC04 watersheds had an R^2 of 0.90 and contained three independent variables: total forest area, average growth of live trees per acre, and range in elevation. These variables seem logical, as more forest area and higher productivity should result in more species. Likewise, elevational diversity gradient (EDG) states that there is a positive correlation between increases in elevation and species richness (Rahbek, 1995). Forest area was present in all species richness models and all-live growth and range of elevation are present in all models except the HUC 10 watershed. As watersheds became smaller, environmental variables became significant and geographic flags appeared in the smallest HUC.

HUC08 and HUC10 species richness OLS models had significant SAC. Spatial error models (SEM) were developed to address this issue. The AICs for both SEMs were smaller than the OLS. The HUC08 model added percent of watershed in agriculture and average temperature along with the spatial error coefficient. Interestingly, percent of watershed in agriculture had a positive

value. HUC10 SEM model added percent of forest area in plantations and two geographic dummy variables representing the Mississippi delta and southern Florida. All three of these variables had negative coefficients.

Shannon's H models for the two larger watersheds revealed that variables pertaining to forest area, elevation and agriculture were significant. The HUC04 watershed had an R^2 of 0.62 and the HUC06 model had an R^2 of 0.53. SAC was not a factor in either of these models.

Shannon's H models for the two smallest watersheds added more local variables such as percent planted and agriculture as well as environmental predictors like PET and softwood growth per acre. Additionally, both had significant SAC. The HUC08 model included a spatially lagged variable for Shannon's H, while the HUC10 model was a SEM model with an autoregressive error term in the equation.

The four models of Simpson's diversity index follow the same pattern as the other indicators, where SAC was not an issue with the two largest watersheds, while the smaller watersheds had significant SAC. Unlike species richness and Shannon's index, there is no set of variables that is significant in every model. Each Simpson's model is unique. However, like the previous models SAC is not an issue with the two larger watersheds and it is a problem with the smaller two HUC sizes. In both cases, SEM was the method used to address SAC in these models.

Discussion

Results of MRPP analysis indicate that watersheds are a viable spatial unit for tree species diversity studies. While this is to be expected, it proves that climatic changes and anthropogenic forces have not altered the landscape to a level where this is no longer the case. Additionally, species-area curves of FIA data confirm that 600,000 acres minimum is a good threshold for FIA data for this study. HUC04 watersheds are definitely suitable for use in this study. One possible problem with HUC04 watersheds is that they may be so large that they mask some of the underlying processes that occur. That is, they may be too coarse for analyzing certain predictors.

HUC06 watersheds definitely fall into the acceptable category. Additionally, models developed from HUC06 watersheds were simpler, consisted of independent variables that behave as they should, and presented little or no problems with normality, heteroscedasticity, and SAC. The R^2 s of HUC04 and HUC06 watersheds were all acceptable as well. HUC08 watersheds were slightly less than the desired size, but they may be useful in some cases, such as where

the forest community structure is more homogeneous and variance between the different communities is low. Problems might arise in HUC08 watersheds that have high degree of variance and/or land-use within them. It is for these reasons that the models developed for HUC08 watersheds begin to have issues such as SAC, heteroscedasticity and lower R^2 and AICs. Models of HUC10 watersheds proved to be complex and wrought with statistical problems.

The primary drivers of all species richness models were related to forest area and productivity. These were followed by elevation variables. All of these are logical and are appropriate. The relationships between scale, productivity, elevation, and species richness and diversity has long been studied and understood and discussed earlier. As area, productivity, and elevation increase (to a certain extent), species richness increases.

The percent of watershed in agriculture and in pine plantations become significant variables in the smaller watersheds, HUC08 and HUC10. This is likely due to the intermediate disturbance hypothesis (IDH) that states diversity will also be low in areas of high and/or frequent disturbance, because only a few species are tolerant of such landscapes (cite?). These are generally pioneer species that first appear after disturbances. Both agriculture and active forest management create edges and small areas of differing land uses. This promotes a wider range of species. Additionally, these areas may be more susceptible to the introduction of new species into an area.

The percent of a watershed in planted stands had a negative impact on species richness at the HUC10 level. There are several possible reasons for this. One explanation is that this not the best scale for this analysis. Earlier results on appropriate scale for FIA data revealed that the area in this watershed size is not sufficient for this study. But if these results are correct, this is the “tipping point” at which plantations have a negative impact on richness. The complete conversion of a natural hardwood stand to a pine plantation, for example, would cause species richness to decline for that particular site.

The primary drivers of Shannon’s richness were variables related to elevation, agriculture, and forest area. All of these were positively related to Shannon’s richness. The reasons for this are the same as those discussed for species richness. Plantations were negatively related to Shannon’s diversity for the two smaller watersheds. Shannon’s differs from species richness in it balances species richness with evenness. While species richness may increase with the presence of anthropogenic disturbance such as establishing pine plantations, they may have a negative impact on evenness. Consequently, there are more species, but the species distribution is dominated by a few species.

Plantation variables were negatively related to all models on Simpson's diversity index for the reasons discussed in the previous paragraph. Elevation was positively related, reflecting the effects of elevational diversity gradient (EDG). Interestingly percent of watershed in agriculture possessed a positive sign. This indicates that while agriculture is a source of disturbance, it does not promote homogenous communities

An interesting artifact of the data and study area became apparent during the development of these models. These models were created with the idea of applying climate change estimates to determine their impacts on species diversity. However, there were many cases where the diversity measures did not behave as anticipated. For example, species richness is not supposed to decrease as rainfall, temperature, or PET increase. Other measure would decrease as productivity increased. This was confounding initially. However, further data interrogation revealed that plantation establishment occurred in fertile areas that have favorable environmental conditions. Humans locate agriculture and managed forests in areas of favorable environmental and site locations. Therefore, plantations are positively correlated to environment and productive sites. Literature detailing the clearing of original forests for agriculture and later the conversion of these farms to managed forests is extensive. USDA (1988), Owen (2002), and Mitchel and Duncan (2009) all detail this transformation in detail with numerous examples and citations.

Conclusions

The southeastern United States "Pine Belt" has been repeatedly disturbed due to anthropogenic activity. This disturbance will confound many long-term studies that focus on either climatic or human impacts on tree species diversity, particularly for the purposes of prediction. However, this type of research can reveal some important information, including: 1) watersheds can be used for nested studies; 2) HUC04 and HUC06 watersheds are an appropriate scale for FIA data and tree diversity studies; 3) forest area, productivity, and elevation are positively correlated to most diversity measures; 4) the Pine Belt has been repeatedly disturbed due to anthropogenic activity; 5) tree species diversity is impacted by these disturbances; 6) these impacts can be both positive and negative, as many disturbances increase species richness; 7) plantations increase the number of species, perhaps due to edge effects from the management; and 8) increasing plantations has a negative impact on both Shannon's diversity index and Simpson's index as both decreases with plantation acreage.

References

- Alatalo, R. V. 1981. Problems in the measurement of evenness in ecology. *Oikos* 37:199–204.
- Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht, The Netherlands
- Anselin, L., 1990. Spatial dependence and spatial structure instability in applied regression analysis. *J. Reg. Sci.* 30: 185–207.
- Anselin, L. 1995. The Local Indicators of Spatial Association LISA, *Geographical Analysis* 27: 93-115.
- Anselin, L. 2005. Exploring Spatial Data with GeoDa: A workbook. Center for Spatially Integrated Social Science. Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana-Champaign. <http://geodacenter.asu.edu/system/files/geodaworkbook.pdf>. 244p.
- Austin M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Model.* 157:101–18
- Cain, S.A. 1938. The species-area curve. *American Midland Naturalist*, 19, 573-581.
- Caspersen, J.P. Pacala, S.W. 2001. Successional diversity and forest ecosystem function. *Ecological Research* 116: 895-903.
- Currie, D.J. 1991. Energy and Large-Scale Patterns of Animal- and Plant Species Richness. *The American Naturalist* 137: 1, 27-49.
- Dale, V.H., Brown, S., Haeuber, R.A., Hobbs, N.T., Huntly, N., Naiman, R.J., Riebsame, W. E., Turner, M.G., and Valone, J.J. 2000. Ecological principles and guidelines for managing the use of land. *Ecological Applications*, 10, 639–670.
- De'ath, G. and Fabricius, K. E. 2000. Classification and regression trees: A powerful yet simple technique for ecological analysis. *Ecology* 81: 3178–3192.
- Ehrlich, P.R. and Wilson, E.O. 1991. Biodiversity studies: science and policy. *Science* 253.5021: 758p.

- Elith, J. and Burgman, M. A. 2002. Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. – In: Scott, J. M. et al (eds), Predicting species occurrences: issues of accuracy and scale. Island Press: 303–314.
- ESRI. 2017. Online help for ArcGIS desktop.
<http://resources.arcgis.com/en/help/>.
- Forman, R. T. T. and Godron, M. 1981. Patches and structural components for a landscape ecology. *Bio. Science* 31: 733-740.
- Franklin, J. 1995. Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. *Prog. Phys. Geogr.* 19, 474-499.
- Getis, A., and J. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24, 189-206.
- Gleason, H.A. 1922. On the relation between species and area. *Ecology*, 3, 158-162.
- Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135, 147-186.
- Heywood, V.H. (Ed.). 1995. Global biodiversity assessment. Cambridge: Cambridge University Press.
- Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–432.
- Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52: 577–586.
- Loh, W.-Y. (2011). Classification and regression trees. WIREs Data Mining Knowledge and *Knowledge Discovery* 1: 14–23.
- MacArthur and Wilson. 1967. The Theory of Island Biogeography. Princeton University Press: Princeton, NJ.
- Magurran, A.E. 1988. Ecological Diversity and Its Measurements. Chapman & Hall, London, 179 p.

- Matthews, S. 2006. GeoDa and spatial regression. Center for Spatially Integrated Social Sciences. Presentation at GIS for Population Scientist, Pennsylvania State University. June 9, 2006.
- McCune, B and M.J. Mefford. 1999. PC-ORD. Multivariate Analysis of Ecological Data, Version 4. MjM Software Design, Gleneden Beach, Oregon, USA.
- McIntosh, R. I. 1967. An index of diversity and the relation of certain concepts to diversity. *Ecology* 48:392–404.
- McNaughton, S. J. 1977. Diversity and stability of ecological communities: a comment on the role of empiricism in ecology. *American Naturalist* 111:515–525.
- Moisen, G. G. 2008. Classification and regression trees. In: Jorgensen, Sven Erik; Fath, Brian D., eds. Encyclopedia of Ecology, Volume 1. Oxford, U.K.: Elsevier. p. 582-588.
- Naeem, S., et al. 1996. Biodiversity and Plant Productivity in a Model Assemblage of Plant Species. *Oikos* vol. 76, no. 2: 259–264. JSTOR, www.jstor.org/stable/3546198.
- O'Neill, R.V. and Hunsaker, C.T. 1997. Monitoring environmental quality at the landscape scale. *BioScience*, 47, 513–519.
- Paliy, O and V. Shankar. 2016. Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* 25: 1032-1057.
- Peck, J.E. 2010. Multivariate Analysis for Community Ecologists: Step-by-Step using PC-ORD. MjM Software Design, Gleneden Beach, Oregon, USA.
- Routledge, R. D. 1980. Bias in estimating the diversity of large, uncensused communities. *Ecology* 61:276–281.
- Schmidhammer, J.L. July 2009. Class notes and personal correspondence.
- Tilman D., Knops J., Wedin D., Reich P., Ritchie M. & Sieman E. 1997. The influence of functional diversity and composition on ecosystem processes. *Science* 227: 1300–1302.
- U.S. Geological Survey. 2013. National Hydrography Geodatabase: The National Map viewer available on the World Wide Web (<http://viewer.nationalmap.gov/viewer/nhd.html?p=nhd>), accessed [July 28, 2014]

- Vitousek, P.M., Flannigan, M.D., Wotton, B.M., and Suffling, R. 1997. Human domination of the Earth's ecosystems. *Science*, 277, 494-504.
- Ward, M.D., Gleditsch, K.S. 2008. Spatial Regression Models. Vol. 155. Sage Publications Inc.
- Wilson, Edward O. 1993. The Diversity of Life. W. W. Norton & Company. 448p.
- Zimmerman, G.M., H. Goetz. P. W. Mielke, Jr. 1985. Use of an Improved Statistical Method for Group Comparisons to Study effects of prairie fire. *Ecology* 66:606-611.
- Zimmerman, R. K., Balasubramani, G. K., Nowalk, M. P., Eng, H., Urbanski, L., Jackson, M. L., et al. Wisniewski, S. R. 2016. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. *BMC Infectious Diseases*, 16, 503.
<http://doi.org/10.1186/s12879-016-1839-x>

Appendix

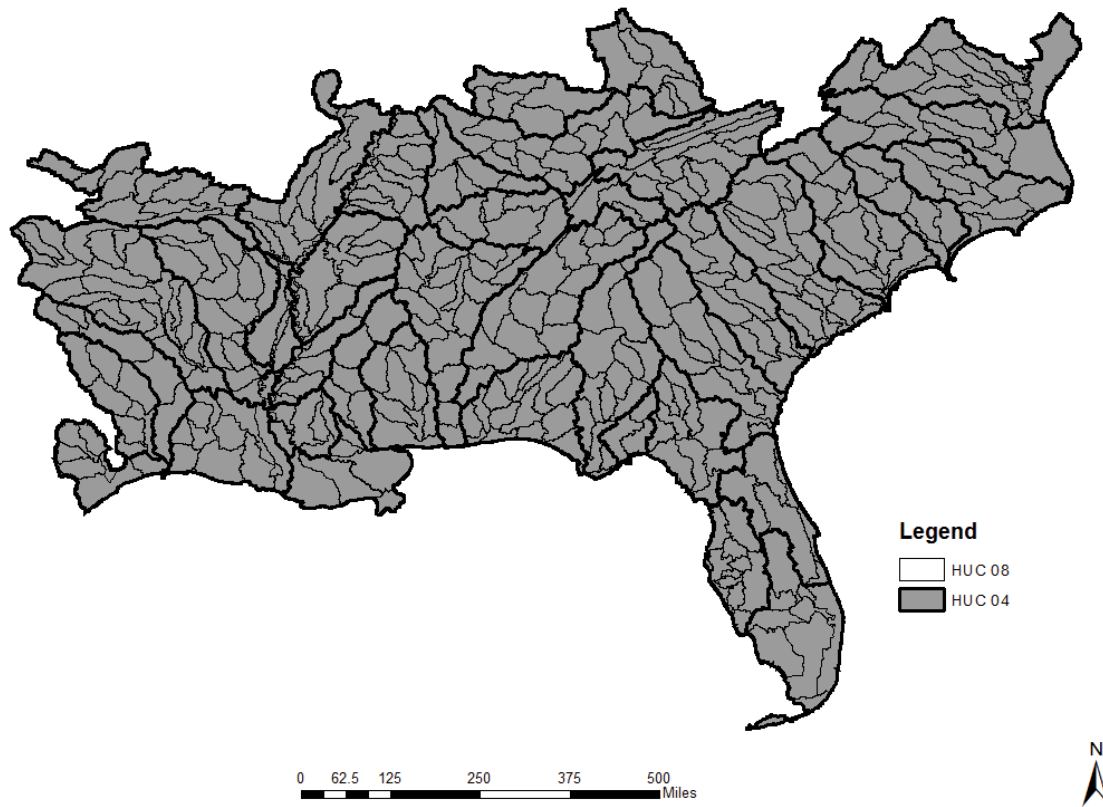


Figure 14. HUC08 basins nested within HUC04 watersheds

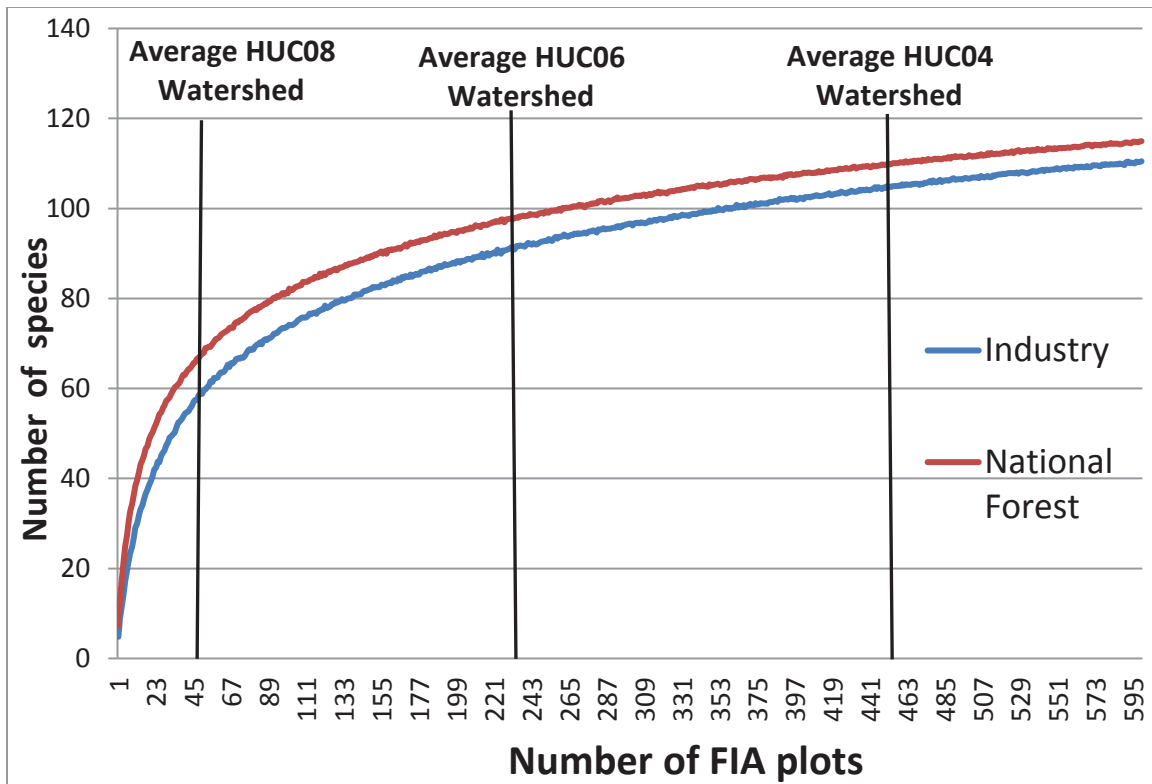


Figure 15. Species - number of plots curves for all plots on forest industry and national forest lands with lines denoting the average size of HIC04, HOC06, and HUC08 watersheds

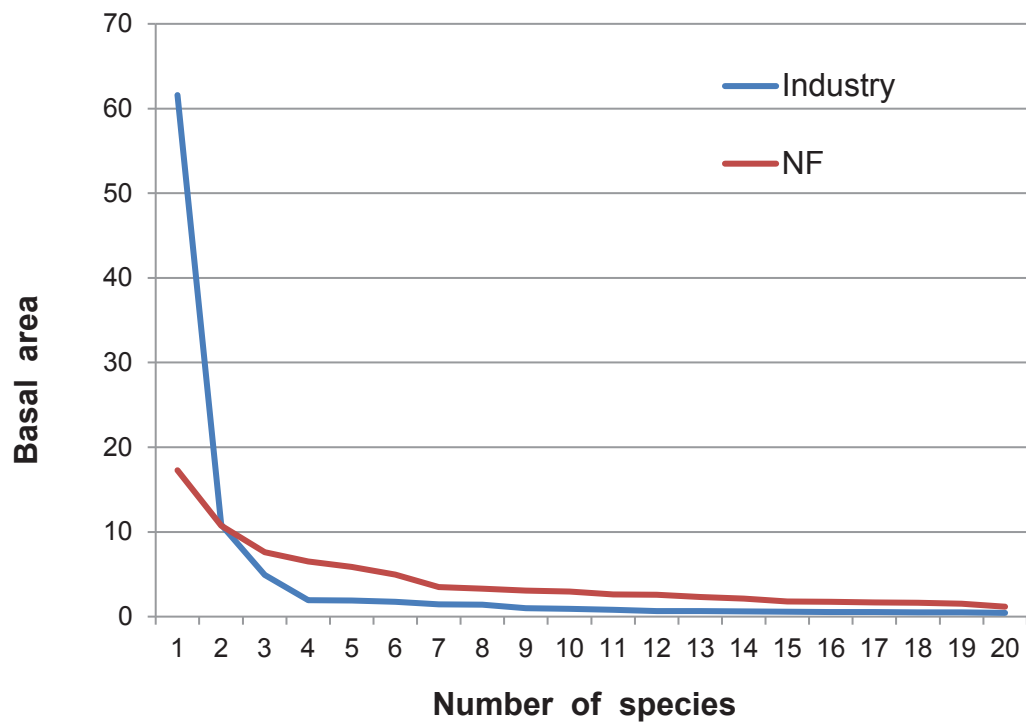
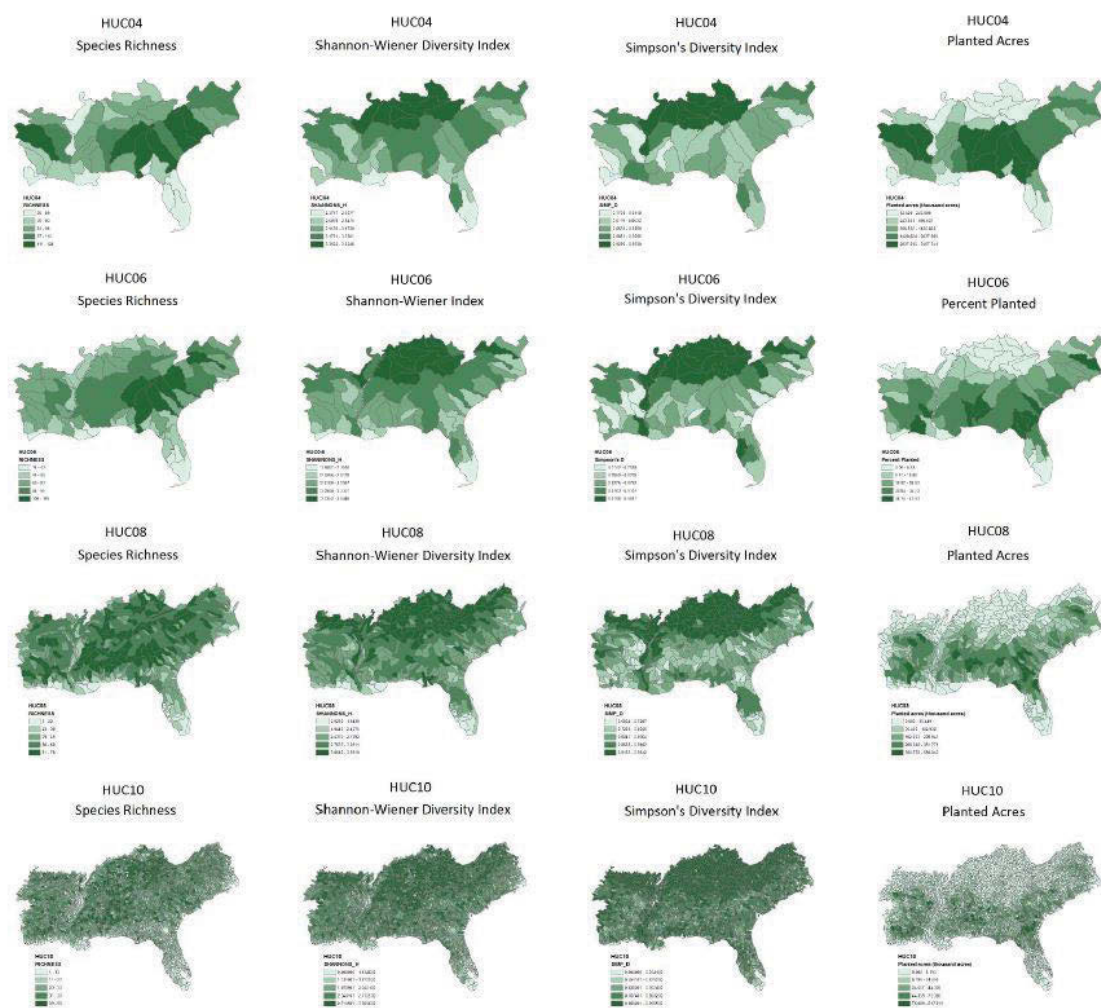


Figure 16. Species basal-area curves for plots that occur on either national forests or forest industry



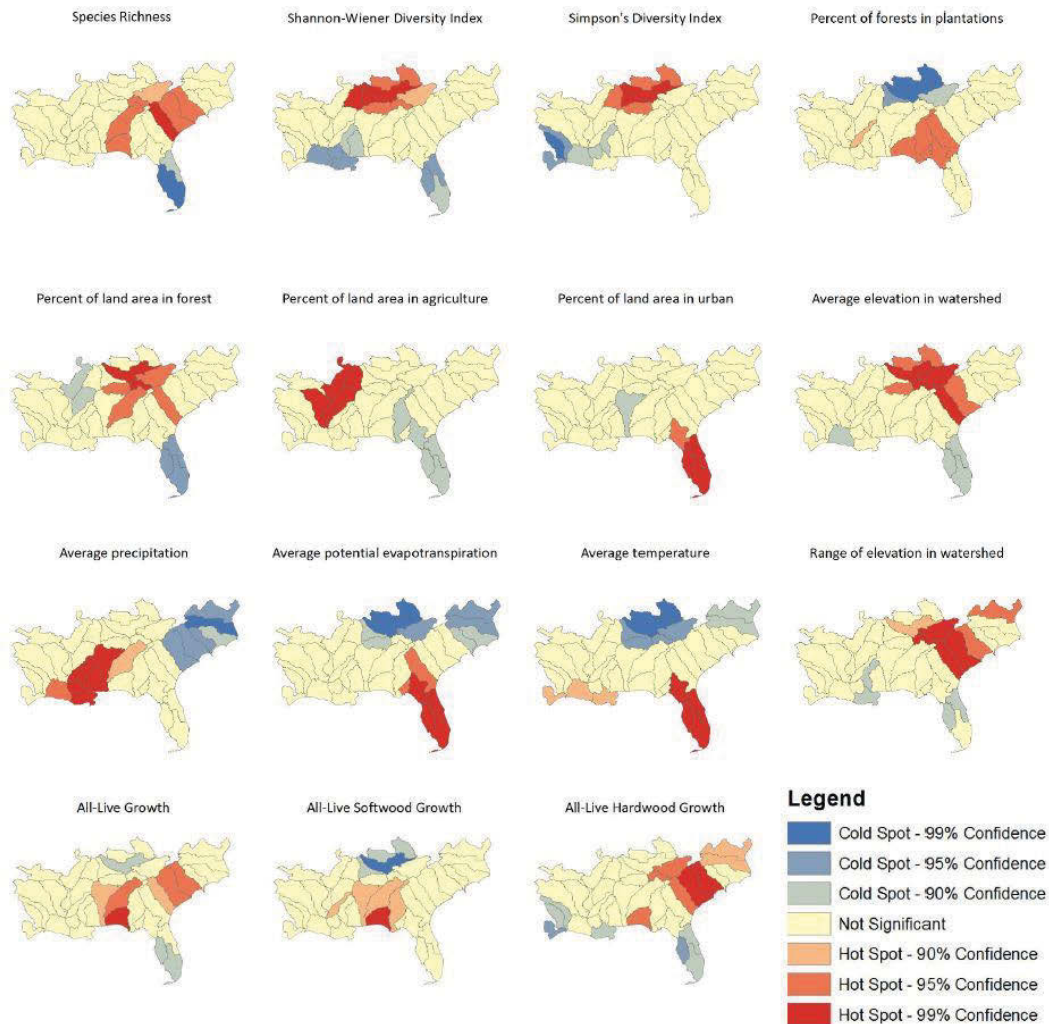


Figure 18. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HUC04 watersheds

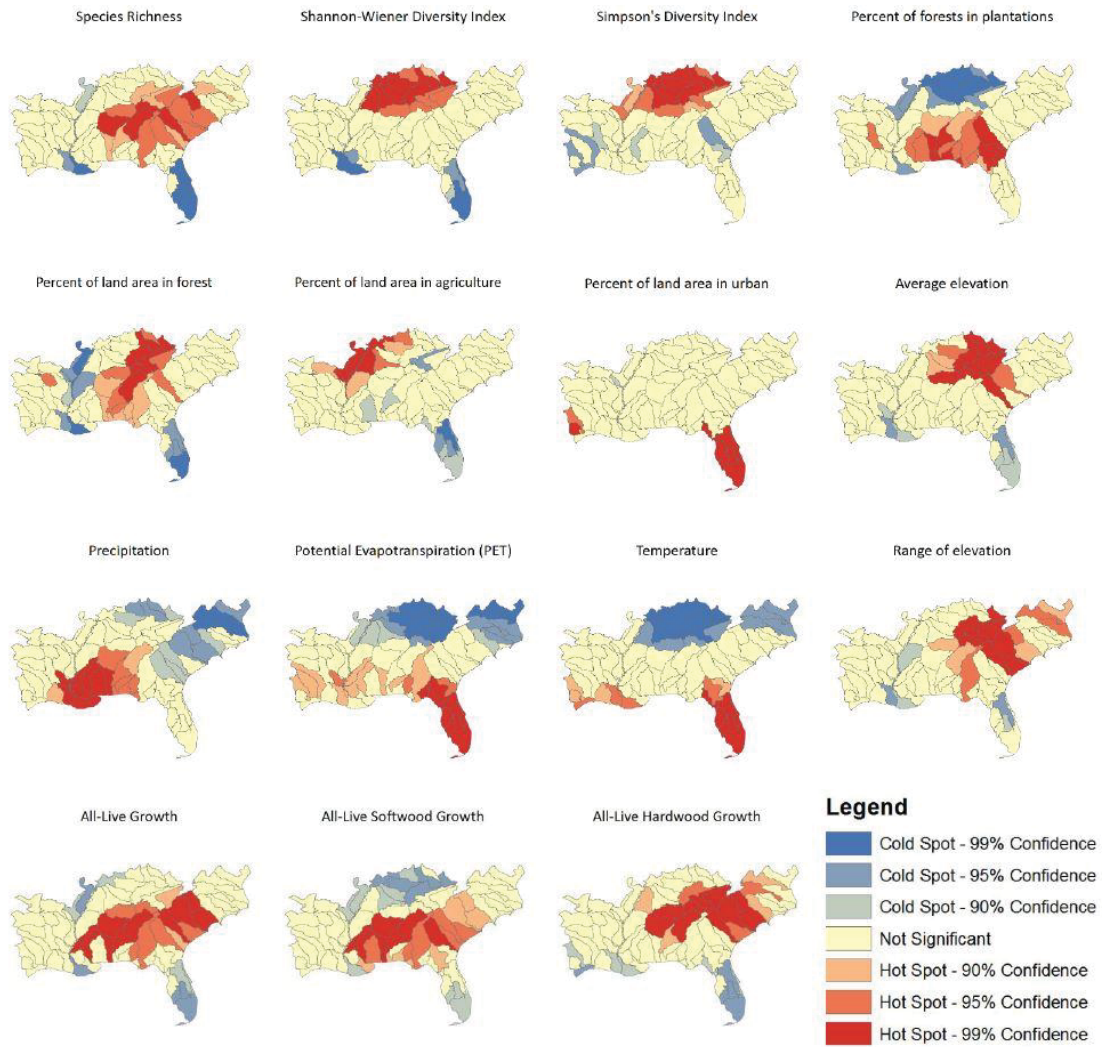


Figure 19. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HUC06 watersheds

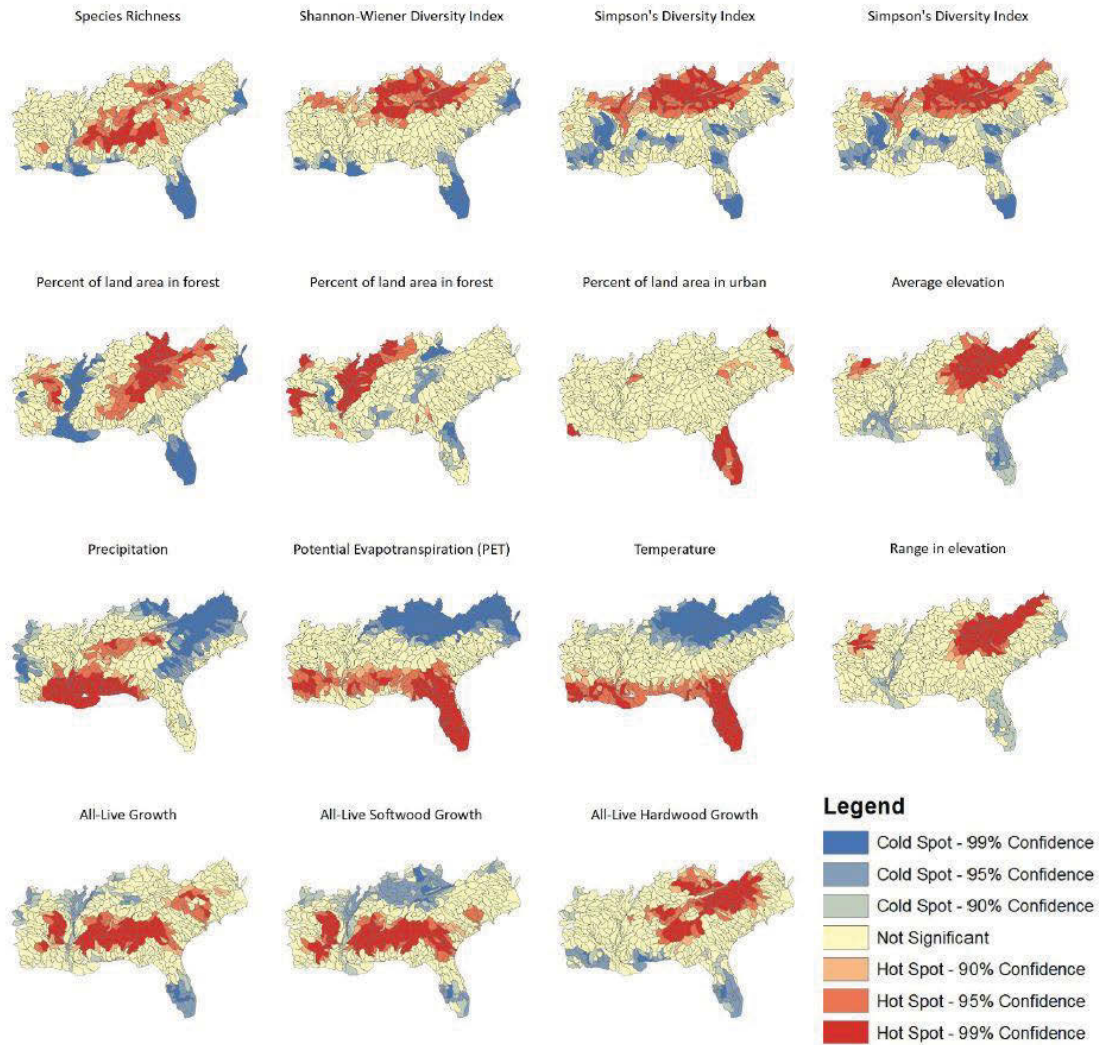


Figure 20. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HOC08 watersheds

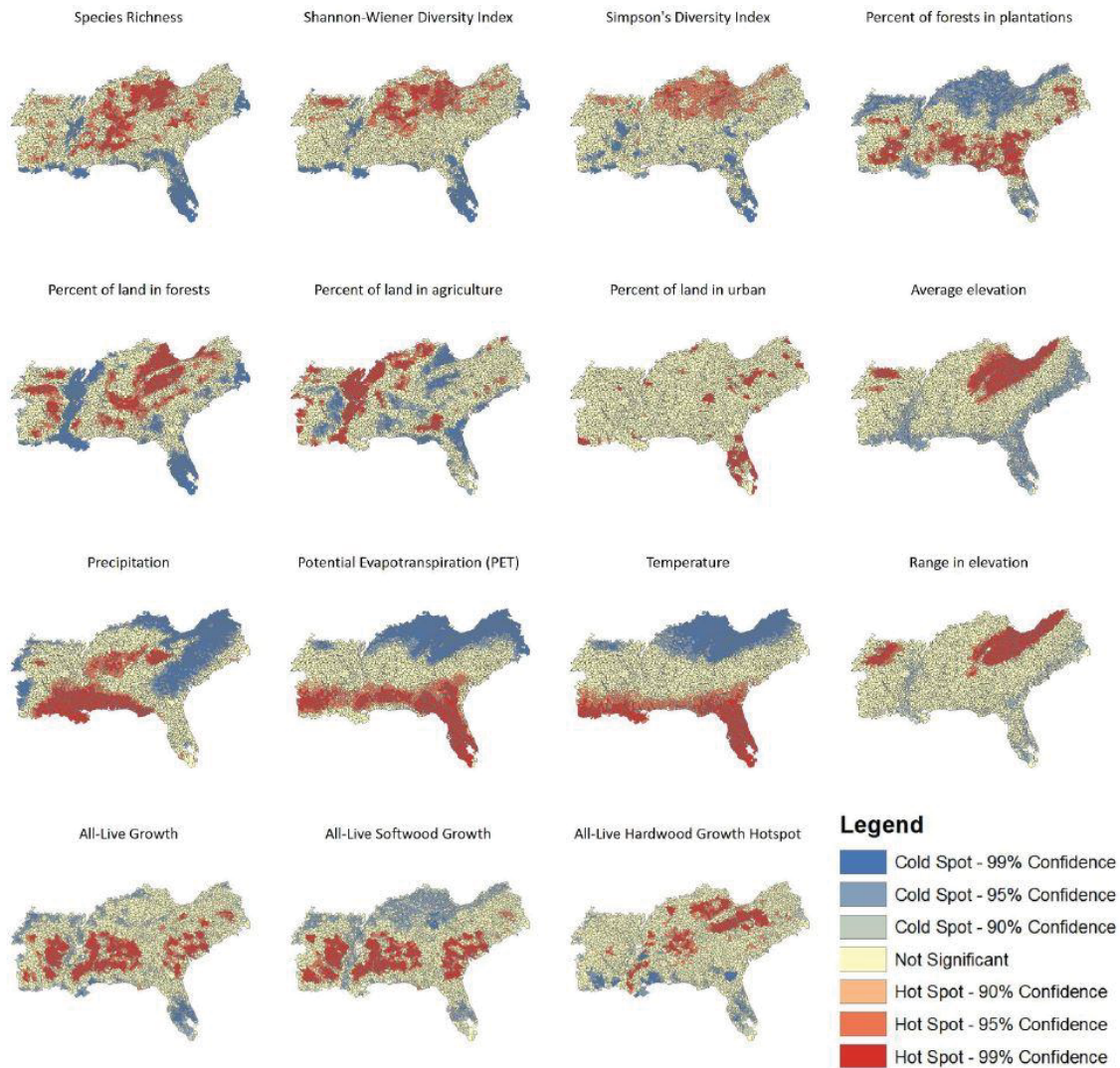
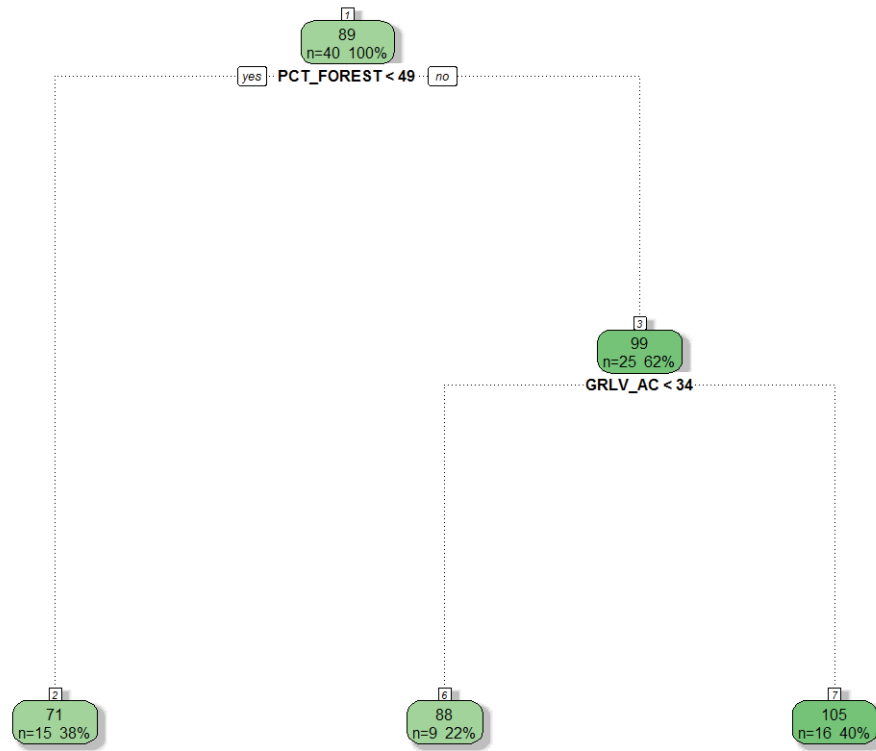


Figure 21. Gettis-Ord GI* Hot Spot maps of species richness, Shannon-Wiener index, Simpson's index, percent of forests in plantations, percent of land area in forests, percent of land in agriculture, percent of area in urban land uses, average elevation, average precipitation, average PET, average temperature, range in elevation, all-live growth, all-live softwood growth, and all-live hardwood growth for HUC10 watersheds

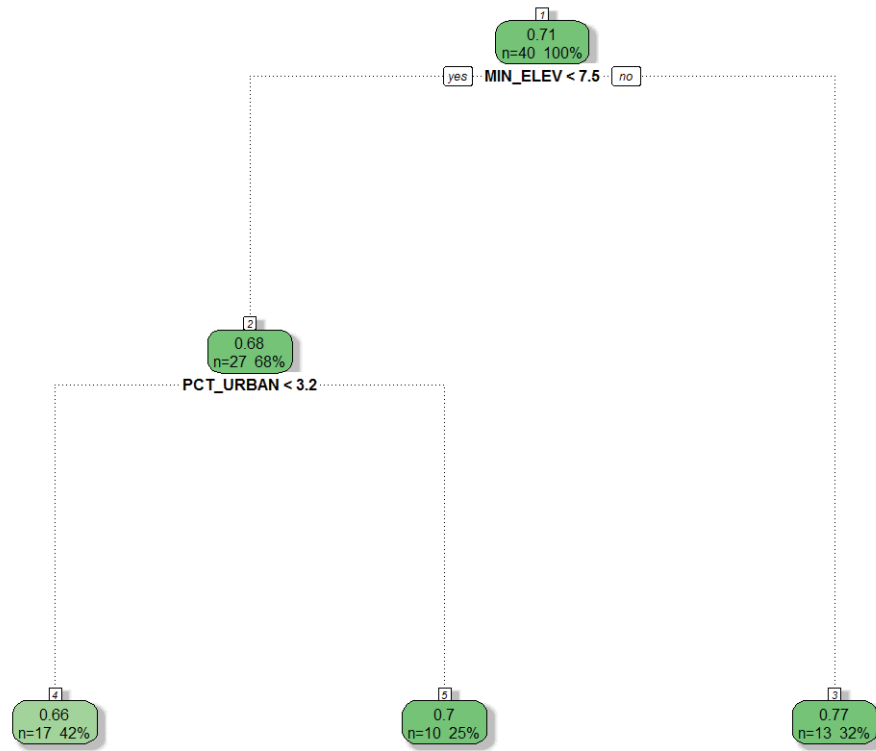
Classification Tree for HUC04 Richness



Rattle 2017-Aug-17 10:32:44 ahartsell

Figure 22. Regression tree for species richness, HUC04 watersheds

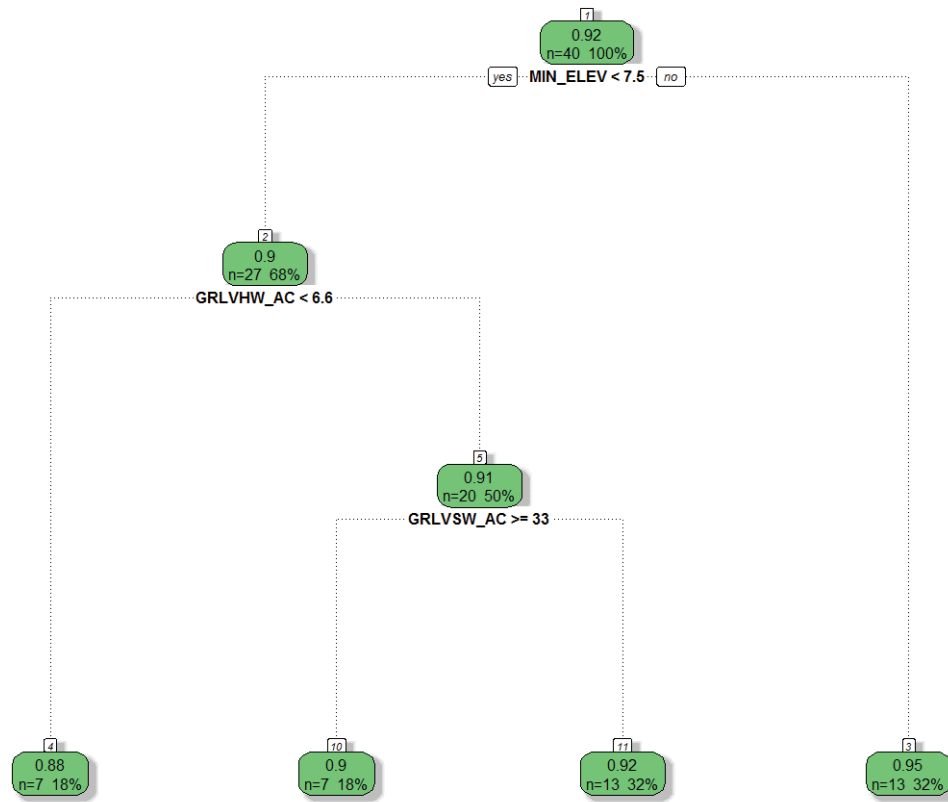
Classification Tree for HUC04 Shannons E



Rattle 2017-Aug-17 10:32:49 ahartsell

Figure 23. Regression tree for Shannon-Wiener diversity index, HUC04 watersheds

Classification Tree for HUC04 Simpsons



Rattle 2017-Aug-17 10:32:58 ahartsell

Figure 24. Regression tree for Simpson's diversity index, HUC04 watersheds

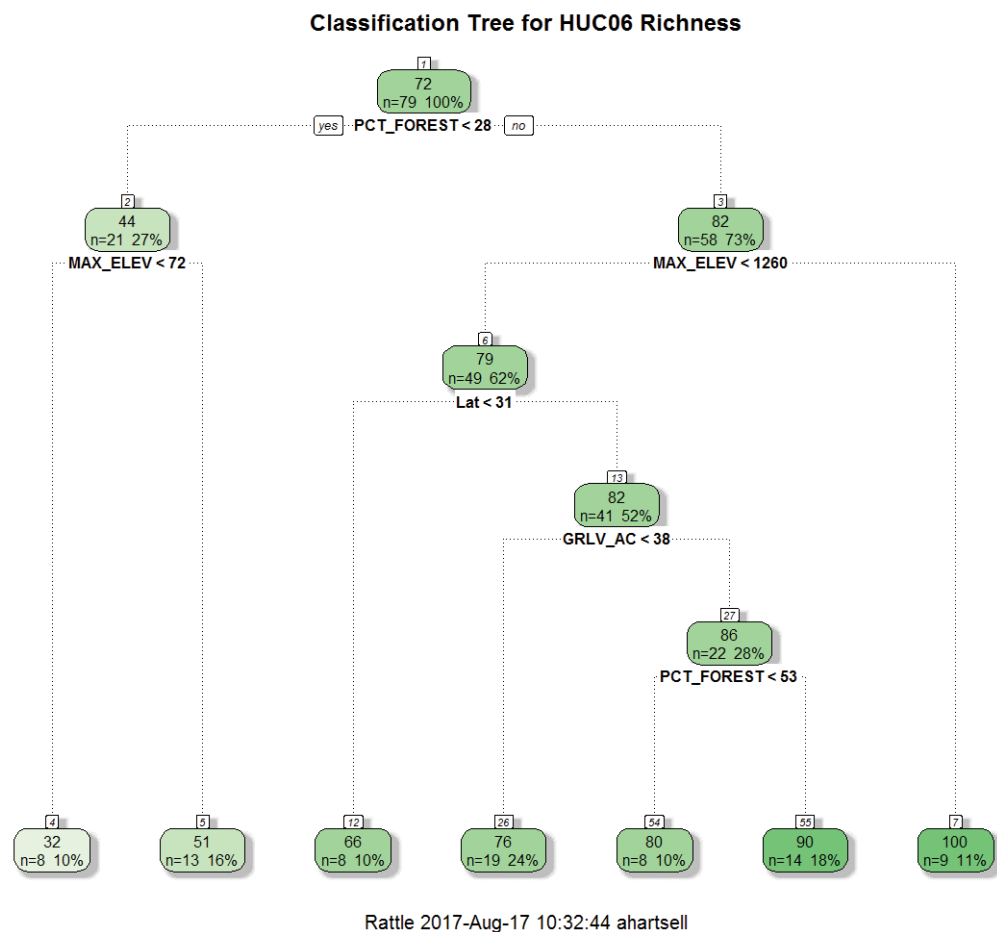


Figure 25. Regression tree for species richness, HUC06 watersheds

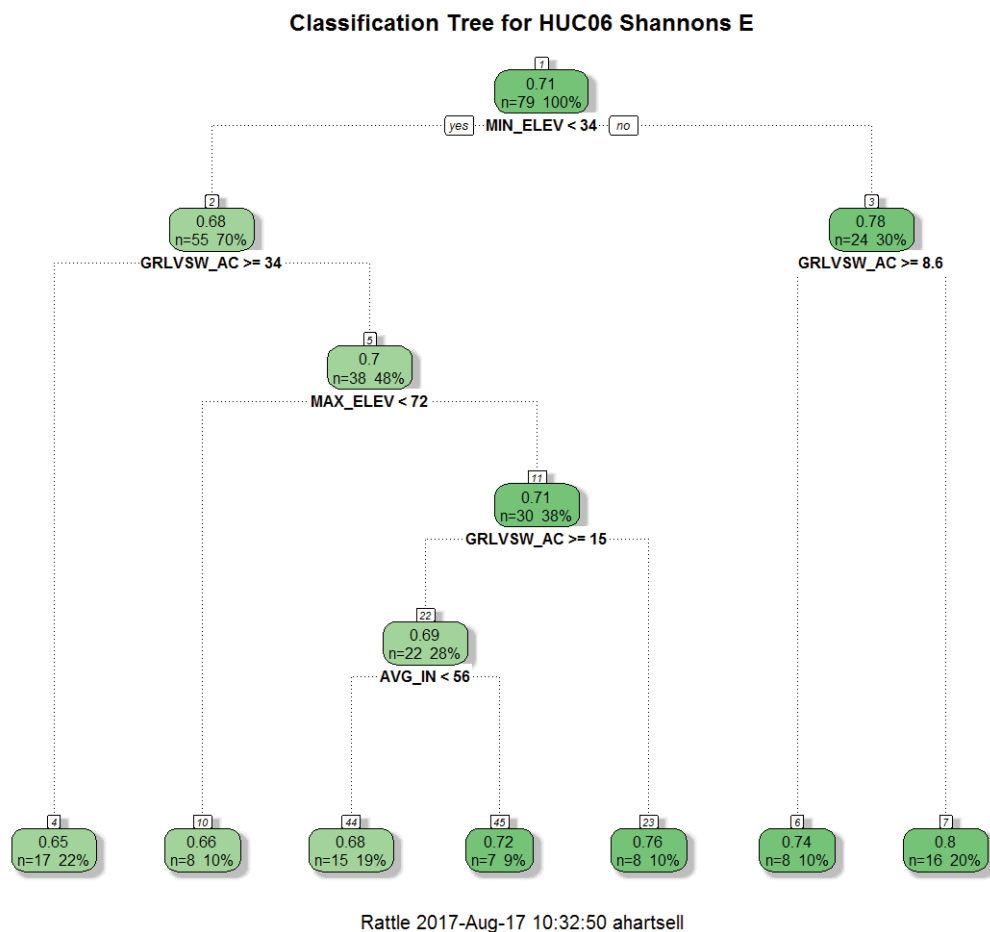


Figure 26. Regression tree for Shannon-Wiener diversity index, HUC06 watersheds

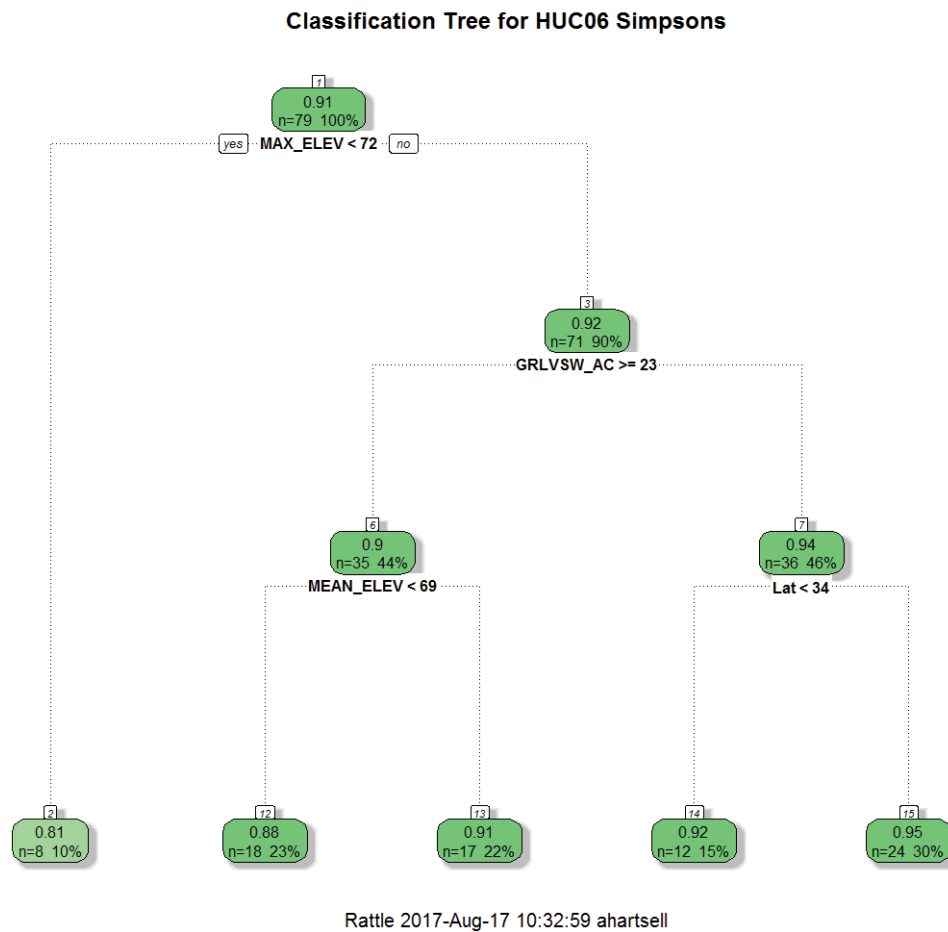


Figure 27. Regression tree for Simpson's diversity index, HUC06 watersheds

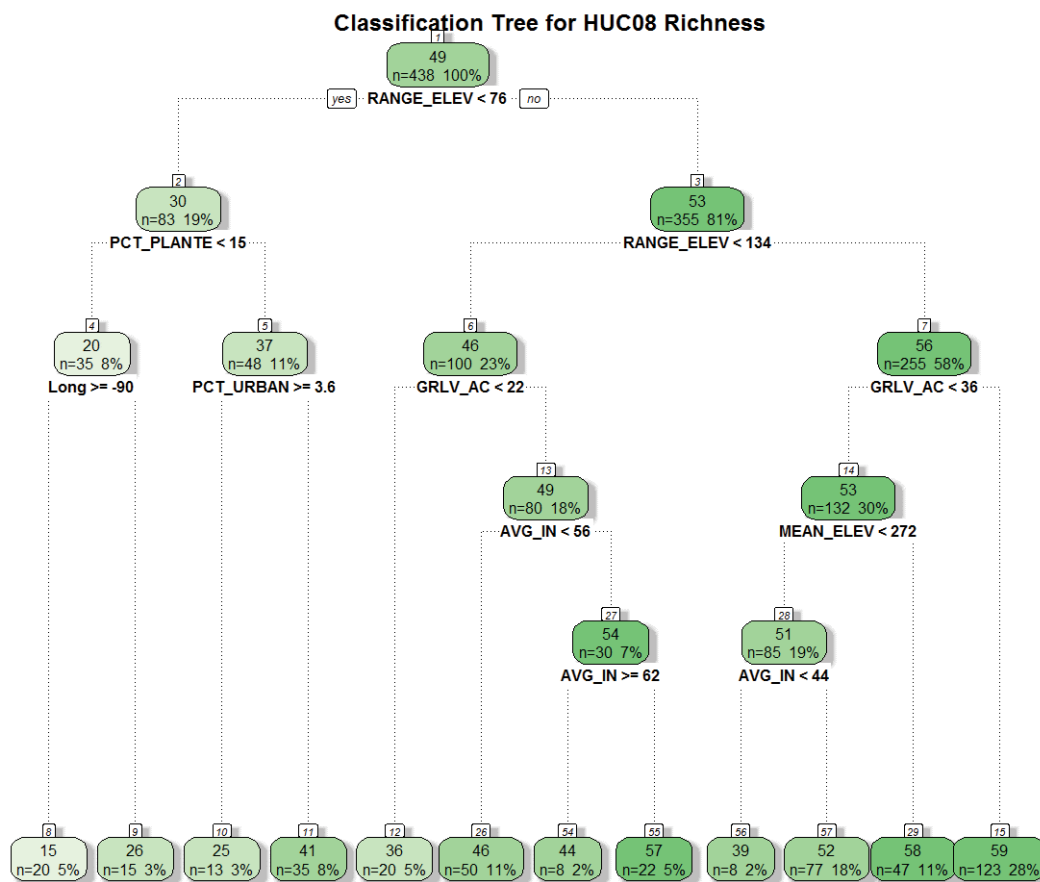


Figure 28. Regression tree for species richness, HUC08 watersheds

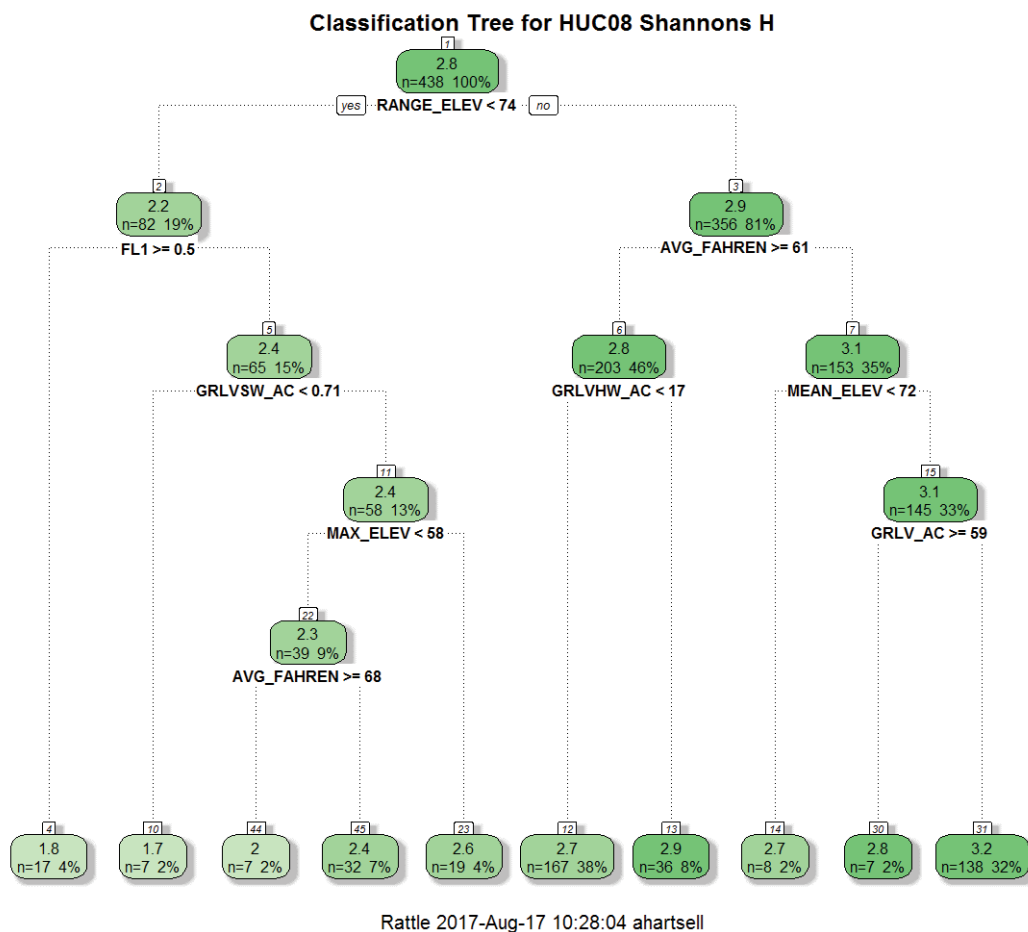
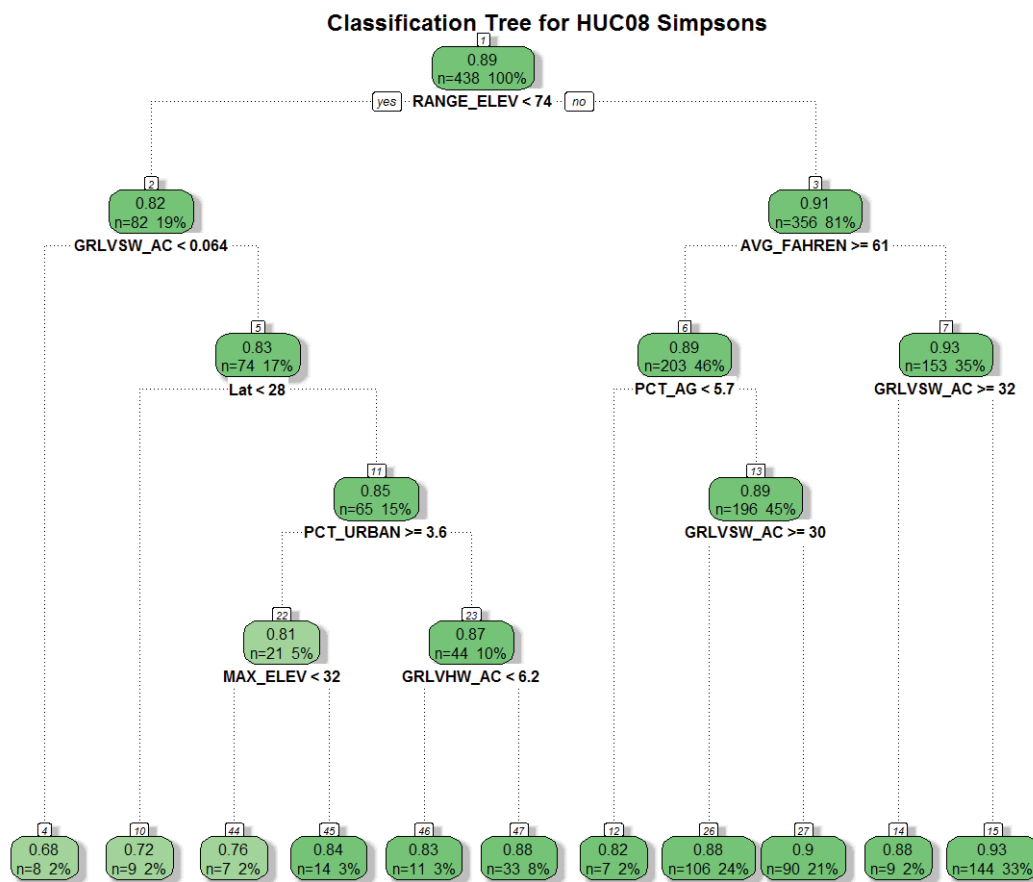


Figure 29. Regression tree for Shannon-Wiener diversity index, HUC08 watersheds



Rattle 2017-Aug-17 10:33:00 ahartsell

Figure 30. Regression tree for Simpson's diversity index, HUC08 watersheds

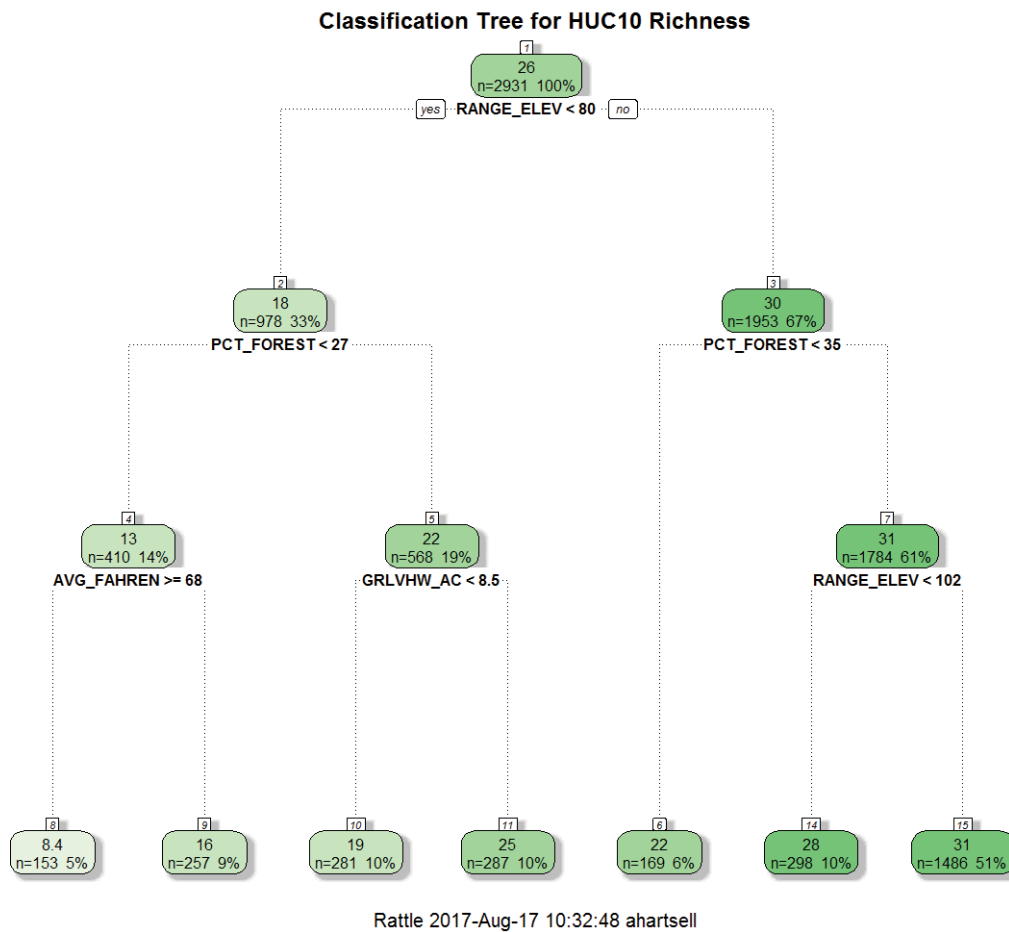


Figure 31. Regression tree for species richness, HUC10 watersheds

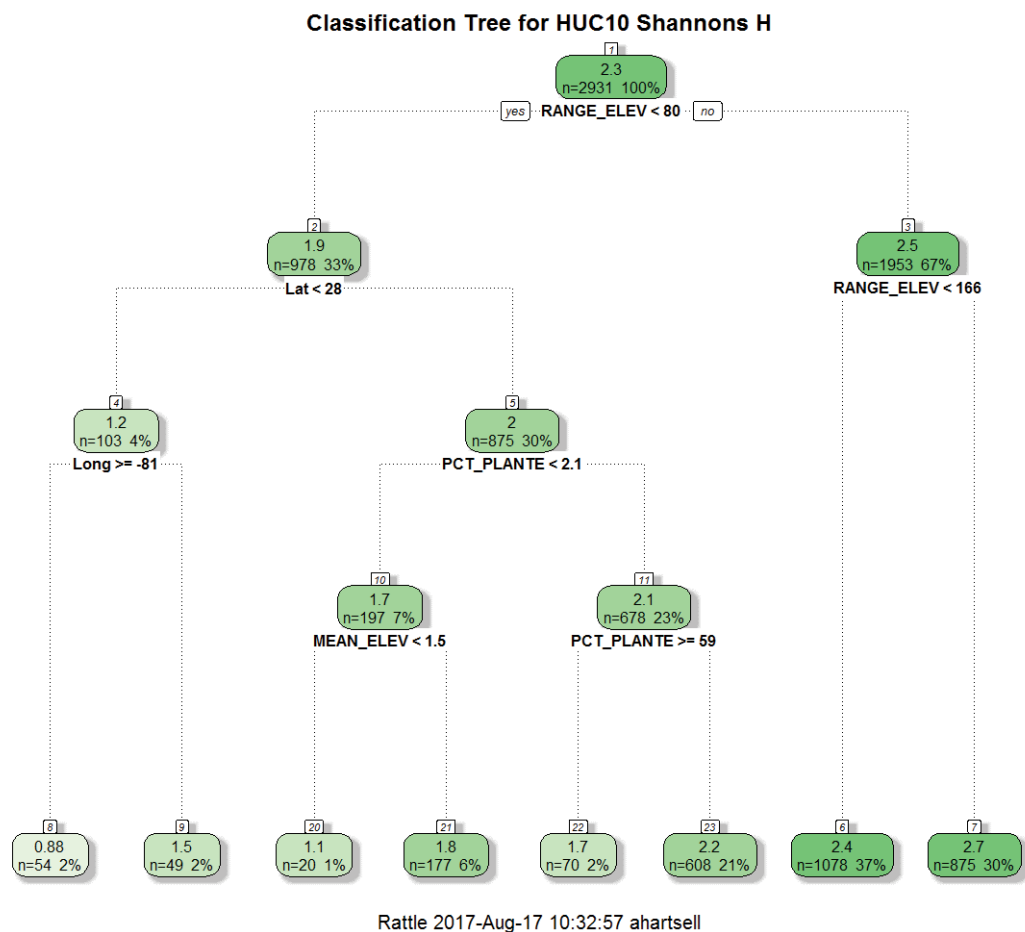


Figure 32. Regression tree for Shannon-Wiener diversity index, HUC10 watersheds

Table 13. Watershed sizes and the average, maximum, minimum, and total number of FIA plots in each HUC class

Unit	Sample size	Average (acres)	Plots			FIA Plots (n)
			Average	Max	Min	
HUC04	40	2,734,800	455.8	1000	76	18,232
HUC06	79	1,420,680	236.78	760	7	18,706
HUC08	430	271,800	45.3	143	4	19,478
HUC10	2828	41,100	6.85	38	1	19,371
HUC12	11554	11,400	1.9	28	1	19,579

Table 14. Variable list and description for all watersheds

Variable	Description
OBJECTID	Object ID
AreaAcres	area of polygon in acres
HUC##	Hydrologic Unit Code (HUC) number
TOTAL_LAND	Total land area (acres)
TOTAL_FORE	Total forest area (acres)
TOTAL_NONF	Total non-forest area (acres)
PLANTED_AC	Area of planted acres
SYP_PLANTE	Area of southern yellow pine plantations (acres)
PCT_SYP	Percent of forests in southern yellow pine plantations
RICHNESS	tree species richness
SHANNONS_E	Shannon-Wiener equitability index
SIMPSONS	Simpson's index
PCT_AG	Percent of landbase in agricultural uses (2011)
PCT_FOREST	Percent of landbase in forests (2011)
PCT_URBAN	Percent of landbase in urban (2011)
MAX_ELEV	Maximum elevation in watershed
MIN_ELEV	Minimum elevation in watershed
RANGE_ELEV	range in elevation in watershed
MEAN_ELEV	Average elevation in watershed
PET_IN	Average annual PET (mm/year)
AVG_IN	Average annual precipitation (inches/year)
AVG_FAHREN	Average annual temperature (Fahrenheit)
Long	Average longitude of FIA plots in watershed
Lat	Average latitude of FIA plots in watershed
PCT_PLANTE	Percent of the forests classified as planted

Table 14 Continued. Variable list and description for all watersheds

Variable	Description
FL1	Geographic (dummy) variable for southern half of Florida
GCP1	Geographic (dummy) variable for all watersheds that touch Gulf of Mexico
GCP2	Geographic (dummy) variable for all watersheds that touch Gulf of Mexico or Atlantic ocean
East_Flag	Geographic (dummy) variable for al watersheds on eastern side of Mississippi river
Delta	Geographic (dummy) variable for watersheds in Mississippi delta
GRLV	Total all-live growth of all-live trees (cubic feet)
GRLVSW	Total all-live growth of all-live softwood tree species (cubic feet)
GRLVHW	Total all-live growth of all-live hardwood tree species (cubic feet)
GRLV_AC	Average annual all-live growth of all-live tree species (cu.ft./year)
GRLVSW_AC	Average annual all-live growth of all-live softwood tree species (cu.ft./year)
GRLVHW_AC	Average annual all-live growth of all-live hardwood tree species (cu.ft./year)

Table 15. MRPP results of HUC04 watersheds

	Richness		Shannon's		Simpson's	
	<i>A</i>	<i>p</i>	<i>A</i>	<i>p</i>	<i>A</i>	<i>p</i>
HUC06	0.2374	0.0072	0.2607	0.0038	0.2065	0.0210
HUC08	0.1917	0.0000	0.2653	0.0000	0.2955	0.0000

Table 16.MRPP results of HUC06 watersheds

	Richness		Shannon's		Simpson's	
	<i>A</i>	<i>p</i>	<i>A</i>	<i>p</i>	<i>A</i>	<i>p</i>
HUC08	0.2484	0.0000	0.3227	0.0000	0.348	0.0000

Table 17. Two dimensional NMS ordination solution of HUC04 watersheds

Axis 1		Axis 2	
variable	score	variable	score
SYP_PLANT	-0.63	GCP2	-0.219
PLANTED_AC	-0.61	SYP_PLANT	-0.194
GRLV	-0.577	PCT_SYP	-0.194
GRLVHW	-0.512	PLANTED_AC	-0.182
RANGE_ELEV	-0.469	GRLVSW_AC	-0.163
MAX_ELEV	-0.439	PCT_PLANTED	-0.163
East_Fla	-0.217	GRLV	-0.117
PCT_SYP	-0.208	PCT_FORE	-0.099
GRLVSW_AC	-0.15	GRLV_AC	-0.089
PCT_FORE	-0.15	East_Fla	-0.049
RICHNESS	-0.149	RICHNESS	-0.032
MEAN_ELE	-0.147	AVG_IN	-0.01
PCT_PLANTED	-0.145	PET_IN	-0.007
FL1	-0.12	RANGE_ELEV	-0.005
GRLV_AC	-0.109	GRLVHW	-0.004
GRLVHW_AC	-0.042	Long	-0.001
Lat	-0.015	SIMP_D_T	0.0002
SIMP_D_T	0.0024	AVG_FAHR	0.0007
PET_IN	0.0067	MAX_ELEV	0.0007
Long	0.0121	Lat	0.0008
AVG_FAHR	0.0129	MEAN_ELE	0.0042
SHANNONS	0.028	SHANNONS	0.0121
AVG_IN	0.03	GRLVHW_AC	0.03
PCT_AG	0.1319	PCT_AG	0.0755
PCT_URBAN	0.1938	PCT_URBAN	0.148
GCP1	0.2124	Delta	0.3263
GCP2	0.3727	GCP1	0.4861
Delta	0.8966	FL1	1.058

Table 18. Two dimensional NMS ordination solution of HUC06 watersheds

Axis 1		Axis 2	
variable	score	variable	score
GRLVSW	-0.678	FL1	-0.258
SYP_PLAN	-0.668	MEAN_ELEVATION	-0.102
PLANTED_ACRES	-0.654	TOTAL_NONFOREST	-0.094
GRLV	-0.644	Delta	-0.089
GRLVHW	-0.582	MAX_ELEVATION	-0.088
TOTAL_FOREST	-0.571	PCT_URBAN	-0.082
TOTAL_LAND	-0.486	RANGE_ELEVATION	-0.08
RANGE_ELAVATION	-0.477	PCT_AG	-0.07
MAX_ELEVATION	-0.453	GRLVHW_ACRE	-0.045
TOTAL_NON_FOREST	-0.387	GRLVHW	-0.026
PCT_SYP	-0.362	TOTAL_LAND	-0.021
GRLVSW_ACRE	-0.34	SHANNONS	-0.007
MEAN_ELEVATION	-0.336	East_Flag	-0.004
PCT_FOREST	-0.331	Lat	-0.004
PCT_PLANTED	-0.32	SIMP_D	-0.001
RICHNESS	-0.245	Long	-1E-03
GRLV_ACRE	-0.204	AVG_TEMP_F	0.0041
East_Flag	-0.19	AVG_PPT_IN	0.0064
GCP2	-0.157	PET_IN	0.0067
SHANNONS	-0.047	RICHNESS	0.0117
Lat	-0.013	PCT_FOREST	0.031
SIMP_D	-0.012	TOTAL_FOREST	0.0318
GRLVHW_ACRE	-0.001	GRLV_ACRE	0.0669
PET_IN	0.0006	GRLV	0.0789
Long	0.0102	GCP1	0.1289
AVG_TEMP_F	0.0153	GRLVSW	0.1388
AVG_PPT_IN	0.021	PLANTED_ACRES	0.145
PCT_AG	0.0639	GRLVSW_ACRE	0.1468
PCT_URBAN	0.1681	SYP_PLANTED	0.1536
GCP1	0.2957	PCT_PLANTED	0.1544
Delta	0.5601	PCT_SYP	0.1762
FL1	0.6822	GCP2	0.2345

Table 19. Three dimensional NMS ordination solution of HUC08 watersheds

Axis 1		Axis 2		Axis 3	
variable	score	variable	score	variable	score
FL1	-1.045	GCP2	-0.431	FL1	-0.565
Delta	-0.606	SYP_PLANTED	-0.402	Delta	-0.543
GCP2	-0.305	PLANTED_AC	-0.389	TOTAL_NONF	-0.268
PCT_URBAN	-0.29	GRLVSW	-0.344	PCT_URBAN	-0.173
PCT_AG	-0.063	PCT_SYP	-0.317	PCT_AG	-0.15
AVG_FAHREN	-0.021	PCT_PLANTE	-0.3	TOTAL_LAND	-0.087
AVG_IN	-0.009	GRLVSW_AC	-0.262	GRLVHW_AC	-0.023
PET_IN	-0.008	GRLV	-0.202	SIMPSONS	-0.013
Long	0.0002	FL1	-0.197	SHANNONS	-0.01
SHANNONS	0.0023	TOTAL_FORE	-0.118	Long	-0.004
SIMPSONS	0.0086	Delta	-0.112	AVG_FAHREN	-0.003
Lat	0.0157	TOTAL_LAND	-0.105	PET_IN	0.0016
PCT_PLANTE	0.0406	GRLV_AC	-0.087	Lat	0.0023
PCT_SYP	0.0528	TOTAL_NONF	-0.078	SHANNONS	0.0028
GRLVHW_AC	0.0586	East_Flag	-0.031	AVG_IN	0.0061
SHANNONS	0.0644	PET_IN	-0.028	GRLVHW	0.0215
GRLVSW_AC	0.0662	AVG_FAHREN	-0.027	East_Flag	0.0263
GRLV_AC	0.0675	AVG_IN	-0.014	RICHNESS	0.0316
East_Flag	0.1242	Long	-0.002	MEAN_ELEV	0.0343
RICHNESS	0.1755	RICHNESS	0.0001	TOTAL_FORE	0.0444
PCT_FOREST	0.2058	SIMPSONS	0.0129	MAX_ELEV	0.0456
TOTAL_NONF	0.238	Lat	0.0252	RANGE_ELEV	0.0557
TOTAL_LAND	0.3149	SHANNONS	0.0294	GRLV_AC	0.1047
PLANTED_AC	0.3525	PCT_URBAN	0.0298	PCT_FOREST	0.1148
SYP_PLANTED	0.3579	PCT_AG	0.0311	GRLV	0.1233
GRLVSW	0.3688	SHANNONS	0.0312	GCP2	0.1339
TOTAL_FORE	0.3792	GRLVHW	0.0377	PLANTED_AC	0.1798
GRLV	0.4046	PCT_FOREST	0.0843	GRLVSW	0.1834
MEAN_ELEV	0.4191	GRLVHW_AC	0.1558	PCT_PLANTE	0.1886
GRLVHW	0.4603	RANGE_ELEV	0.4394	GRLVSW_AC	0.1954
MAX_ELEV	0.4737	MEAN_ELEV	0.4448	SYP_PLANTED	0.1985
RANGE_ELEV	0.4873	MAX_ELEV	0.4546	PCT_SYP	0.223

Table 20. Three dimensional NMS ordination solution of HUC10 watersheds

Axis 1		Axis 2		Axis 3	
variable	score	variable	score	variable	score
RANGE_ELEV	-0.48711	MAX_ELEV	-0.4548	FL1	-0.56653
MAX_ELEV	-0.47352	MEAN_ELEV	-0.44495	Delta	-0.54389
GRLVHW	-0.46027	MIN_ELEV	-0.4396	TOTAL_NONF	-0.26759
MEAN_ELEV	-0.41885	GRLVHW_A	-0.15612	PCT_URBAN	-0.17212
GRLV	-0.40471	PCT_FOREST	-0.08423	PCT_AG	-0.1498
TOTAL_FOREST	-0.37923	GRLVHW	-0.03785	TOTAL_LAND	-0.08739
GRLVSW	-0.36904	PCT_AG	-0.0312	GRLVHW_AC	-0.02325
SYP_PLANT	-0.35809	PCT_URBAN	-0.02997	SIMPSONS	-0.01259
PLANTED_AC	-0.35268	SIMPSONS	-0.02945	SHANNONS	-0.00954
TOTAL_LAND	-0.31492	SHANNONS	-0.02936	Long	-0.00355
TOTAL_NONF	-0.2379	Lat	-0.02524	AVG_FAHREN	-0.0034
PCT_FOREST	-0.20575	RICHNESS	-0.00016	PET_IN	0.00159
RICHNESS	-0.17544	Long	0.00164	Lat	0.00236
East_Flag	-0.12427	AVG_IN	0.01435	AVG_IN	0.00603
GRLV_AC	-0.06759	AVG_FAHREN	0.02749	GRLVHW	0.02134
GRLVSW_AC	-0.06635	PET_IN	0.02761	East_Flag	0.02634
GRLVHW_AC	-0.05863	East_Flag	0.03092	RICHNESS	0.03163
PCT_SYP	-0.05302	TOTAL_NONF	0.07798	MEAN_ELEV	0.03438
PCT_PLANT	-0.04084	GRLV_AC	0.0865	TOTAL_FORE	0.04421
Lat	-0.01569	TOTAL_LAND	0.10454	MAX_ELEV	0.04566
SIMPSONS	-0.00854	Delta	0.11165	RANGE_ELEV	0.05576
SHANNONS	-0.00233	TOTAL_FORE	0.11801	GRLV_AC	0.10463
Long	-0.0002	FL1	0.19893	PCT_FOREST	0.11481
PET_IN	0.00778	GRLV	0.20212	GRLV	0.12316
AVG_IN	0.00938	GRLVSW_AC	0.26161	GCP2	0.13388
AVG_FAHREN	0.02118	PCT_PLANTE	0.3005	PLANTED_AC	0.17964
PCT_AG	0.06279	PCT_SYP	0.31681	GRLVSW	0.18315
PCT_URBAN	0.29048	GRLVSW	0.34393	PCT_PLANTE	0.18859
GCP2	0.30444	PLANTED_AC	0.389	GRLVSW_AC	0.19541
Delta	0.60539	SYP_PLANTE	0.40181	SYP_PLANTE	0.19832
FL1	1.04675	GCP2	0.43142	PCT_SYP	0.22294

Table 21. Two dimensional NMS ordination solution of HUC08 watersheds categorized as being 'cold spots' for Simpson's diversity index

Axis 1		Axis 2	
variable	score	variable	score
Delta	-1.426	AVG_TEMP (F)	0.019
FL1	-0.926	AVG_PPT_IN	-0.006
PCT_URBAN	-0.541	Delta	0.2349
GCP2	-0.274	East_Fla	-0.134
PCT_AG	-0.048	FL1	0.9005
AVG_TEMP (F)	-0.021	GCP2	0.0075
PET_IN	-0.005	GRLV	-0.107
AVG_IN	0.0007	GRLV_ACRE	-0.151
Long	0.0092	GRLVHW	-0.093
GRLVHW_A	0.0132	GRLVHW_A	-0.091
Lat	0.0241	GRLVSW	-0.109
GRLV_ACRE	0.1436	GRLVSW_ACRE	-0.165
East_Fla	0.1537	Lat	-0.024
GRLVSW_ACRE	0.169	Long	-0.007
PCT_PLANTED	0.1877	MAX_ELEVATION	-0.108
PCT_SYP	0.2073	MEAN_ELEVATION	-0.128
PCT_FOREST	0.2116	MIN_ELEVATION	-0.166
TOTAL_NON_FOREST	0.2187	PCT_AG	0.0348
MEAN_ELEVATION	0.3335	PCT_FOREST	-0.142
RANGE_ELAVATION	0.358	PCT_PLANTED	-0.151
MAX_ELEVATION	0.3641	PCT_SYP	-0.163
TOTAL_LAND_AREA	0.3765	PCT_URBAN	0.4421
MIN_ELEVATION	0.3853	PET_IN	0.0114
GRLVHW	0.4522	PLANTED_ACRES	-0.08
TOTAL_FOREST_AREA	0.459	RANGE_ELAVATION	-0.099
GRLV	0.4633	SYP_PLANTED	-0.083
GRLVSW	0.4673	TOTAL_FOREST_AREA	-0.014
SYP_PLANTED	0.507	TOTAL_LAND_AREA	0.1369
PLANTED_ACRES	0.5083	TOTAL_NON_FOREST	0.4313

Table 22. Regression models for species richness, HUC04 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	47.8345	2.9967	15.9624	0.0000
GRLV_AC	0.4683	0.0813	5.7578	0.0000
TOTAL_FOREST	0.0034	0.0005	7.1538	0.0000
RANGE_ELEV	0.0117	0.0023	5.0595	0.0001
n, DF		4, 36		
r-square		0.8955		
AIC		263.744		
Jarque-Bera				0.9264
Breusch-Pagan				0.4997
Moran's I (error)				0.3095

Table 23. Regression models for Shannon-Weiner diversity index, HUC04 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.6614	0.0193	34.3144	0.0000
PCT_AG	0.0018	0.0004	4.1804	0.0002
MEAN_ELEV	0.0003	0.0000	5.3105	0.0001
TOTAL_FOREST	0.0001	0.0000	-3.3315	0.0020
n, DF		4, 36		
r-square		0.6263		
AIC		-146.8220		
Jarque-Bera				0.2181
Breusch-Pagan				0.0927
Moran's I (error)				0.9359

Table 24. Regression models for Simpson's diversity, HUC04 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.8812	0.0214	41.1483	0.00000
PCT_AG	0.0011	0.0004	2.8360	0.00745
MIN_ELEV	0.0003	0.0001	2.3256	0.02579
PCT_PLANTED	-0.0018	0.0006	-2.8529	0.00714
n, DF		4, 36		
r-square		0.6084		
AIC		-148.9400		
Jarque-Bera				0.1592
Breusch-Pagan				0.2462
Moran's I (error)				0.1511

Table 25. Regression models for species richness, HUC06 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	80.3046	17.7382	4.5272	0.0000
GRLV_AC	0.4096	0.0789	5.1896	0.0000
TOTAL_FOREST_AREA	0.0076	0.0009	8.9323	0.0000
RANGE_ELEV	0.0071	0.0033	2.1631	0.0338
PET_IN	-0.8411	0.3200	-2.6280	0.0104
n, DF		4, 75		
r-square		0.8338		
AIC		-284.24		
Jarque-Bera				0.0524
Breusch-Pagan				0.2395
Moran's I (error)				0.4196

Table 26. Regression models for Shannon-Weiner diversity index, HUC06 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.689077	0.0175	39.3706	0.00000
PCT_AG	0.00141	0.0004	3.9233	0.00002
MIN_ELEV	0.00044	0.0001	3.7278	0.00004
Percent Planted	-0.00144	0.0005	-3.08496	0.00285
n, DF		4, 75		
r-square		0.5326		
AIC		-254.222		
Jarque-Bera				0.6914
Breusch-Pagan				0.0511
Moran's I (error)				0.9335

Table 27. Regression models for Simpson's diversity, HUC06 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.8788	0.0146	60.2360	0.0000
PCT_PLANTE	-0.00198	0.0004	-4.8304	0.0000
GRLVHW_AC	0.0016	0.0006	2.5380	0.0132
MIN_ELEV	0.0003	0.0001	2.4594	0.0162
n, DF		4, 75		
r-square		0.5708		
AIC		-271.376		
Jarque-Bera				0.0814
Breusch-Pagan				0.5960
Moran's I (error)				0.4741

Table 28. Regression models for species richness, HUC08 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	55.5168	7.7127	7.1982	0.0000
PCT_AG	0.2098	0.0268	7.8437	0.0000
PCT_FOREST	0.1909	0.0262	7.2860	0.0000
TOTAL_FOREST_AREA	0.0285	0.0015	19.3629	0.0000
GRLV_AC	0.1130	0.0197	5.7500	0.0000
AVG_PPT_INCHES	0.1465	0.0714	2.0519	0.0408
AVG_TEMP_F	-0.7533	0.1166	-6.4610	0.0000
n, DF		7, 431		
r-square		0.7334		
AIC		3031.63		
Moran's I (error)				0.00000

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	60.2532	9.9239	6.0715	0.0000
PCT_AG	0.1801	0.0310	5.8185	0.0000
PCT_FOREST	0.1911	0.0298	6.4070	0.0000
TOTAL_FOREST_AREA	0.0287	0.0015	19.5977	0.0000
GRLV_AC	0.0788	0.0224	3.5107	0.0004
AVG_TEMP_F	-0.6788	0.1385	-4.9023	0.0000
LAMBDA	0.3318	0.0643	5.1587	0.0000
n, DF		7, 431		
r-square		0.7491		
AIC		3012.78		

Table 29. Regression models for Shannon-Weiner diversity index, HUC08 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.7024	0.0100	70.5821	0.0000
PCT_FOREST	0.0009	0.0001	7.6774	0.0000
PCT_AG	0.0011	0.0002	6.8233	0.0000
GRLVSW_AC	-0.0018	0.0002	-7.2744	0.0112
PCT_PLANTED	-0.0007	0.0003	-2.5477	0.0000
n, DF		5, 433		
r-square		0.4495		
AIC		-1342.34		
Moran's I (error)				0.00000

Variable	Coefficient	Std.Error	t-Statistic	Probability
W_SHANNONS_E	0.5214	0.0503	10.3564	0.0000
CONSTANT	0.3297	0.0361	9.1387	0.0000
PCT_FOREST	0.0004	0.0001	3.7866	0.0001
PCT_AG	0.0007	0.0002	4.4493	0.0000
PCT_PLANTED	-0.0012	0.0002	-6.7882	0.0000
n, DF		5, 433		
r-square		0.5208		
AIC		-1376.61		

Table 30. Regression models for Simpson's diversity, HUC08 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.8326	0.0109	76.1354	0.0000
PCT_PLANTED	-0.0024	0.0002	-10.8882	0.0000
PCT_AG	0.0013	0.0002	7.4231	0.0000
MEAN_ELEV	0.0001	0.0000	3.5594	0.0004
PCT_FOREST	0.0004	0.0002	2.3668	0.0184
n, DF		5, 433		
r-square		0.5012		
AIC		-1258.63		
Moran's I (error)				0.0000

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.8425	0.0152	55.4385	0.0000
PCT_PLANTED	-0.0030	0.0002	-12.3414	0.0000
PCT_AG	0.0012	0.0002	5.2979	0.0000
PCT_FOREST	0.0007	0.0002	3.7676	0.0002
LAMBDA	0.6261	0.0467	13.4169	0.0000
n, DF		5, 433		
r-square		0.6416		
AIC		-1364.81		

Table 31. Regression models for species richness, HUC10 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	13.6942	0.4014	34.1200	0.0000
PCT_FOREST	0.1223	0.0064	19.2065	0.0000
TOTAL_FOREST_AREA	0.1181	0.0031	37.6159	0.0000
PCT_PLANTED	-0.0961	0.0062	-15.4614	0.0000
FL_FLAG	-8.9472	0.5510	-16.2392	0.0000
Delta	-1.5024	0.5372	-2.7966	0.0052
n, DF		6, 2925		
r-square		0.5800		
AIC		19367.6		
Moran's I (error)				0.0000

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	14.4503	0.5337	27.0763	0.0000
PCT_FOREST	0.0805	0.0083	9.7531	0.0000
TOTAL_FOREST_AREA	0.1296	0.0028	45.4976	0.0000
PCT_PLANTED	-0.0892	0.0071	-12.5702	0.0000
FL_FLAG	-5.0938	0.7743	-6.5782	0.0000
Delta	-3.8887	0.7205	-5.3970	0.0000
LAMBDA	0.6008	0.0197	30.5527	0.0000
n, DF		7, 2924		
r-square		0.7005		
AIC		18608.00		

Table 32. Regression models for Shannon-Weiner diversity index, HUC10 watershed

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	-0.5239	0.1055	-4.9674	0.0000
MEAN_ELEV	0.0001	0.0000	10.1955	0.0000
PCT_PLANTED	-0.0007	0.0001	-7.2055	0.0000
PET_IN	0.0104	0.0011	9.6264	0.0000
Latitude	0.0209	0.0015	13.6318	0.0000
n, DF		5, 2926		
r-square		0.1553		
AIC		-5630.78		
Moran's I (error)				0.0000

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	-0.5963	0.1401	-4.2578	0.0000
MEAN_ELEV	0.0001	0.0000	7.4811	0.0000
PCT_PLANTED	-0.0007	0.0001	-6.9281	0.0000
PET_IN	0.0111	0.0014	7.8087	0.0000
Latitude	0.0219	0.0021	10.6338	0.0000
LAMBDA	0.3310	0.0258	12.8195	0.0000
n, DF		6, 2925		
r-square		0.2175		
AIC		-5793.06		

Table 33. Regression models for Simpson's diversity, HUC10 watersheds

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.8502	0.0059	144.0720	0.0000
PCT_PLANTE	-0.0031	0.0001	-30.2126	0.0000
PCT_FOREST	0.0007	0.0001	8.1798	0.0000
PCT_URBAN	-0.0018	0.0003	-5.5157	0.0000
FL_FLAG	-0.1269	0.0089	-14.3239	0.0000
GRLVHW_AC	0.0003	0.0001	4.1835	0.0000
n, DF		6, 2925		
r-square		0.3272		
AIC		-4762.44		
Moran's I (error)				0.0000

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	0.8427	0.0081	103.7020	0.0000
PCT_PLANTE	-0.0031	0.0001	-25.7466	0.0000
PCT_FOREST	0.0008	0.0001	6.1384	0.0000
PCT_URBAN	-0.0009	0.0004	-2.4065	0.0161
FL_FLAG	-0.0955	0.0125	-7.6576	0.0000
LAMBDA	0.5214	0.0217	23.9767	0.0000
n, DF		6, 2925		
r-square		0.4528		
AIC		-5203.30		

CHAPTER IV
THE IMPACTS OF VARYING SPATIAL SCALE UPON
PREDCTIONS OF THREE MEASURE OF TREE SPECIES
DIVERSTIY

Abstract

Understanding the forces that affect tree species richness is important for biological conservation and forest management. One of the most topical and controversial of these involve human perturbations into ecological systems. Pine plantation forestry, urban sprawl, and agricultural establishment are examples of anthropogenic actions that can potentially alter tree species diversity. The goal of this study was to develop models of three measures of tree species diversity (species richness, Shannon-Wiener index, and Simpson's index) by incorporating these factors as well as climatic and environmental variables. Additionally, nested watersheds from the watershed boundary dataset (WBD) were used as spatial bounds to determine the impacts of varying spatial scale upon covariates of tree diversity. Ordinary least squares and geographically weighted regression models were developed for each diversity measure and watershed size. Final results reveal that models of larger scale watersheds contained fewer independent variables that are generally global in nature, with little spatial autocorrelation (SA), heteroscedasticity, multicollinearity, or other issues. SA and other issues of model fit became greater as spatial scale decreased. Additionally, the models contained more local variables, human-caused disturbance in particular. The effects of southern pine plantation establishment in particular proved noteworthy, as plantations increased species richness but were negatively related to diversity measures that incorporated species evenness.

Introduction

Trees play a vital role in maintaining forest ecosystem functions. Water flow, water quality, energy flow, and nutrient cycling are all affected by trees and their structure within a forest. There are many examples of studies that link species diversity to forest ecosystem function and productivity (Caspersen & Pacala 2001; Tilman *et al.* 1997). Species diversity is frequently used as a surrogate for forest stability, health, and productivity (McIntosh 1967; McNaughton 1977; Tilman 1996; Austin 2002).

However, the world's landscapes are undergoing rapid change, mainly due to anthropogenic forces (Vitousek *et al.* 1997). Changes in forest structure may have significant bearing on ecosystem integrity (Forman, 1995; O'Neill and Husaker, 1997; Dale *et al.* 2000; Nagendra, 2002). The main factors responsible for biodiversity loss include: land use change; habitat change such as forest fragmentation and conversion; invasive alien species; overexploitation; climate change; and pollution (Hanski 2005).

Forest fragmentation continues to increase in the United States, causing loss of plant and animal habitat while potentially reducing local and regional biodiversity (Reed *et al.* 1996; Riitters *et al.* 2011; Griffith *et al.* 2003). Today,

only 45% of the eastern United States' forest area is not fragmented and that fragmentation threatens forest sustainability as smaller fragments are more likely to shift in composition to edge-adapted and invasive species (Riitters et al. 2011). The impacts of intensive southern yellow pine plantation forestry on biodiversity are a contentious subject. Linder (2004) explains how this type of management could negatively impact biodiversity of southern forests.

A key component to diversity studies is spatial scale. Patterns of species diversity differ as a function of spatial scale. The relationship between scale and species richness and diversity has long been studied (Gleason 1922; Cain 1938; MacArthur & Wilson 1967). As the size of area in question increases, so does species richness and indices of diversity and evenness. Additionally, the factors that influence diversity can differ with scale. Regional variables such as solar radiation, temperature, precipitation, and elevation may prove to be more significant at larger scales (Currie 1991), while local factors such as disturbance may play a larger role in smaller areas. Determining what factors impact tree species diversity at various scales and understanding the underlying processes involved in these relationships is critical for long-term land management and sustainability.

This study investigates the impact of anthropogenic forces such as urbanization, agricultural establishment, and plantation forestry upon three measures of species diversity: species richness, Shannon-Wiener index, and Simpson's index. Climatic and environmental variables were included and investigated as well. Ordinary least squares (OLS) and geographically weighted regression (GWR) models were created for each of the three diversity measures for three different watershed sizes with the goal of determining not only the factors that influence tree species diversity, but how area of interest affects these relationships.

Hypothesis

The results from the predictions were hypothesized to reveal that anthropogenic factors such as agriculture and plantation forestry affect species diversity. Moreover, these factors are likely to positively impact species richness and negatively affect two other diversity indicators; Shannon-Wiener index and Simpson's index.

Data

Tree and Forest Data

The USDA Forest Service Forest Inventory and Analysis (FIA) data comprise the foundation of the dataset. This database stores inventory data from

sample plots taken across the South. Information such as location, species, DBH, height, basal area, standing volume, and average annual growth, removals, and mortality are easily derived from this system. The 1998 Farm Bill requires that a systematic grid of plots is located across the county, and that a standard set of variables and measurements are performed on each plot. The FIA plot grid assigns one plot to approximately every 6,000 acres. The Farm Bill further states that a systematic number of these plots be inventoried for each state every year (Bechtold & Patterson 2005). The southern states average about 6 years for each cycle. Thus, every 5-7 years, a complete set of plots has been re-inventoried for every state.

FIA estimates used include: average latitude and longitude of plots, percent forest/non-forest, acres of pine plantations, acres southern pine plantations expressed as a percentage of total forest area, total forest area, total land area, average site index, average all-live growth per acre, average all-live softwood growth per acre and average all-live hardwood growth per acre in each spatial unit. Additionally, species richness (or a count of species), Shannon's equitability (for evenness), and Simpsons diversity index were calculated for each spatial unit.

GIS Data Layers

Yearly average estimates of temperature, rainfall and potential evapotranspiration (PET) were derived from map layers containing polygons of average annual estimates of the variable in the contiguous United States, for the climatological period 1990-2009. These Parameter Regressions on Independent Slopes Model (PRISM) derived raster data published by the PRISM Climate Group at Oregon State University. 100-meter resolution elevation raster data was obtained from the U.S. Geological Survey. A GeoTIFF file of 2011, 30-meter national land cover data (NLCD) was used to calculate percent forest, urban, and agriculture for each county.

The Watershed Boundary Dataset (WBD) forms the areal extent for this study. The WBD was developed by the U.S. Geological Survey. The Dataset is composed of the entire United States which is divided and sub-divided into successively smaller regions or hydrologic units. These nested units are classified into four levels: regions, sub-regions, accounting units, and cataloging units, with regions being the largest and cataloging units the smallest. Each hydrologic unit is identified by a unique hydrologic unit code (HUC) consisting of two to eight digits based on the four levels of classification in the hydrologic unit system. The smaller the HUC the larger the area. A two digit HUC can be decomposed into a four digit HUCs, which can be further sub-divided into six digit HUCs (U.S. Geological Survey. 2013). Fig 34 illustrates the nesting effect of the WBD. Table 34 reveals the number of HUCs, average, maximum and minimum

number of FIA plots in each HUC size, and the total number of FIA plots in each HUC dataset.

Nested watersheds were selected over other possible boundaries such as ecoregions because they can be continuously sub-divided into successively smaller units. Ecoregions may prove to be better units of delineation, but they are not always the same size and cannot be broken down as the smallest basin in the WBD dataset. As one of the goals of this study was to investigate the impact varying spatial scale, WBD was deemed to accomplish this goal.

All raster data was aggregated to the HUC level using the zonal statistics tool in ArcGIS. Average estimates of temperature, rainfall, PET, average elevation, minimum elevation, maximum elevation, range in elevation, and percent of the watershed in three main land uses: forest, urban and agriculture were computed for each HUC. Table 35 details the variables in the dataset.

Methods

The model developing procedure developed by Anselin (2005) was followed in the creation of the models with one exception. Anselin's solution to the presence of spatial autocorrelation was to develop spatial lag (SLM) or spatial error models (SEM). Geographically weighted regression performed on ArcGIS version 10.3.1 versus spatial lag and error models developed in GeoDA 1.8.14.8. The reason for this is that it is difficult to perform projections using either SLM or SLE models due to the autoregressive nature of the variables. Changing the estimate of one location in a spatially auto correlated dataset should result in changes of neighboring values, which in turn impact the spatially lagged variable used in the model. GWR corrects for spatial dependence and allows for predictions to be made. The SLM and SLE models outlined in previous chapters served as the basis for the GWR models. Small changes in the model occurred only after diagnostics revealed problems or the results did not behave as anticipated. All GWR models were performed in ArcGIS version 10.3.1. with an adaptive kernel type (Gaussian) and Akaike Information Criterion (AICc) bandwidth settings.

Predictions were made by varying environmental and anthropogenic variables in each model up to 15 percent (plus and minus) in 5 percent increments. If total forest area is a significant variable, then six projections are made by adjusting forest area $\pm$ five percent each time, with a maximum of 15%.

Results

Projections of HUC04 species richness were based on the ordinary least squares model described in table 36. Projections are based on varying all-live tree growth per acre 5, 10, and 15 percent and total forest area by the same amount. The model predicts that on average, each HUC04 watershed will contain 88.7 different tree species. A 15 percent decrease in average all-live growth will result in an estimate of 86.1, and a 15 percent increase in growth results in 91.3 species. Likewise, a 15 percent increase or decrease in total forest area yields similar estimates of 86.2 and 91.2 respectively. The third row represents what can be expected if both growth and area increase by the same amount, while the last row has the two variables moving in opposite directions, with the -15% column representing a 15 percent decrease in growth combined with a 15 percent increase in forest area. These values change by 5 percent in opposite directions across the table, where the last column contains the estimate for the maximum value for forest growth and 15 percent decrease in forest area.

Table 37 illustrates the impact of varying the percent of the watershed in agriculture and total forest area in each HUC04 watershed for the OLS model on Shannon-Wiener index found in table 22. Likewise, table 38 reflects how changes in the percent of watershed in agriculture and forests in plantations affects Simpson's diversity.

Tables 39, 40, and 41 reveal the impacts of the sensitivity analysis on HUC06 watersheds for species richness, Shannon-Wiener index, and Simpson's index. The models for Shannon-Wiener and Simpson's includes minimum elevation as a variable, but it is not included in the tables below, as that is not a variable that is likely to change.

HUC08 diversity models from the previous chapter were derived from either SLM or SEM procedures. As these methods do not allow forecasting, GWR models were created. The GWR model for HUC08 species richness is similar to table 27 in that it includes percent of watershed in agriculture, percent of watershed in forests, total forest area and average all-live growth per acre. The R^2 for this model was 0.75, with an AIC of 3015.7 (table 42).

The ArcGIS GWR procedure produces a series of maps that reveal the distribution of residuals as well as the impact of the coefficients across the study area (figure 35).

The GWR model for Shannon-Wiener index included the following variables: percent forest, percent agriculture, average all-live growth of softwood species per acre, and percent of forests in plantations. It has an R^2 of 0.75 and an AIC of 3015.73 (table 44).

Table 45 reveals the impacts of varying the coefficients of the model in table 41 by ± 15 percent. Changes in percent of watershed in forests has the greatest impact upon this diversity measure, followed by percent of forests in plantations and percent of watershed in agriculture respectively.

The GWR series of maps that reveal the distribution of residuals as well as the impact of the coefficients across the study area for the Shannon-Wiener diversity index (figure 36).

The final model for this study is a GWR model of Simpson's diversity index for HUC08 watersheds. Percent of forests in plantation, percent of watershed in agriculture and percent of watershed in forests are the dependent variables (table 43). It has an R^2 of 0.57 and an AIC of -1305.21. The sensitivity table (table 47) and map of residuals and coefficients (figure 37) reveal that changes in percent of the land in agriculture has the greatest impact upon Simpson's index for this spatial scale.

Discussion

Patterns of predictors emerged for each diversity measure. For species richness, total forest area and average all-live growth per acre appeared in all three spatial scale models. This seems intuitive, as the largest, most productive watersheds will contain the most tree species. Elevational variables were important for the larger watersheds, as range in elevation was significant for HUC04 and HUC06 watersheds. This elevational element disappeared for the smallest watershed. The absence of elevation in HUC08 basins may suggest that this is the scale at which the importance of global parameters begins to decrease and regional or local variables gain significance.

Investigating the sensitivity tables for species richness (tables 36, 39, and 43) reveals patterns within each spatial scale. For HUC04 basins, variations in range of elevation had the smallest impact on number of tree species, while average all-live growth and total forest area had similar effects. The net impact on species richness was almost negligible if both vary the same amount in different directions.

The sensitivity table for species richness for HUC06 watersheds (table 39) illustrated many similarities to those reported in the previous paragraph. Again, changes in either growth or forest area resulted in a corresponding change in species richness of around 2.5 species. The main difference was that growth was 2.1 species while total forest area was almost 3. PET has the greatest impact, but it was negatively correlated to species richness, which seems counter

intuitive. One possible explanation may be the long-term, human-caused disturbances that occurred in the Gulf Coastal Plain.

The species change for a 15 percent change in forest area of HUC08 watersheds was again about 2 species. However, this is actually a greater change, as the average HUC08 watershed contains 49 species, while the HUC04 and HUC06 average 89 and 72, respectively. Percent of land area in agriculture was significant in the HUC08 model and positive. Agriculture can be associated with disturbance, land-use edges, and introduced species, any of which are likely to increase the total number of species in an area. This factor may explain why agriculture is currently positively correlated to species richness. However, as most of the study area is in forested land-uses, this relationship may reach a tipping point if agriculture ever becomes the dominant land use.

The three predictors of Shannon-Wiener diversity index were found across all three spatial scales. Percent of the land in agriculture land use and a variable representing forest area, either total forest area in acres or percent of basin in forests, appeared in all three models (tables 37, 40 and 45). In all three HUCs, these variables were positively correlated. As with species richness, elevation variables were in the two largest watersheds but not significant in the smallest, HUC08. Total growth of all-live softwoods and percent of forests in plantations were in the HUC08 Shannon-Wiener model. There is a correlation between all-live softwood growth and plantation forestry, as plantations are going to be established in areas where they can be most productive, i.e. areas with high potential for softwood growth.

The sensitivity tables for Shannon-Wiener reveal that changes in the amount of land under agricultural and forests had the greatest impact on this measure. A variable representing plantations was significant in the smallest watershed size, HUC08, and negatively correlated to Shannon-Wiener. This lends credence to the idea that plantations impact species evenness, as this indicator balances species richness with species evenness and it becomes apparent at this scale.

Percent of forests in plantations was significant for all three spatial scales for Simpson's diversity index model (tables 38, 41 and 45). Simpson's index emphasizes species evenness, suggesting that plantations negatively impact this measure. Predictions using less agriculture produced higher estimates of Simpson's diversity in all three scales. As with species richness and Shannon-Wiener, elevation variables were present in HUC04 and HUC08 watersheds, but not included in HUC08 basins.

Conclusions

Global and regional variables were more important at larger scales in the study. Factors such as forest area, elevation, and major land uses are key for this level of investigation. As spatial scale decreases, elevation disappears. The models tend to become more complex, as variables relating to local factors such as plantation forestry become more significant.

One interesting but intuitive relationship highlighted by this study is the role forest area plays in species richness. The species-area relationship has long been understood by ecologists and is the reason why species-area curves are highly utilized in diversity studies (MacArthur & Wilson 1967; Browne & Peck 1996). What is intriguing about these results is that the results from the species richness models (tables 37, 40, & 44) suggest that the curve may be initially steeper with a more dramatic tapering off than the curve in figure 15. A 15 percent increase or decrease in forest area of HUC04, HUC06 and HUC08 basins results in approximately 2.5 species change in the same direction. This seems remarkable as the average number of species in HUC04, HUC06 and HUC08 basins is 88, 72, and 49 respectively.

Urban land-uses were not significant in any model or at any scale. One possible explanation for this is that the region, while under constant urbanization and development, is still dominated by forest land uses, and urban development has yet to reach a point that impacts these measures. Agriculture also had a positive effect on almost every single indicator. Again, forests tend to be the dominant land use in most of the study area. There may be a tipping point at which either or both agriculture establishment or urbanization may negatively impact tree species diversity, but it has not been reached yet.

Finally, all three indicators used in this study are numeric values. They do not reflect how the population may actually be changing. It is possible to have a dominant species or forest community decrease and be replaced by another species or community. This process can happen at multiple levels. Hartsell and Rosson (2007) report on how many upland forest communities have changed over time. In particular several species of oak (*quercus sp.*) and shortleaf pine (*Pinus echinata*) and are being replaced by loblolly pine (*pinus taeda*) and sweetgum (*Liquidambar styraciflua*). This may not be picked up by a numeric value that is just tracking the numbers and distributions of species.

References

- Alatalo, R. V. 1981. Problems in the measurement of evenness in ecology. *Oikos* 37:199–204.
- Anselin, L. 2005. Exploring Spatial Data with GeoDa: A workbook. Center for Spatially Integrated Social Science. Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana-Champaign. <http://geodacenter.asu.edu/system/files/geodaworkbook.pdf>. 244p.
- Austin M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Model.* 157:101–18.
- Bechtold, W.A., Patterson, P.L. Editors. 2005. The enhanced Forest Inventory and Analysis Program – national sampling design and estimation procedures. Gen. Tech. rep. SRS 80. Ashville, NC: USDA Forest Service, Southern Research Station. 85p.
- Browne, J. B. and Peck, S.B. 1996. The long-horned beetles of south Florida (Cerambycidae: Coleoptera): biogeography and relationships with the Bahama Islands and Cuba. *Canadian Journal of Zoology*, 74:2154-2169.
- Cain, S.A. 1938. The species-area curve. *American Midland Naturalist*, 19, 573-581.
- Caspersen, J.P. Pacala, S.W. 2001. Successional diversity and forest ecosystem function. *Ecological Research* 116: 895-903.
- Currie, D.J. 1991. Energy and Large-Scale Patterns of Animal- and Plant Species Richness. *The American Naturalist* 137: 1, 27-49.
- Dale, V.H., Brown, S., Haeuber, R.A., Hobbs, N.T., Huntly, N., Naiman, R.J., Riebsame, W. E., Turner, M.G., and Valone, J.J. 2000. Ecological principles and guidelines for managing the use of land. *Ecological Applications*, 10, 639–670.
- Ehrlich, P. R. and Wilson, E.O. 1991. Biodiversity studies: science and policy. *Science* 253.5021: 758p.
- Forman, R.T.T. 1995. Land mosaics: the ecology of landscapes and regions. Cambridge: Cambridge University Press.
- Gleason, H.A. 1922. On the relation between species and area. *Ecology*, 3, 158-162.

- Griffith, J.A., Stehman, S.V., Loveland, T.R. 2003. Landscape trends in Mid-Atlantic and southeastern United States ecoregions. *Environ Manag* 32:572–588.
- Hanski, Ilkka. 2005. Landscape fragmentation, biodiversity loss and the societal response. EMBO reports VOL 6 - NO 5 – 2005
- Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–432.
- Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52: 577–586.
- Linder, E.T. 2004. Chapter 25. Biodiversity and Southern Forests. In- Southern forest science: past, present, and future. Gen. Tech. Rep. SRS–75. Rausdher, H.M & Johnsen, K. eds. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 394 p
- MacArthur and Wilson. 1967. The Theory of Island Biogeography. Princeton University Press: Princeton, NJ.
- Magurran, A.E. 1988. Ecological Diversity and Its Measurements. Chapman & Hall, London, 179 p.
- McIntosh, R. I. 1967. An index of diversity and the relation of certain concepts to diversity. *Ecology* 48:392–404.
- McNaughton, S. J. 1977. Diversity and stability of ecological communities: a comment on the role of empiricism in ecology. *American Naturalist* 111:515–525.
- Naeem, S., et al. 1996. Biodiversity and Plant Productivity in a Model Assemblage of Plant Species. *Oikos* vol. 76, no. 2: 259–264. JSTOR, www.jstor.org/stable/3546198.
- Nagendra, H. 2002. Opposite trends in response for Shannon and Simpson indices of landscape diversity. *Applied Geography*. 22: 175-186.
- O'Neill, R.V. and Hunsaker, C.T. 1997. Monitoring environmental quality at the landscape scale. *BioScience*, 47, 513–519.
- Reed, R., Lemak, D. J., & Montgomery, J. C. 1996. Beyond process: TQM content and firm performance. *Academy of Management Review*: 21, 173-202.

- Riitters H.H., Wickham, J.D., O'Neill, R.V., Jones, K.B., Smith, E.R., Coulston, J.R., Wade, T.G., Smith, J.H. 2002. Fragmentation of Continental United States Forests. *Ecosystems* 5:8, 815-822.
- Routledge, R. D. 1980. Bias in estimating the diversity of large, uncensused communities. *Ecology* 61:276–281.
- U.S. Geological Survey. 2013. National Hydrography Geodatabase: The National Map viewer available on the World Wide Web (<http://viewer.nationalmap.gov/viewer/nhd.html?p=nhd>), accessed [July 28, 2014]
- Vitousek, P.M., Flannigan, M.d., Wotton, B.M., and Suffling, R. 1997. Human domination of the Earth's ecosystems. *Science*, 277, 494-504.
- Wilson, Edward O. 1993. The Diversity of Life. W. W. Norton & Company. 448p.

Appendix

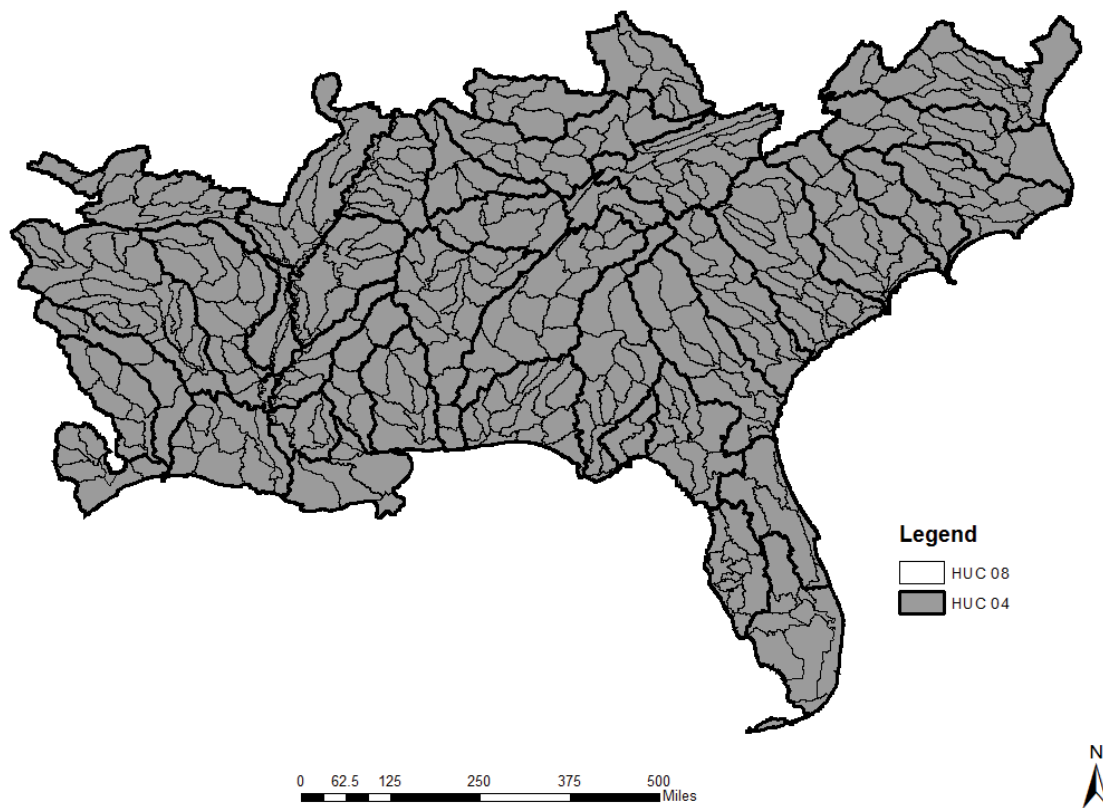


Figure 34. HUC08 basins nested within HUC04 watersheds

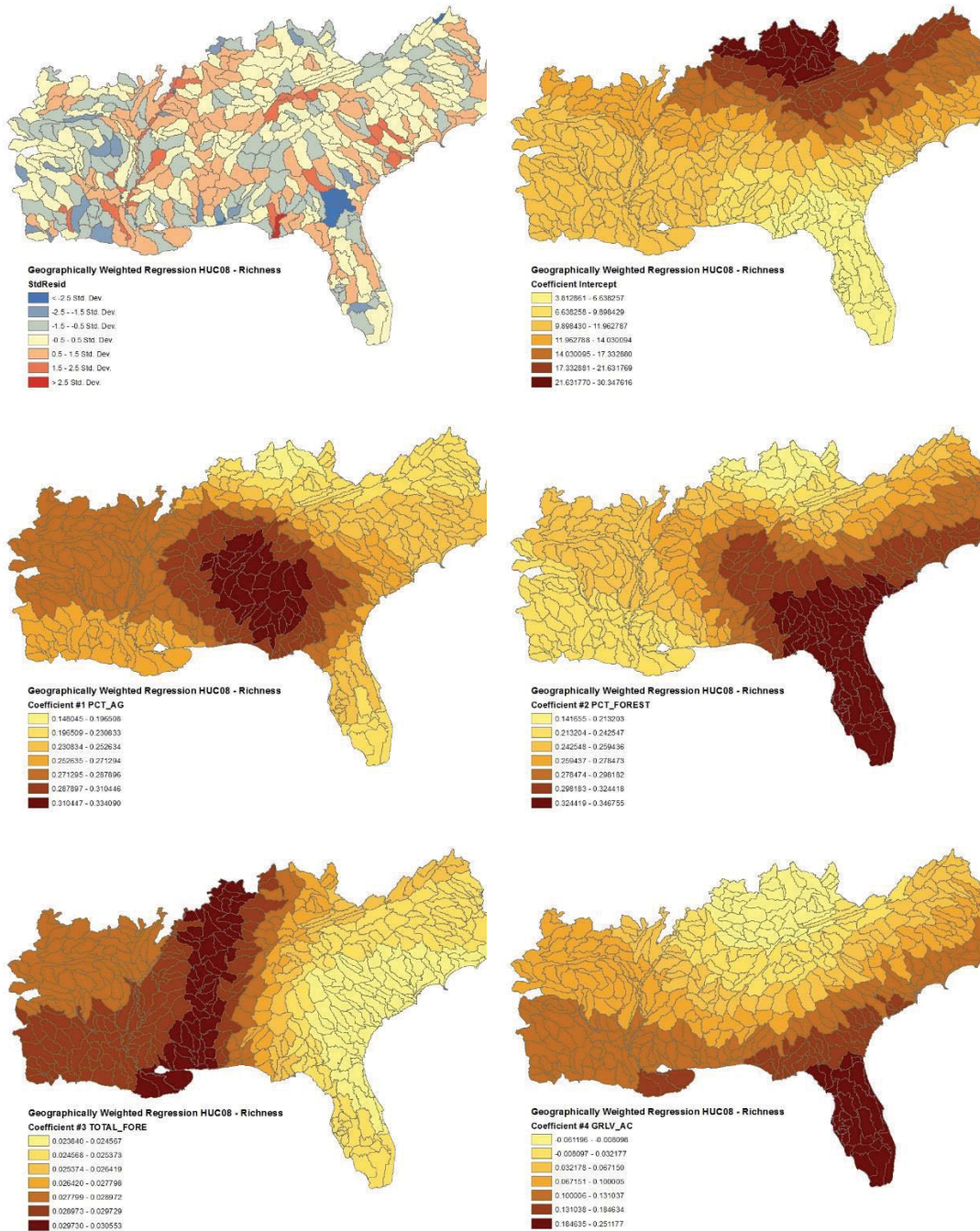


Figure 35. Map of residuals, intercept, percent agriculture, percent forest, total forest, and all-live growth coefficients from GWR model of species richness for HUC08 watersheds.

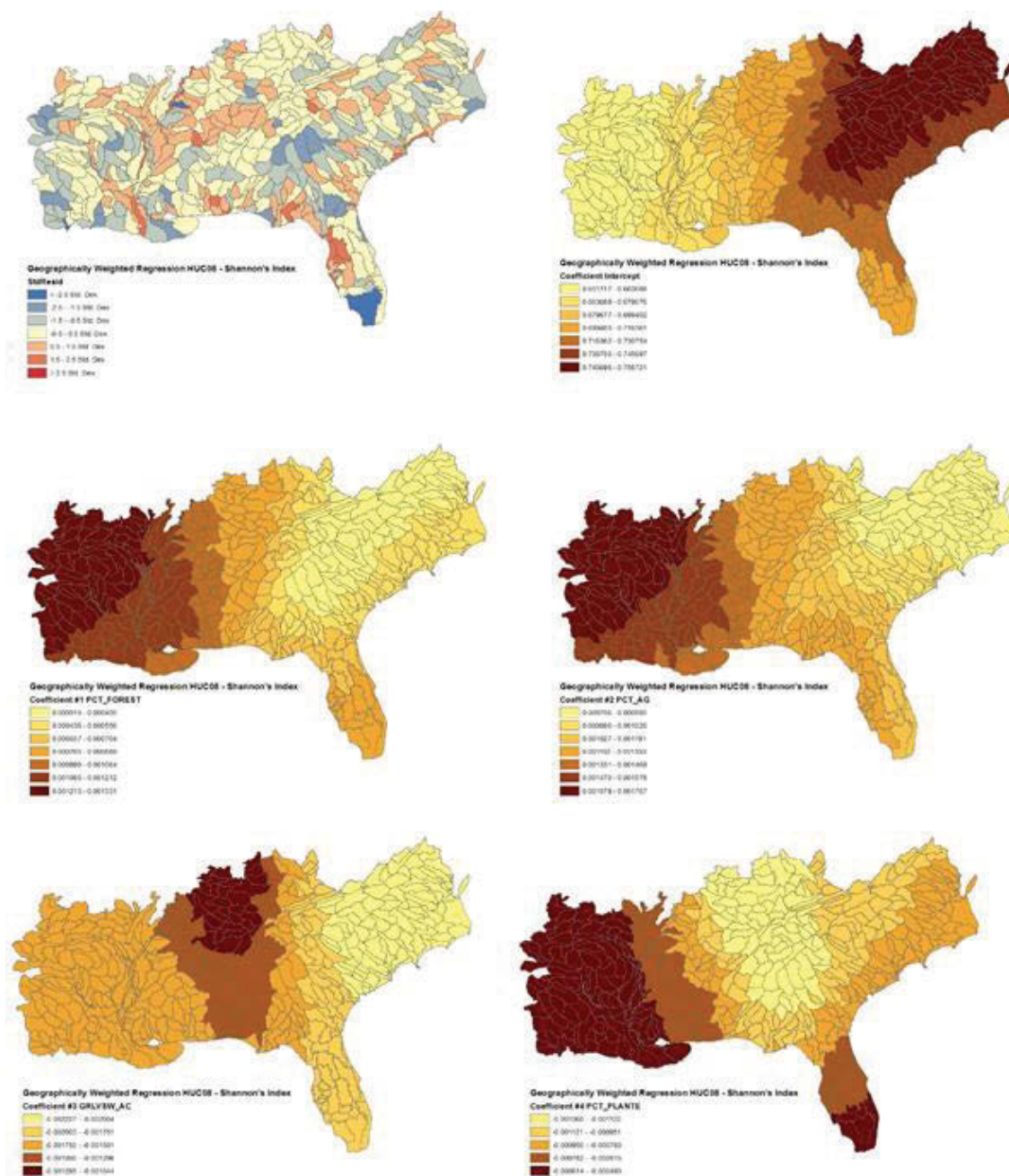


Figure 36. Map of residuals, intercept, percent agriculture, percent forest, total forest, and all-live softwood growth, and percent of forests in plantation coefficients from GWR model of Shannon-Wiener index for HUC08 watersheds.

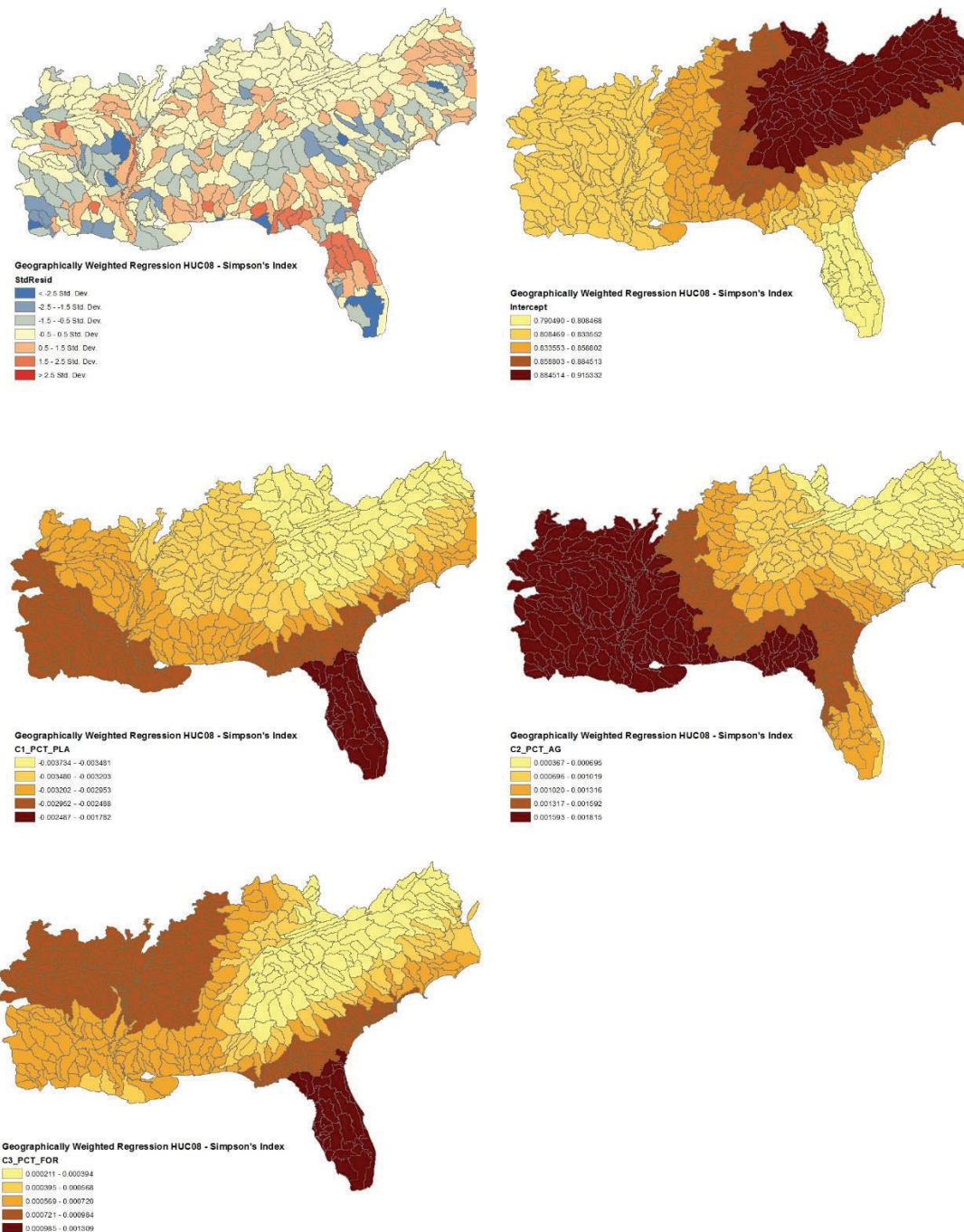


Figure 37. Map of residuals, intercept, and percent of forests in plantations and percent of watershed in agriculture and forest coefficients from GWR model of Shannon-Wiener index for HUC08 watersheds.

Table 34. Watershed sizes and the average, maximum, minimum, and total number of FIA plots in each HUC class

Unit	Sample size	Average (acres)	Plots			FIA Plots (n)
			Average	Max	Min	
HUC04	40	2,734,800	455.8	1000	76	18232
HUC06	79	1,420,680	236.78	760	7	18706
HUC08	430	271,800	45.3	143	4	19478
HUC10	2828	41,100	6.85	38	1	19371
HUC12	11554	11,400	1.9	28	1	19579

Table 35. Variable list and description for all watersheds

Variable	Description
OBJECTID	Object ID
AreaAcres	area of polygon in acres
HUC##	Hodrologic Unit Code (HUC) number
TOTAL_LAND	Total land area (acres)
TOTAL_FORE	Total forest area (acres)
TOTAL_NONF	Total non-forest area (acres)
PLANTED_AC	Area of planted acres
SYP_PLANTE	Area of southern yellow pine plantations (acres)
PCT_SYP	Percent of forests in southern yellow pine plantations
RICHNESS	tree species richness
SHANNONS_E	Shannon-Wiener equitability index
SIMPSONS	Simpson's index
PCT_AG	Percent of landbase in agricultural uses (2011)
PCT_FOREST	Percent of landbase in forests (2011)
PCT_URBAN	Percent of landbase in urban (2011)
MAX_ELEV	Maximum elevation in watershed
MIN_ELEV	Minimum elevation in watershed
RANGE_ELEV	range in elevation in watershed
MEAN_ELEV	Average elevation in watershed
PET_IN	Average annual PET (mm/year)
AVG_IN	Average annual precipitation (inches/year)
AVG_FAHREN	Average annual temperature (Fahrenheit)
Long	Average logitude of FIA plots in watershed
Lat	Average latitude of FIA plots in watershed

Table 35 Continued. Variable list and description for all watersheds

Variable	Description
PCT_PLANTE FL1	Percent of the forests classified as planted Geographic (dummy) variable for southern half of Florida
GCP1	Geographic (dummy) variable for all watersheds that touch Gulf of Mexico
GCP2	Geographic (dummy) variable for all watersheds that touch Gulf of Mexico or Atlantic ocean
East_Flag	Geographic (dummy) variable for al watersheds on eastern side of Mississippi river
Delta	Geographic (dummy) variable for watersheds in Mississippi delta
GRLV	Total all-live growth of all-live trees (cubic feet)
GRLVSW	Total all-live growth of all-live softwood tree species (cubic feet)
GRLVHW	Total all-live growth of all-live hardwood tree species (cubic feet)
GRLV_AC	Average annual all-live growth of all-live tree species (cu.ft./year)
GRLVSW_AC	Average annual all-live growth of all-live softwood tree species (cu.ft./year)
GRLVHW_AC	Average annual all-live growth of all-live hardwood tree species (cu.ft./year)

Table 36. Effects of varying average all-live growth per acre and total forest area on OLS model on species richness for HUC04 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Average all-live growth/acre	86.1	87.0	87.8	88.7	89.6	90.4	91.3
Total forest area	86.2	87.1	87.9	88.7	89.5	90.4	91.2
Growth and area	83.6	85.3	87.0	88.7	90.4	92.1	93.8
Minus growth Plus area	88.6	88.7	88.7	88.7	88.7	88.7	88.8

Table 37. Effects of varying average percent agriculture and total forest area on OLS model on Shannon-Wiener diversity index for HUC04 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Percent agriculture	0.7760	0.7784	0.7809	0.7833	0.7857	0.7881	0.7906
Total forest area	0.7781	0.7799	0.7816	0.7833	0.7850	0.7867	0.7885
AG(-) Forest Area (+)	0.7854	0.7847	0.7840	0.7833	0.7826	0.7819	0.7812

Table 38. Effects of varying average percent agriculture and percent of watershed in plantations of OLS model on Simpson's diversity index for HUC04 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Percent agriculture	0.8760	0.8775	0.8790	0.8805	0.8819	0.8834	0.8849
Percent in plantations	0.8861	0.8842	0.8823	0.8805	0.8786	0.8767	0.8748
AG(-) Plantations (+)	0.8704	0.8737	0.8771	0.8805	0.8838	0.8872	0.8906

Table 39. Effects of varying average all-live hardwood growth/acre, total forest area (acres), and potential evapotranspiration (PET) of OLS model on tree species richness for HUC06 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
All-live growth/acre	69.9	70.6	71.3	72.0	72.7	73.4	74.1
Total forest area	69.1	70.0	71.0	72.0	72.9	73.9	74.9
PET in inches	78.7	76.4	74.2	72.0	69.7	67.5	65.2
Forest(-) PET (+)	62.3	65.5	68.8	72.0	75.2	78.4	81.6

Table 40. Effects of varying percent of watershed in agriculture and total forest area (acres) of OLS model on tree Shannon-Wiener index for HUC06 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Percent in agriculture	0.7073	0.7092	0.7110	0.7129	0.7148	0.7167	0.7186
Total forest area	0.7170	0.7157	0.7143	0.7129	0.7116	0.7102	0.7088
Ag and Forest	0.7227	0.7195	0.7162	0.7129	0.7097	0.7064	0.7032

Table 41. Effects of varying percent of forests in plantations and average hardwood tree growth/acre of OLS model on tree Simpson's index for HUC06 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Percent in plantations	0.8876	0.8858	0.8839	0.8820	0.8801	0.8783	0.8764
Hardwood growth/acre	0.8785	0.8797	0.8808	0.8820	0.8832	0.8844	0.8855
Ag (-) and Forest (+)	0.8912	0.8881	0.8851	0.8820	0.8790	0.8759	0.8729

Table 42. Geographically weighted regression model for species richness of HUC08 watersheds.

Constant	12.677
Percent in agriculture	0.2671
Percent in forest	0.2735
Total forest area (acres)	0.0273
Average all-live growth/acre	0.0842
AIC	3015.7
r-square	0.7524

Table 43. Effects of varying percent of watershed in agriculture, percent of watershed in forests and total forest area (acres) of GWR model on tree species richness for HUC08 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Percent agriculture	48.1976	48.5504	48.9031	49.2559	49.6086	49.9614	50.3141
Percent forest	47.2090	47.8913	48.5736	49.2559	49.9382	50.6205	51.3028
Total forest area	47.3186	47.9644	48.6101	49.2559	49.9016	50.5474	51.1931

Table 44. Geographically weighted regression model for Shannon-Wiener diversity index of HUC08 watersheds.

Constant	0.708038
Percent forest	0.000805
Percent Ag	0.001247
GRLVSW	-0.00169
Pct Planted	-0.000848
AIC	3015.73
r-square	0.7524

Table 45. Effects of varying percent of watershed in agriculture, percent of watershed in forests and percent of forests in plantations of GWR model on Shannon-Wiener diversity index for HUC08 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Percent forest	0.724	0.726	0.728	0.730	0.732	0.734	0.736
Percent agriculture	0.725	0.726	0.728	0.730	0.732	0.733	0.735
Percent in plantations	0.732	0.731	0.731	0.730	0.729	0.728	0.728

Table 46. Geographically weighted regression model for Simpson's diversity index of HUC08 watersheds.

Constant	0.8519
Pct Planted	-0.003
Percent Ag	0.0013
Percent forest	0.0006
AIC	-1305
r-square	0.5664

Table 47. Effects of varying percent of watershed in agriculture, percent of watershed in forests and percent of forests in plantations of GWR model on Simpson's diversity index for HUC08 watersheds.

	Percent change in parameter						
	-15%	-10%	-5%	0%	5%	10%	15%
Percent of forests in plantations	0.867	0.864	0.861	0.858	0.855	0.852	0.850
Percent agriculture	0.853	0.855	0.857	0.858	0.860	0.862	0.864
Percent forest	0.854	0.855	0.857	0.858	0.860	0.861	0.863

CHAPTER V CONCLUSIONS

The results from Chapter 1 indicate that tree species diversity has changed over time. This study suggests that anthropogenic forces have impacted tree species diversity. Processes such as plantation forestry affect species diversity in terms of richness and evenness (Simpson's diversity index). Plantations were found to increase the number of species in an area. This may be due to edge effect. This combined with the fact that many of these managed forests occur in rural areas where forest acreage is high, and area of forests is the greatest predictor of species richness.

The first study revealed that urban development and agriculture have an impact as well. Both have the potential to increase species richness and other indicators of diversity by promoting patches of disturbance and introducing new species. However, this study is based on the number of species and does not track species changes in species composition. Species often favored by plantation forestry, urbanization, and agriculture typically are early successional and/or non-native invasive species, and are introduced into the study area via anthropogenic means. These new species would add to the list of species already present in these areas.

Increasing plantation area has a negative impact on both Shannon's diversity index and Simpson's index. As Simpson's is a measure of evenness, this indicates that the species are being dominated by a few species, as shown in Table 4. So, overall species numbers have yet to be impacted by plantation forestry, but it has affected the distribution of species negatively.

How spatial scale impacts predictors of tree species diversity was investigated as well. This required determining the appropriate scale for FIA data. In Chapter 2, I determined that spatial units needed to be at least 600,000 to 900,000 acres in size. This makes HUC04 and HUC06 watersheds entirely appropriate for this type of study, and HUC08 occurring on the edge.

How factors change with scale was another topic under investigation. Larger scales, and global or regional factors such as forest area and elevation tended to be highly important. As area decreased, factors such as productivity and land-use increased in significance. Additionally, models of the larger units tended to contain fewer predictor variables and "behaved" in an appropriate manner, meaning that issues such as multicollinearity, variance inflation factors, heterogeneity, independence, and spatial autocorrelation were not serious. Larger scale models usually had higher R^2 s and could be solved with traditional statistical methods such as OLS. These became problems as

spatial scale decreased. The models also became more complex and contained more independent variables. Advanced modeling techniques under the category of spatial statistics needed to be utilized to develop proper models.

Chapters 2 and 3 revealed that plantation forestry can have a negative impact on evenness, particularly at smaller scales. When the area of interest is large enough, the patchwork of plantations will not impact the numeric estimate of diversity. However, as the scale decreases, this is not the case. Rosson and Hartsell (2007) describe how forest structure changes with timber management. As noted earlier, the numeric values assigned to diversity indexes will fail to capture this type species changes.

An interesting artifact of the data and study area became apparent during the development of these models. These models were created with the idea of applying climate change estimates to determine their impacts on species diversity. However, diversity measures often did not behave as anticipated. For example, species richness is not supposed to decrease as rainfall, temperature, or PET increases. Other measures would decrease as productivity increases. These occurred in this study. This was confounding at first. However, further data interrogation revealed that plantation establishment occurred in fertile areas that have favorable environmental conditions. Humans locate agriculture and managed forests in areas of favorable environment and site locations. Therefore, plantations are positively correlated to productive sites. Literature detailing the clearing of original forests for agriculture and later the conversion of these farms to managed forests is extensive. USDA (1988), Owen (2002), and Mitchel and Duncan (2009) detail this transformation. The southeastern United States “Pine Belt” has been repeatedly disturbed due to anthropogenic activity. This disturbance will confound many long-term studies that focus on climatic impacts on tree species diversity, particularly for the purposes of prediction.

Thus, what are the long-term consequences of plantation forestry? The results presented here demonstrate that this type of management affects diversity at smaller scales. This is likely to continue and will impact larger scales as this type of management continues. However, current productivity of these forests suggests that this is not irreversible. If left alone and unmanaged, the forests would eventually become more diverse. While the forests will never become the forests of pre-European time, as many species such as American chestnut (*Castanea dentata*) have become virtually extinct, the forest structure can revert to a similar form if left alone.

Land-use change from forests to urban or agriculture is the biggest threat facing US forests. It interesting and confusing that these factors did not appear in this study. One possible explanation is that the scale to which these can be detected is smaller than the appropriate scale of the FIA data that forms the

basis of this work. There may be tipping point in the future were either farming or urban development increase to the size that FIA data can provide further insight into this. Another possibility is to expand the study area and time frame to include more urban and agricultural areas.

APPENDIX – DATA & METHODS

Forests of the Southern United States

The South, as defined in this study, ranges across 13 states from Virginia to Texas. Forests dominate the landscape, as this land use accounts for 40 to more than 80 percent in many counties with a few exceptions in the western grasslands of Texas and Oklahoma, and in key agricultural areas in the Mississippi Valley, southern Florida, and parts of Kentucky and Tennessee (fig. 1). Total forest area has been stable since the 1970s. However, this statistic does reflect the amount of change that has occurred on the southern landscape. Throughout this time, agriculture lands have been converted to forests at about the same rate that forests have been cleared for urban development. Urban areas have replaced forests at a multi-county level in a few places, most notably the corridor between Atlanta, GA and Charlotte, NC. (fig. 38). Strong population growth is expected to continue in these areas, and possibly have greater impacts on other areas of the South (Wear & Greis, 2012).

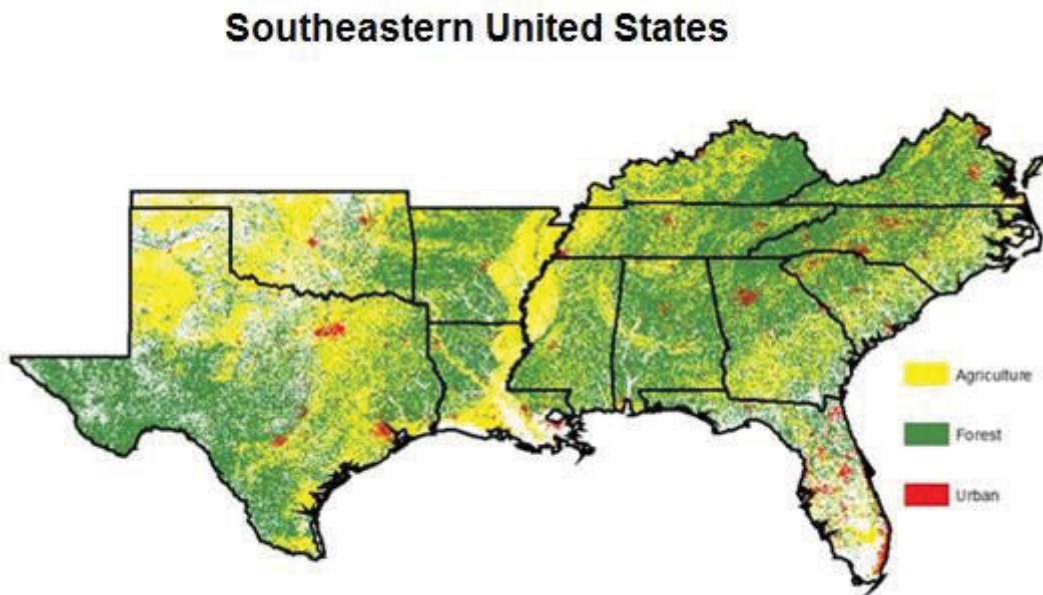


Figure 38. Land uses of the southeastern United States, 2011 NLCD

The forests of the South support a diverse biota that contain 1,027 native terrestrial vertebrate species; 178 amphibians, 504 birds, 158 mammals, and 187 reptiles. Recent increases in nonnative, invasive plants, insects, and diseases have affected the forest landscape as well (Wear & Gries, 2012). Emerald ash borer and laurel wilt have impacted many sites, causing a loss in host species. Additionally, nonnative tree species such as Chinese tallowtree (*Triadica*

sebifera), chinaberry (*Melia azedarach*), and other have been spreading across the landscape replacing or competing with native species (Miller, 2003).

Forest types ranging from upland hardwoods, natural and artificial pine stands, lowland gum-cypress swamps, high elevation spruce-fir, and coastal mangrove/live-oak forests can be found within the southern region. Figure 3 reveals the distribution of these forests and how they are often comingled together based on microclimate, disturbance, and other local forces. Upland hardwoods dominate the Piedmont and northern gulf coastal plain. Pine is the dominant type for most of the gulf coastal plain, while lowland forest types dominate the Mississippi Alluvial Valley (Jin et al, 2013)

Tree and plot data

USDA Forest Service Forest Inventory and Analysis (FIA) data comprises the foundation of the dataset. This database stores inventory data from sample plots taken across the South. Information such as location, species, DBH, height, basal area, standing volume, and average annual growth, removals, and mortality are easily derived from this system. The 1998 Farm Bill requires that a systematic grid of plots is located across the county, and that a standard set of variables and measurements are performed on each plot. The FIA plot grid assigns one plot to approximately every 6,000 acres. The Farm Bill further states that a systematic number of these plots be inventoried for each state every year (Bechtold & Patterson 2005). The southern states average about 6 years for each cycle. Thus, every 5-7 years, a complete set of plots has been re-inventoried for every state.

FIA estimates used include: average latitude and longitude of plots, percent forest/non-forest, acres of pine plantations, acres of southern pine plantations expressed as a percentage of total forest area, total forest area, total land area, average site index, average all-live growth per acre, average all-live softwood growth per acre and average all-live hardwood growth per acre in each spatial unit. Additionally species richness (or a count of species), Shannon-Wiener's equitability (for evenness), and Simpson's diversity index were calculated for each spatial unit.

GIS Data Layers

Yearly average estimates of temperature and rainfall are derived from the parameter regressions on independent slopes model (PRISM) derived raster data is published by the PRISM Climate Group at Oregon State University. One hundred meter resolution elevation raster data was obtained from the U.S. Geological Survey. The average annual yearly potential evapotranspiration (PET) raster dataset from the Global Aridity and PET database was used to calculate

average annual PET for each county. A GeoTIFF file of 2011, 30 meter national land cover data (NLCD) was used to calculate percent forest, urban, and agriculture for each spatial unit.

The Watershed Boundary Dataset (WBD) forms the areal extent for the last two studies. The WBD was developed by the U.S. Geological Survey. The Dataset is composed of the entire United States which is divided and sub-divided into successively smaller regions or hydrologic units. These nested units are classified into four levels: regions, sub-regions, accounting units, and cataloging units, with regions being the largest and cataloging units the smallest. Each hydrologic unit is identified by a unique hydrologic unit code (HUC) consisting of two to eight digits based on the four levels of classification in the hydrologic unit system. The smaller the HUC the larger the area. A two digit HUC can be decomposed into a four digit HUCs, which can be further sub-divided into six digit HUCs (U.S. Geological Survey. 2013). Figure 14 illustrates the nesting effect of the WBD. Table 12 details the number of HUCs, average, maximum and minimum number of FIA plots in each HUC size, and the total number of FIA plots in each HUC dataset.

All raster data was aggregated to the HUC level using the zonal statistics tool in ArcGIS. Average estimates of temperature, rainfall, PET, elevation and range of elevation were computed for each HUC. Additionally, range of elevation was calculated as well. The percent of each county in the three main land uses (forest, urban, and agriculture) were calculated in a similar fashion.

Diversity Indices

This study investigates three measures of species diversity. These include: species richness, Shannon-Wiener index and Simpson's index. These indices were calculated with the "diversitycomp" command in the "BiodiversityR" package in R-Studio version 1.0.136. FIA tree and plot level data were used to create the data matrix of species counts by plot and watershed.

Species Richness

This study focuses on two types of biological diversity measurements: richness and evenness. Richness is simply the number of different organisms or species living in an area (Wilson 1993; Ehrlich & Wilson 1991). It does not take into account proportion or distribution of each subspecies within the area of interest. All one must do is simply count the number of subspecies found in a community. It is the simplest and easiest diversity measurement to calculate and describe. Species richness is a key component in many the plant ecology community (Naeem et al 1996).

Relative species abundance within a community is often measured with an index ranging from 0 to 1 and equated to species evenness (Hulbert 1971; Alatalo 1981). Evenness refers to how the subspecies are distributed throughout the whole community (Routledge 1980; Magurran 1988). It, in effect, is evenness normalized. For example, there are two communities composed of five different subspecies. The first one has counts of 20 for each subspecies for a total of 100 individual subspecies counted. The second one has one subspecies with a count of 96, and the other four have one observation each. The second community also has a total of 100 individual subspecies counted. The first community would be considered more even, as each subspecies accounts for the same percentage of the total, while the second one is dominated by a single species.

Shannon-Wiener Index

Proportional abundance indices depend on both richness and relative abundance (Hill 1973). Shannon-Wiener (Shannon's for short) evenness and diversity index (H') can from information theory and measure the order and disorder within a population. This index is derived by calculating the proportion of species i relative to the total number of species (p_i), and then multiplying by the natural logarithm of this proportion ($\ln(p_i)$). The result is summed across species, and multiplied by -1:

$$H = - \sum_{j=1}^s p_i \ln p_i$$

Simpson's Index

Simpson's evenness and diversity index (D) was developed in 1949. Like Shannon's index, it also takes into account richness and diversity. Simpson's D can be found in three forms. The one used in this study is derived by taking the proportion of species i relative to the total number of species (p_i) and then squaring. The squared proportions for all the species are summed, and the reciprocal is taken.

$$D = \frac{1}{\sum_{j=1}^s p_i^2}$$

An evenness or equitability index can be calculated for Simpson's D , just as one can be derived from Shannon's H . Simpson's equitability index (ED) is calculated by taking Simpson's index (D) and dividing it by the total number of

subspecies groups (S). Thus the formula for ED is: $ED = D / S$, where S = total number of subspecies groups. Evenness or equitability takes a value between 0 and 1, with 1 being complete evenness.

Multiple Comparison Procedures

Analysis of Variance (ANOVA)

The traditional tool for comparing two groups and testing for statistically significant differences is the t-test. Analysis of variance (ANOVA) is the standard statistical technique used when comparing three or more groups. Simply stated, ANOVA compares variance within groups to variance between groups. The null hypothesis is that there is no difference between groups. If the variance within groups is significantly greater than the variance between groups, then the null hypothesis is usually not rejected. However, if between groups variance is greater than within group variance, the null hypothesis is usually rejected.

ANOVA is useful in determining if there are differences in means between groups. However, if the ANOVA null hypothesis is rejected and the means are not equal, the test statistic does not indicate which groups are different or similar to each other. Many statistical procedures have been developed over time to address this issue. This study investigates three ANOVA post-hoc procedures: Scheffe, Bonferroni, and Tukey's HSD (Honest Significant Difference). These were developed to help control the alpha, the significance level used to reject the null hypothesis. When multiple comparisons of means is performed the overall error rate increases with each test. Therefore, alpha needs to be controlled for multiple comparison procedures.

Scheffe's method is designed to look at any possible linear contrast of the means and still control the overall alpha. That is any and all comparisons of the means is performed. It is a general method of performing multiple comparisons. Bonferroni's method involves performing a series of t-tests and dividing alpha by the number of comparisons being made. Tukey's method is sometimes referred to as the all pairwise comparisons test, because this is what it does. Tukey's compares each mean to the other while controlling the overall alpha. One benefit of Tukey's method is the output can display all groupings that are statistically similar. This study briefly investigated ANOVA with post-hoc Tukey HSD method to determine if smaller watersheds nested within a larger watershed are similar. One problem with ANOVA techniques is that they are dependent upon assumptions of normal data, equality of variance, and univariate outliers. Since not all assumptions held, additional methods were investigated.

Multi-Response Permutation Procedures (MRPP)

Multi-Response Permutation Procedures (MRPP) is a non-parametric procedure that tests for differences between or among groups of sample units based on permutation-based randomization tests of among- and within-group dissimilarities. The groups must be a priori. (Mielke 1991; McCune & Mefford 1999; Peck 2010). MRPP starts off by creating a distance matrix using the distance measure selected. This study used Euclidean distance measure as prescribed by McCune & Mefford (1999) and Zimmerman *et al.* (1985). Alternative distance measures included Sorensen, Relative Sorensen, Jaccard, Chi-squared, and squared Euclidean. Of these, Euclidean is the most frequently used distance measure (McCune & Mefford, 1999). The average distance is calculated for each group resulting in 'within group' average distance. These distances are then weighted by the relative size of the group then randomly assigned to another group. Then new distances are created. This process is repeated numerous times and the resulting distances are approximated by a Pearson type III continuous distribution. The comparison of the observed test statistic to the distribution of test statistics derived from the iterative randomization assess the statistical significance of the observed within-group average distances. A small p-value indicates that the sample units are more similar to each other than would be expected if they belonged to other groups. A 'within group agreement' statistic ('A') is calculated and is an indicator of how similar the sample units within a group are to one another. The agreement statistic has values that range from zero to one. Completely homogeneous communities composed of just one species have an agreement statistic of one, while completely heterogeneous communities have a score of 0. (Peck 2010).

Examples of MRPP

Biondini *et al.* (1985) used MRPP for statistical analysis of species composition, rate of change, and successional rate. Groupings created by MRPP were further interrogated by canonical correspondence analysis and multiple regression techniques.

Zimmerman *et al.* (1985) describe how MRPP is similar to the t-test and one-way ANOVA and how p-value is derived from permutation, thus making distribution requirements needed for other tests unnecessary. MRPP analysis indicated that standing crop phytomass did not differ after one year of burning, but burned sites had statistically higher yield after the second.

MRPP was used to evaluate the effects of logging on species composition in stands with various logging histories (Brown and Gurevitch 2004). Final results indicated that logging history significantly altered species composition, native species presence and abundance were different on logged and unlogged sites.

Altitude proved to be correlated with species composition as well, but that was due to the fact that sites on lower elevations were more likely to be harvested. Logging and plant invasion were strongly and positively correlated when altitude was held constant.

The influences of canopy species and topographic variables on understory species diversity and composition in coniferous forests was performed by Huo et al. (2014). Significant differences in understory vegetation were found based on MRPP. Out of 36 total understory species, four were unique to spruce forests and 14 occurred exclusively in juniper forests.

Multivariate Analysis

Ordination statistical procedures are used by ecologists and other researchers to try to find order in objects found in large sample datasets based on the values of variables associated with these objects (Goodall 1954). This is accomplished by interrogating a multidimensional dataset to find main gradients of variation, then rearranging the objects in the dataspace along each gradient which serve as new axis. For this reason, ordination methods are often referred to as gradient analysis. Often large, complex datasets can be described by two to four gradients or dimensions, which aid in interpretation. Thus, ordination techniques are popular in biological sciences (Paliy and Shankar 2016).

There are two main classes of multivariate analysis: classification and ordination. Classification puts samples into classes, usually hierarchical classes. Classification assumes that the data set or community group has assemblages that fall into discrete groups. Ordination is based on the concept that assemblages exist on a gradient and seeks to summarize data by creating a low-dimensional ordination space where similar objects are plotted close together and dissimilar objects farther apart (Peet 1980). Ordination results can be displayed in two dimensions facilitating interpretation.

Classification Analysis

In forestry, ecology and other biological sciences the objects being grouped are often samples or communities. Almost all forms of classification fall into one of two groups: divisive or agglomerative. Divisive classification divides a large dataset into smaller and smaller groups, while agglomerative systems begin with smaller groups and progressively lump them into the larger dataset (Pielou 1984).

Cluster analysis is one of the most common forms of classification analysis. Clustering reveals associations in data by graphically showing the clusters and the relationship with the data points. Related objects that can be linked together form the same group, while the remaining objects form another

group. Confidence is dependent on levels of similarity within the dataset and the similarity index used (Chahouki 2013).

Ordination Analysis

Gauch (1982) stated that “Ordination primarily endeavors to represent sample and species relationships as faithfully as possible in a low-dimensional space”. Reasons for performing ordination analysis include (Palmer, accessed July, 2017):

- 1) Visualizing multiple dimensions simultaneously is difficult or impossible.
- 2) Multivariate analysis saves time when compared to multiple iterations of individual univariate analysis.
- 3) Identifies low dimensional space(s) that represent important and interpretable gradients.
- 4) Allows user to determine the importance of resulting gradients.
- 5) Population patterns may differ from community patterns.
- 6) Species-environment relationships can be more readily visible from associated graphical results.

Ordination analysis methods are either a form of direct or indirect gradient analysis. With direct gradient analysis, data pertaining to the environment is used after analysis as an interpretive tool. Essentially, indirect gradient analysis asks the species in the dataset, “What are the most important environmental gradients?” In contrast, direct gradient analysis includes ancillary environmental data in a manner similar to regression analysis. This reveals if species composition is related to environmental variables. Direct gradient analysis allows for testing of the null hypothesis that species composition is unrelated to environmental variables (Gauch 1982; ter Braak and Prentis 1988; Palmer 1993; Palmer, accessed July, 2017). For these reasons, only direct ordination methods were considered in this study. These included: canonical correspondence analysis (CCA), principle components analysis (PCA), and non-metric multidimensional scaling (NMS). The following section explores these three methods.

Non-metric Multidimensional Scaling (NMS)

Non-metric multidimensional scaling (NMS) is a relatively new free ordination tool that is becoming more popular in species datasets. Peck (2010) suggests that it is even beginning to replace CCA as the “tool of choice for ecological analysis.” The reasons for this are many. First, NMS is truly a non-parametric tool. NMS handles data that is nonnormal, arbitrary, unimodal, linear, discontinuous, and/or has questionable scales (McCune and Mefford 1999; Peck

2010). It is useful when communities exist on a long gradient. Paliy and Shankar (2016) indicate that NMS is one of the best analysis techniques when seeking similarities among objects using distance based calculations. Also, NMS is useful when searching for patterns within a matrix of multiple responses. No specific sample design is required, and NMS can be used to explore the structure of almost any dataset (Peck 2010). For these reasons, NMS was selected to be one of the primary multivariate analysis tools used in this study. The following section describes the technique.

As mentioned earlier, PCA has a horseshoe effect flaw that comes from the fact that PCA maximizes linear correlations. NMS solves this by maximizing the rank order correlation. This tends to linearize relationships that exist between environmental and sociological distances (Palmer 2017; McCune & Mefford 1999). One disadvantage of NMS is that it is distance-based, and this masks all information about tree species as it is hidden in the distance matrix. Because NMS is numerical and the species are hidden, the results of NMS are not unique. A “stress” parameter is produced that measures the lack of fit between object distances within the matrix (Peck 2010).

NMS is an iterative search and ranking process where the placement of n entities on k dimensions (axis) is accomplished in a way to minimize stress of the k -dimensional configuration. The final output is an NMS ordination score that has the same rank as the original data matrix. Once this has been achieved, NMS works backward to create a final ordination score. These final scores are then refined through an iterative process until their rank order matches the original rank as close as possible.

Palmer (2017) describes the NMS procedure in six steps.

- 1) The user selects the number of dimensions (N) for the solution, and chooses an appropriate distance metric.
- 2) The distance matrix is calculated.
- 3) An initial configuration of samples in N dimensions is selected. This configuration can be random, though the chances of reaching the correct solution are enhanced if the configuration is derived from another ordination method.
- 4) A measure of ‘stress’ (mismatch between the rank order of distances in the data, and the rank order of distances in the ordination) is calculated
- 5) The samples are moved slightly in a direction that decreases the stress.
- 6) 4 and 5 are repeated until ‘stress’ appears to reach a minimum. The final configuration of points may be rotated if desired.

Palmer goes on to state that the final solution is a configuration of points. This pattern is dependent upon the number of dimensions selected. For this reason, it is important to note that the first two axis of a 3-dimensional solution will not be the same as a 2-dimensional solution.

Another final output of NMS is an NMS score. NMS scores represent the stress derived from increasingly refining placements of each new sample unit on each calibration axis. Stress values less than 10 are associated with interpretable and reliable ordination (Peck 2010). A file containing the variable list, axis, and stress for each variable within the axis is produced. This file can be sorted by axis to look for patterns within the axis. Additionally, 2- and 3-dimensional plots can be created to help visualize these relationships. For this study, NMS was performed on all watershed sizes. Only one provided a reliable ordination. The final NMS score and resulting files and plots will be discussed in the results section.

Examples of NMS

NMS was used to investigate the relationship between climate and deforestation in Puerto Rico and the U.S. Virgin Islands (Brandeis et al. 2009). The authors used PC-Ord version 5.0 using autopilot mode to find the best dimensionality and starting coordinates. A three dimensional ordination axis solution was found for each of the three island configurations studied. Only environmental variables were found significant to be significant. For example, the all-islands ordination axis-1 was negatively correlated with spring and annual moisture stress and positively correlated to mean monthly precipitation in April and May and distance from coast. Axis-2 was positively correlated with minimum temperatures in March, January and November and negatively correlated to higher elevations. Axis-3 was negatively correlated with mean maximum temperatures in September and June and positively correlated with elevation. Species assemblages were identified as being positively or negatively correlated with each axis.

Davies (2007) applied NMS to determine if any correlation existed between species composition and environmental variables. PC-ORD version 4.25 was set for a random starting location and Sorensen's distance measure were used in conjunction with NMS autopilot set for slow and thorough. A Monte Carlo Test of 30 runs yielded a 2-dimensional solution, where the stress for each axis was 27.32 and 0.00001. The first axis captured 41.9% of the dataset variation and the second accounted for 31.8%. The two axis in total accounted for 73.7% of the variance in the dataset.

Sidehurst *et al.* (2012) performed NMS analysis using PC-OD version 5 to investigate tree species composition and abundance of exotics. They used tree

species basal area as the measure of abundance in the NMS analysis. Environmental variables included landuse, overstory tree species diversity, stand age, geologic substrate, soil texture and properties, and absence/presence of exotic species. A random starting configuration was used along a Sorensen (Bray-Curtis) distance measure. A three dimensional solution was recommended, and the final solution. The final ordination graph revealed clear clustering of plots by forest fragment.

NMS using the Bray-Curtis index and Sorensen's distance measure was used by Meng et al. (2013) to look for relationships between the distributions of the longhorn beetle (Coleoptera cerambycidae), tree diversity, and changing landscapes. This ordination was performed by PC-ORD software. The NMS procedure was set for a maximum of 500 iterations and random starting coordinates. A total of 12 vegetation and land use assemblages were included in the model. The final three dimensional solution accounted for 78.5% of the variation. Axis 1 accounted for 41.0% of the variation, while the other two accounted for 19.2% and 18.4%, respectively. Axis 1 was positively correlated with vegetation height, canopy coverage, succession age, and variables representing number of trees and species assemblages. Axis 1 was negatively correlated with ground coverage and number of grass and forges species found at the sample sights.

In 2008, Flinn et al. used NMS to develop a three axes solution that represented 77% of the variance in plant species composition. The three axis accounted for 33%, 27%, and 17% of the dataset variability respectively. This ordination was performed by PC-ORD software using a Sorensen (Bray-Curtis) similarities. Fifty iterations of real data and 50 iterations of randomized data were compared to select a dimensionality. 500 iterations were used to find a stable solution. The authors concluded that species assemblages comprised of wetland species make disproportionately important contributions to landscape-level diversity.

Classification and Regression Trees (CART)

Analyzing ecological data can be difficult for numerous reasons. Datasets are often very large, unbalanced and can contain numerous missing values. Relationships can be nonlinear, involving high-order interactions among explanatory variables. A process that solves these and many other issues in a manner that is easy to implement and understand while producing easily interpretable results is needed. Prediction trees are well suited for ecological analysis because they accomplish all of these goals with a single tool. Prediction trees were developed by Brieman et al. (1984). There are two types of prediction trees: classification trees for when the target variable is discrete and regression

trees for when the target variable is continuous (De'Ath and Fabricius 1999; Moisen 2008; Loh 2011).

Linear regression and many other forms of analysis used in ecological studies are often forms of global models where there is a single response variable. This response is the basis of a formula that contains many explanatory variables that interact in complex, nonlinear fashions. Prediction trees approach the problem in a different manner. The data is subdivided or partitioned into smaller boxes. Creating these splits is analogous to variable selection in regression analysis. The split occurs on a single explanatory variable with the goal of maximizing homogeneity between the two groups. This is referred to as binary partitioning because it is always a two-way split. The point of the split is called a node. A parent node is always split into two child nodes. Each of the resulting child nodes are then subjected to the same splitting criteria as the parent node. Thus each child node becomes a parent. The term recursive is used to describe this continuous splitting. Recursive partitioning is the phrase used to describe the prediction tree process (De'Ath and Fabricius 1999; Moisen 2008; Schmidhammer 2009; and Loh 2011).

A typical prediction tree is displayed growing upside down. The root is the initial split. Each node or split is can be viewed as a leaf or branch. The tree will “grow” downward as more leaves are created. Ultimately, an observation passes down through the tree and its series of decision splits until it resides in a terminal node composed of similar observations that form a homogenous group. The goal of prediction trees is to partition the response into the groups than maximize homogeneity (or minimize impurity), but also keep the tree size small (De'Ath & Fabricius 1999; Moisen 2008).

The advantages of prediction trees include:

1. Suitable for non-parametric, non-normal data.
2. Can handle missing values
3. Computations are simple, just require huge numbers of calculations
4. Makes computations fast
5. Results are easily interpretable

Several measures of impurity can be used to determine the splitting criteria. Classification trees for discrete or categorical data typically use either the Gini index or the Twoing index. Regression trees built on continuous targets typically use either least-squared deviation or least absolute deviations. This study investigates continuous data and used the least-squared deviation method for splitting criteria. However, a discussion of the Gini index is included as it aids in understanding the concept of seeking impurity.

The general formula for the Gini index is:

$$g(t) = \sum_{j=1}^k p(j|t)p(i|t)$$

where

$$p(j|t) = \frac{p(j,t)}{p(t)}$$

$$p(j,t) = \frac{\pi(j)N_j(t)}{n_j}$$

$$p(t) = \sum_{j=1}^k p(j,t)$$

$\pi(j)$ = prior value for category j

The maximum Gini index value is $1-1/k$, where k is the number of categories. The minimum Gini index value is zero, as this occurs when all cases occur in a single category. A successful input variable split is going to drive observations into a single category. The more an input variable accomplishes this, the greater the change in the Gini index from the root node to the leaf. A higher Gini index denotes greater purity. Thus, the variable that has a split with the highest Gini index is used as the node. What Gini is attempting to do is search for the largest class within the target and seek a split that will isolate it from the other categories. A perfect split would end up with k pure child nodes, one for each of the target categories (Schmidhammer, 2009).

The Gini index is not used for targets measured on a continuous scale. A regression tree is similar to Gini, except the measure for impurity used is the least-squared deviation (LSD). The LSD measure is the weighted within-node variance for node t . The splits are chosen to minimize the sum of squared error between the observations on the mean of each node (Loh, 2011). The LSD formula is the following:

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} W_n f_n [y_i - \bar{y}(t)]^2$$

where

$N_w(t)$ = (weighted) number of cases in node t
 w_n = value of the weighting variable for case i (if any)
 y_i = value of the target variable
 f_n = value of the frequency variable (if any)
 $\bar{y}(t)$ = weighted mean for node t

One problem with prediction trees is that they grow large. This can result in having just one observation in each leaf. This is an indication of overfitting the model, which can lead to poor predictions on independent datasets (Moisen 2008). Breiman et al. (1984) introduced the idea of tree pruning to overcome this and to find the best tree. Pruning a tree removes “weak” branches, i.e. a branch with a high misclassification rate. Pruning the tree will increase the overall error rate for the training set, but the pruned tree will provide better predictive power. The goal is to prune the branches with the least additional predictive power (Schmidhammer 2009).

One pruning method is the v -fold validation procedure. In the v -fold cross validation procedure, all the data is used to fit an initial tree. The data is then divided into subgroups, usually 10. The model is then fitted with each subset removed in turn, building the tree from the remaining groups and using the model to predict the responses from the omitted subsets. An estimated error is computed after each iteration. These 10 subgroup tests are then combined to produce an independent error rate for the initial large tree. The subtree with the smallest error rate is chosen as the optimal tree size (De'ath & Fabricius 2000).

All classification trees created in this study were derived with the “rpart” package in R-Studio version 1.0.136. The v -fold cross validation method and interrogation of resulting scree plots were the primary pruning method used when pruning was needed.

Examples of Regression Trees

Fang et al. (2016) used regression trees to test if environmental heterogeneity was responsible for the distribution of deciduous and evergreen trees in evergreen broad-leaved forests. Vegetation and environmental variables were collected from a study site using grid cells of three different sizes. Along with vegetation and tree species, environmental variables such as elevation, slope, aspect, soil pH, and total soil N and P concentrations were gathered. The authors used multivariate regression trees to analyze different habitat requirements. A cross-validation relative error method was used to select optimum tree size. The study concluded that elevation and soil phosphorus were the main variables associated with species aggregation.

The relationship between topography and species assemblages was investigated by Wang et al. (2017). A 35-ha plot was divided into 625, 20 meter square plots. From each plot topographic and tree species data was collected. Topographic variables included elevation, slope, convexity, aspect, topographic wetness, and vertical distance to stream channels. Multivariate regression trees were used to group areas with similar species composition based on topographic

variables. The trees were pruned using the cross-validation procedure. Resulting regression trees partitioned the trees into five types. Habitat type was mainly determined by four topographic variables. These were convexity, elevation, TWI (ratio of area upslope from any point on the landscape) and VDC (vertical distance from the channel network).

Salmon et al. (2016) utilized regression trees to determine the best silvicultural method to use to restore forests to a more historic condition. Data from 50 sites across northwestern and central New Brunswick, Canada were collected. These included harvested stands from four ecoregions. Recursive partitioning in the form of regression trees were used to model relative abundance and species dominance responses at the plot level. Species composition and dominance was linked to harvest/treatment type and amount of residual stand left after harvest.

Species Area Curves

The species-area relationship (SAR) plays a vital role in ecological theory. In its most basic form, this concept can be used to answer the question “how many species are found in a given area?” Larger areas tend to contain more species than smaller areas. Arrhenius (1921) defined the relationship to be $S = cA^z$, where S is the number of species found in region A . Additionally, species-area relationships form a key component of almost all theories pertaining to community ecology (Plotkin et al. 2000; Peck 2010; Thompson et al. 2014). For example, Rosenzweig (1995) investigated a wide range of factors including slope and elevation as they pertain to SAR. Species-area relationships are a key concept behind many extinction-immigration studies (MacArthur and Wilson, 1967), predator-prey relationships (Brose et al. 2004), and impacts of disturbance and scale on SAR (MacArthur & Wilson 1967).

This uses SAR to determine the optimal number of FIA plots that can be used for species diversity studies. This is performed by creating species-area curves using the Summary/Species Area Curves tool in PC-Ord version 4.25. The resulting output displayed the number of species accumulated as number of FIA plots increases. Two different curves were fitted, one for plots that fall on National Forest Land and another for plots that exist on lands owned by forest industry. This will determine if management goals and timber management practices impacted the shape of the curves. Additionally, curves of percent of basal area by the top 20 species will be produced to visualize the differences in species composition for these two land owner groups.

Spatial Autocorrelation

Ecological data such as species distributions and plant communities often occur in clusters or gradients along the landscape. Spatial autocorrelation (SAC) is a measure of this clustering. Spatial autocorrelation in a dataset indicates that pairs of subjects that are close together are more similar to one another than a pair that are more distant (Zuur et al. 2009). This phenomenon was identified by Tobler in 1970 with his first law of geography, which states: “everything is related to everything else, but near things are more related than distant things” (Tobler 1970).

Autocorrelation means that a variable is correlated with itself. The simplest definition is a verification of Tobler’s law; two pairs of subjects that are close together are more likely to have the same values or be more similar, and pairs that are far apart are probably going to be dissimilar. This leads to a spatial structure of the dataset resulting in patterns that one would not find if SAC was not present and the data was truly random. Positive SAC indicates that clustering is present, and negative SAC will be revealed through a checkerboard pattern where observations seem to repel each other. When SAC is present it may be possible to predict a value at one location based on a sample from a neighboring location (Zuur et al. 2009).

Autocorrelation can be caused by a response variable having an interaction with itself (univariate) or with other independent variables in the dataset (multivariate). An example of univariate correlation is when all other factors such as soil, precipitation, wind, etc. are not a factor, but a plant can be found in clusters formed by seed fall or coppice from parent trees. Clustering of plants due to soil type, wind and other variables is an example of multivariate SAC. Both univariate and multivariate SAC will produce clusters close together and far apart. While the sources of SAC may differ, the tools for identifying and analyzing them are the same (Legendre and Legendre 1998; Dorman et al. 2007; Dale & Fortin 2009; Zuur et al. 2009).

The consequences of not accounting for SAC is that it may bias parameter estimates and increase the chance of a type I error (falsely rejecting the null hypothesis). This is brought about by the fact that each sample represents a degree of freedom. If the samples are not independent from each other, the degrees of freedom will be overstated, thus increasing the chance of a type I error (Legendre & Legendre 1998).

Global measures of SAC help identify whether a spatial pattern does or does not exist (Scott, 2008). There are three commonly used measures of global SAC. These include: Moran’s I, Getis-Ord General G, and Geary’s C. All three

require a weight matrix to measure “nearness”. All use the basic concept that examines if a polygon or point has neighbors that are similar or different.

Moran's I

The most common measure of detecting global spatial autocorrelation is the Moran's I statistic. Moran was one of the original investigators of SAC (Moran 1947, 1948). Moran's I is a weighted correlation coefficient used to detect departures from randomness such as clusters. It can be used for polygons or points. Moran's I is based on a random permutation procedure where the statistic is calculated multiple times to generate a reference distribution. The obtained statistic is then compared to the reference distribution to obtain a pseudo-significance level (Anselin 2005). The formula for Moran's I follows:

$$I = \frac{N \sum_{h=1}^n \sum_{i=1}^n W_{hi} (x_h - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - \bar{x})^2}$$

Where:

N is the number of observations (points or polygons)
 $\bar{x}$ is the mean of the variable
 X_i is the variable value at a particular location
 X_j is the variable value at another location
 W_{ij} is a weight indexing location of i relative to j

The Zi-score for the Moran's I statistic is computed as follows:

$$Z_i = \frac{I - E[I]}{\sqrt{V[I]}}$$

Where:

$$E[I] = -1 / (n - 1)$$

$$V[I] = E[I^2] - E[I]^2$$

Moran's I takes the form of a correlation coefficient where the mean of a variable is subtracted from each sample value in the numerator. The resulting coefficient has values ranging from -1 to +1. Values from 0 to +1 indicate a positive association, values from 0 to -1 denote a negative association, and a value of 0 indicates that there is no correlation amongst the variables. The null hypothesis is that there is no SAC, i.e. Moran's I = 0. The alternative hypothesis

is that SAC exists. The null hypothesis is rejected if the Z statistic > 1.96 or < -1.96 (Dorman et al. 2007; Scott 2008; Dale & Fortin 2009; ESRI 2017).

Moran's I can also be interpreted as the correlation between a variable and the spatial lag of that variable formed by averaging all the values of neighboring observations. A scatter plot between these two variables (x and $\text{lag-}\bar{x}$) will reveal if SAC is present. A horizontal line indicates no SAC, a sloped line indicates that SAC is present and the greater the slope, the greater the autocorrelation. The slope of the line is Moran's I.

Geary's C

Soon after Moran developed his measure of SAC, Geary developed his (1954). The two are similar, except whereas Moran used the cross-product is based on the deviations from the mean of two locations, Geary used the cross-product of the actual values themselves at each location. Moran's is a measure of similarity, Geary's is a measure of dissimilarity. The formula for Geary's C is:

$$C = \frac{N \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - x_j)^2}{2 \left(\sum_{i=1}^n \sum_{j=1}^n w_{ij} \right) \sum_{i=1}^n (x_i - \bar{x})^2}$$

While Geary's C is used similarly to Moran's I, the interpretation is opposite. Possible values for Geary's C range from 0 to 2. A value of 1 for Geary's C indicates that the data are random, with no SAC. Values less than one indicate positive autocorrelation. Values greater than one are indicative of negative SAC. Thus higher values indicate negative SAC where high values of Moran's I denote positive SAC.

Gettis-Ord General G

In 1992, Getis and Ord devised a type of statistic that describes the extent to which a sample location is surrounded by high, low, or similar values. Thus, these can be used to determine if clustering or spatial autocorrelation exist within the data. The formulas for General G is:

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} x_i x_j}{\sum_{i=1}^n \sum_{j=1}^n x_i x_j}$$

The Zg-score for the General G statistic is computed as:

$$Z_g = \frac{G - E[G]}{\sqrt{V[G]}}$$

Where:

$$E[G] = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}}{n(n-1)}$$

$$V[G] = E[G^2] - E[G]^2$$

The G statistic describes the measure of clustering of like values around a location, regardless of the value at that location. A statistically significant Positive Gz score means that higher/larger values are clustered together than would be expected if the underlying process was random. A statistically significant negative Gz score means that low/smaller values are clustered spatially. General G may be a superior to Moran's I if the process under investigation tends to be random and the goal is to identify a sudden spike of high values (Scott 2008).

Moran's I was the measure used to detect global SAC for this study. It is more widely used, accepted, and more easily interpretable than Geary's C. General G is more useful when it is known that clustering is going to involve either high values or low values, but not both (Scott 2008). As this is not the case, General G was not used in this study. Moran's I for regression residuals is the measure most commonly used to detect spatial autocorrelation. ArcGIS version 10.3.1 was used to compute Moran's I for multiple response and predictive variables in the dataset.

Hot Spot Analysis

Moran's I, Geary's C and General G determine if SAC exists within a data set. However, these techniques do not reveal where or how these groupings of clusters appear spatially. There are two methods used to determine where spatial patterns appear. These two tools are the Getis-Ord Gi* and Local Moran's I.

Getis-Ord Gi (Hotspot Analysis)*

Getis and Ord developed a technique for analyzing high and low clustering within the dataset. The standardized Gi* statistic is (Gettis & Ord 1992; ESRI 2017):

$$G_i^* = \frac{\sum_{j=1}^n W_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{\sum_{j=1}^n w_{i,j}^2 - \left(\sum_{j=1}^n w_{i,j} \right)^2}{n-1}}}$$

Where:

x_j is attributed to the feature j

$w_{i,j}$ is the spatial weight between the feature in i and j

n is equal to the total number of features

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2}$$

The Gi* statistic is a z-score, so no additional computations are required. The Getis-Ord Gi* Hotspot function identifies statistically significant clusters of high values (hot spots) and low values (cold spots). The resulting p-values and z-scores are used to determine if the null hypothesis is rejected or not. The null hypothesis being that neighboring observation are not similar. In order to be statistically significant, the observation needs to be surrounded by others that feature similar high/low values. The larger the z-score, the more intense the clustering (ESRI 2017).

Local Moran's I (Cluster Analysis)

The Local Moran's I was developed by Anselin in 1995 as a local indicator of SAC. Anselin sometimes refers to this as local indicator of spatial autocorrelation (LISA). This statistic provides an indication of significant spatial clustering of similar and values around an observation. The sum of all LISA statistics is proportional to a global indicator of SAC (Anselin, 1995). The Local Moran's I statistic is calculated using the following formula.

$$I_i = \frac{x_i - \bar{X}}{S_i} \sum_{j=1, j \neq i}^n w_{i,j} (x_j - \bar{X})$$

Where:

x_i is an attribute for feature i

$\bar{X}$ is the mean of the corresponding attribute

$w_{i,j}$ is the spatial weight found between feature i and j

$$S_i^2 = \sqrt{\frac{\sum_{j=1, j \neq i}^n (x_j - \bar{X})^2}{n-1}}$$

The z_i –score for the Local Moran’s I statistic is:

$$Z_i = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}}$$

Where:

$$E[I_i] = -\frac{\sum_{j=1, j \neq i}^n w_{ij}}{\sqrt{V[I_i]}}$$

$$V[I_i] = E[I_i^2] - E[I_i]^2$$

Interpretation of I is similar to that of the Getis-Ord Gi*. A positive value indicates that features of similarly high or low values are next to each other. A negative value indicates that an observation has neighbors with dissimilar values. The negative values are useful for outlier analysis.

Getis-Ord Gi* and Local Moran’s I are similar but are typically used for different reasons. The Getis-Ord Hotspot analysis tool is used to find statistically significant hot and cold spots. The Local Moran’s I is used to find spatial outliers (Scott 2008). This study used the Getis-Ord Gi* Hotspot Analysis tool in ArcGIS 10.3 to identify hot spots and cold spots of diversity and predictor variables to aid in seeking correlations and relationships between the two.

Examples of Local Moran’s I (Cluster Analysis)

Gaither et al. 2011 and 2015 used hot spot analysis to investigate the impacts of wildfires and smoke plumes on social vulnerability. The 2011 study

utilized a bivariate LISA (Local Indicator of Spatial Autocorrelation) maps to reveal areas of high and low wildfire risk with areas of high and low social vulnerability. Final conclusion was that wildfire risk was highest in poor areas. The 2015 study produced similar results with smoke plumes and social vulnerability.

Ma et al. (2011) used local Moran's I to detect and visualize clustering of changes in bird species richness at four different spatial scales in the state of New York. Results revealed that a few small areas had a few clusters of negative z-values and large areas had significant positive z-values. This pattern persisted as the spatial scale increased, however, the areas of negative z-values were more prevalent in larger areas. Additionally, they produced hot-spot maps of three key climatic variables at the four scales. These revealed that the variation in rainfall might have different impacts on species richness depending on spatial scale. Similar correlations and patterns were found with maximum and minimum temperatures. Final remarks stated that the finest spatial scale (25 km) detected more locally detailed information and the largest scale (150 km) revealed large scale patterns of bird species richness change.

LISA maps of proportion of mountainous index and land use index were utilized in a Guizhou Province, China by Zhao and Li (2016). The study revealed that counties with high levels of exploitation were clustered around elevation and population.

Model Development

Ordinary Least Squares and Spatial Regression (OLS)

Ordinary least squares (OLS) analysis is used by scientists to model relationships. The goal of OLS is to define the relationship between a dependent variable (y) and a set of explanatory variables (x). The relationship is generally linear in nature. The formula for OLS is: $Y = X\beta + e$. Ordinary least squares is a global model in that it produces one equation for the entire dataset

The OLS process accomplishes two goals. The first is the creation of a properly fitted model that fits predicted values of $\chi\beta$ onto observed values of the explanatory variable y . Discovering how the explanatory variables interact and contribute to the linear relationship of the dependent variable is the second goal (Matthews, 2006). This is accomplished by minimizing the sum of the squared prediction errors (least squares).

The best OLS model is often referred to as the best linear, unbiased estimator (BLUE). BLUE requires certain assumptions and inferences be made

about the population and sample. These assumptions are that the observations are normally distributed, independent and unbiased. If the random error terms have a mean of zero, then there is no systematic bias in the BLUE equation. Additionally, BLUE has random error terms that are uncorrelated and have a constant variance, i.e. they are homoscedastic (Matthews 2006; Ward & Gleitsch 2007).

Ecological processes and datasets often violate the BLUE assumptions. Often times the observations influence each other or are even related. The data is often skewed or bi-modal. One purpose of regression models is to interpret the significance and magnitude of the coefficients of prediction variables. It is important in analyzing spatial data to determine if the relationships revealed are random or linked together by underlying forces. Analysis conducted on spatial data may be incorrect if the variables show have significant spatial autocorrelation (SA) and steps were not made to correct for this SA. The best method to determine if SA is an issue is the Global Moran's I described previously. If Global Moran's I indicates that SA is an issue, then OLS may not be the best modeling process (Matthews 2006; Ward & Gleitsch 2007).

Spatial dependence can be considered either a nuisance or substance. If the goal is to derive proper statistical inference, estimation, hypothesis testing, and prediction, then the presence of SA is a nuisance. This is because the parameters estimated may be biased and incorrect. Including a spatially dependent variable in the regression error term can alleviate some of this nuisance. However, if the aim of the study is to discover if interactions and spatial spillover does occur, then the SA can be considered substantive. The solution for this is to incorporate the structure of the dependence into the model which will allow for correct estimation and interpretation (Anselin 1988, 1990, 1995; Matthews 2006; Ward & Gleditsch 2007; ESRI 2017). A third option for addressing SA is the use of dummy variables representing areas of known similarities in the OLS regression.

Dummy Variables

Scientists of all disciplines have recognized that there is often a large amount of heterogeneity within their study areas. Countries of similar economic wealth or education exist in similar regions. Productive and unproductive forests occur alongside each other. Regions of radioactivity or toxic waste are within a certain distance of a contaminant site. One way to address these issues is to include a dummy variable in the OLS process. These variables represent different regions across the geographic landscape that the researcher believes have similar underlying processes or relationships with each other. The number of dummy variables used is one less than the total number of regions.

Interpretation of the OLS model is dependent upon knowing the specific variable without that does not have a dummy variable. For example, if a simple one variable model is created for six regions, then the OLS would produce a beta for the intercept and the variable in question. These two represent β_0 and β_1 for the unspecified region. All other coefficient estimates indicates the predicted difference of the value in question to the un-specified region.

Assume that a study is investigating the impacts of precipitation on plant growth. These are the only variables in the dataset. However, the scientist believes that there are three ecoregions in the study area and that they have an impact on plant growth. These three ecoregions are the southeastern mixed forest (SEMF), the outer gulf coastal plain (GCP), and eastern broadleaf forest. Dummy variables for southeastern mixed and gulf coastal plain are created. The resulting OLS model would have an intercept, and betas for growth, SEMF, and GCP. The intercept and growth coefficient are the model for broadleaved forest. The coefficient for SEMF and GCP represent the growth of the regions above or below the estimate for the broadleaf forest (Matthews 2006; Schmidhammer 2009).

Spatial Lag Model (SLM)

The spatial lag model (SLM) incorporates the influence of not only the independent variable, but the effect that neighboring attribute values by incorporating a “spatially lagged” variable of y on the right hand side of the regression equation (Anselin, 1988, 1990, 1995). The formula for the SLM model is: $Y = \rho W_y + X\beta + e$, where W_y is the weighted average of neighboring y 's, i.e. the spatially lagged variable and ρ is the spatial autoregressive coefficient. The SLM model primarily controls spatial autocorrelation in the dependent variable. The SLM treats spatial dependence as substance rather than a nuisance. If the SLM model is properly specified, there should be little or no spatial dependence remaining in the residuals (Anselin 1988, 1990, 1995).

The inclusion of the lagged y variable is appropriate in cases where the values of y interact and influence each other. Examples of possible dependence among observations include; the presence of a species is more likely found in a neighboring site due to seed fall or coppice; an individual home value is likely impacted by the value of neighboring houses; a country is more likely to have similar wealth and levels of education to its neighbors do to interactions between the two. The SLM incorporates the influence of unmeasured independent variables plus the effects of neighboring attribute values. This model is often referred to as the simultaneous spatial autoregression model (Anselin 1988; Ward & Gleditsch 2007; Matthews 2009).

OLS produces biased coefficients in cases where spatial lag is the correct model. Therefore, the reason for using a SLM model over OLS is to derive the proper inference on the coefficients of the other covariates in the model. This is because as one y is based on a neighboring y , which is then in turn based on its own value of y , creating a feedback or dependency that is ignored by OLS. The feedback will increase as the size of the data frame increases, and the resulting OLS estimates will not converge on the true values as the number of observations increases. This dependency means that $\hat{\beta}$ no longer measures the treatment effect of χ on y , but is probably biased downward. The lagged model coefficients assess the effect of $\hat{\beta}$ on the observation while controlling for spatial dependence among y . The spatially lagged variable represents that amount of y that can be attributed the value of y in neighboring observations (Anselin 1988; Ward & Gleditsch 2007; Matthews 2009).

Spatial Error Model (SEM)

The spatial error model (SEM) captures the influence of unmeasured independent variables upon each other. Spatial error models describes the extent to which the spatial clustering of a variable is not explained by the measures of independent variables but accounted for the clustering of errors. The formula for SEM is: $Y = X\beta + \lambda Wu + e$, where Wu is the spatial lag of the error terms and λ is the autoregressive coefficient and indicates the extent to which the spatial component of the errors are correlated with nearby observations (Anselin 1988, 1990, 1995; Ward & Gleditsch 2007).

If there is no SA between the errors for neighboring observations, then λ will be 0. If $\lambda \neq 0$ then there is SA between the error terms and the connected observations. There are consequences of failing to address SA between the errors and performing on OLS instead. The OLS estimates would actually be unbiased, but the standard errors of the coefficients would be wrong. This due to the fact that the OLS model depends on an estimate of variance than assumes independent observations. If this assumption is violated, then the OLS estimate of variance will tend to underestimate the actual variance. Additionally, the estimated coefficients may not be efficient or “close” to the true values (Anselin 1988, 1990, 1995; Ward & Gleditsch 2007; Matthews 2009).

Model Comparison and Goodness of Fit

A suite of regression modeling techniques are available for ecological and spatial analysis. These include OLS, OLS with dummy variables, SLM and SEM. What are the advantages and disadvantages of each, and when should one be preferred of the others? If spatial dependence or autocorrelation are not present, then OLS is the preferred model. In instances where the researcher recognizes

a few areas of clustering and can identify and discuss the forces for this clustering and is not interested in investigating these reasons, then OLS with dummy variables is a valid choice.

If SA is an issue, it must be addressed and OLS removed as a modeling option. This due to the fact that one of the key uses of modeling is for specification testing. The presence of SA will produce misspecified models. Of the two spatial regression models, SLM and SEM, the failure to account for spatially lagged dependence is the most serious, as it is similar to an omitting a key variable from the model. The end result would be biased and inconsistent estimates for all coefficients and lead to incorrect inference (Matthews 2006). However, both are serious issues and a procedure for choosing between the two is needed.

One method of choosing which spatial model to use depends upon the researcher's understanding of the area, topic of interest and the data. If it is understood beforehand that the observations influence each other, then the SLM model may be the best choice. This is a form of nuisance dependence and can be seen in studies involving home values or species migration. Spatial error model may be best in studies where it is recognized that the values of the dependent variable are not influenced by each other, but by connecting factors. This would be made evident by a correlation of the error terms.

Baller et al. (2001) illustrated how the presence or absence of SA and the types of SA impact the estimation of dependent variables. They go on to state that the lag model primarily controls SA in the dependent variable, while the error model controls SA in the both residuals. As a result, the error model controls autocorrelation in both the dependent and independent variables. This makes the SEM a more robust and generally better model (Ward and Gleditsch 2007 and Baller et al. 2008).

The criterion for deciding the best fit model for OLS is r-square (R^2) or adjusted R^2 . Sometimes referred to as the coefficient of determination, R^2 measures how close the data are to the regression line. It is a measure of how much variation in the dataset is explained by the model. However, neither of these two is appropriate in comparing different spatial regression models. The Akaike Information Criteria (AIC) is used to measure the model with the best fit. Unlike R^2 , where the larger the estimate the better the model, the smaller the AIC value the better the model. The formula for AIC is: $AIC = 2k + n [\ln(\text{Residual Sum of Squares})]$, where k is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept term. AIC can only be used to compare models with the same dependent variable (Akaike 1974).

Choosing the best regression model

Choosing the appropriate model is easy when either the reasons for clustering are known or spatial dependence is absent from the data. However, in ecological studies, these scenarios are rare. In most studies the reasons for clustering are only partially known and captured by the suite of variables. Additionally, there is autoregression between both the dependent variables and error terms, so neither the SLM or SEM are obvious solutions. This study employed the model decision tree developed by Luc Anselin for use with the spatial analysis software package GeoDa that he developed (Anselin 2005).

The GeoDa decision process begins by performing an Ordinary Least Squares (OLS) analysis on the data. Spatial autocorrelation diagnostics based on Moran's I is included in this process. If this measure is insignificant, then the process can end and the results of the OLS procedure are used. However, if spatial dependence is found to be a problem, then a maximum likelihood regression is performed on either the spatially lagged variable or the spatial error terms. The decision on which model (SLM or SEM) depends on the results of OLS diagnostics (Anselin 2005).

The test diagnostics for determining the best spatial model are based on Monte Carlo simulation experiments that use Lagrange multipliers (LM) of spatial lag and error statistics. The spatial LM-lag and LM-error specifications are related, so that tests against one form of dependence will also have power against the other. The basic approach used in the GeoDa software developed by Anselin uses LM tests that are based on the residuals of OLS regression. Separate tests for LM-lag and LM-error are produced. If the LM-error diagnostics are significant, then a SEM procedure is performed. Likewise, if LM-lag diagnostics are significant, then a SLM procedure is carried out. If both LM-lag and LM-error statistics are significant and reject the null-hypothesis then a more Robust LM-lag and LM-error statistics are produced. Like the non-robust test procedure, the statistic that is significant or most significant will indicate that the matching model should be used in the analysis (Anselin 1998, 2005; Mathews 2007).

All spatial analysis in this study was performed using either ArcGIS version 10.3 or GeoDA version 1.8.14.8. ArcGIS was used for detecting global spatial autocorrelation (Moran's I), hot-spot analysis, and producing general maps. All OLS and spatial regression was performed using GeoDA and the model decision tree developed by Anselin in 2005 was followed. AIC was used to determine the best fit model.

Examples of spatial regression

Poudyal et al. (2009) followed Anselin's 2005 methodology in assessing the impacts of natural resource amenities in human life expectancy production functions. The LM on OLS residuals indicated that a SEM was the appropriate regression. AIC was used to choose the best model. Where the OLS model revealed 22 of 30 variables in the dataset were significant, the SEM model reduced this to 18. The lambda, the spatially autoregressive error lag term, was significant and positive. With a value of 0.13, the lambda indicated that a county experienced a 1.3% increase in life expectancy if surrounding counties increased by 10%. Conclusions were that life expectancy was increased by natural resource amenities, and agencies might see a benefit by introducing outdoor recreational opportunities.

Zhang et al. (2009) used GeoDa and the spatial regression decision tree to create SLM and SEM models that investigated clustering of trees in stands. Local Moran's I was used to map out "hot spots" before the models were developed. Regression diagnostics indicated that SEM was the appropriate model. The SEM fitted the data better than OLS.

Pattanayak and Butry (2005) used the spatial regression procedures to investigate the impacts of ecosystem services on economic benefits to local people in Flores, Indonesia. Both SEM and SLM models were developed and both had better fit and explanatory power over OLS. Independent variables behaved similarly in both models. All estimated models revealed that river baseflow had a positive impact on income and erosion had a negative impact. The SLM suggested that a 5% increase in ecosystem services could increase agricultural profits by \$9-\$11 per household. A 10% increase would increase profits by \$19-\$24.

Geographically Weighted Regression (GWR)

Geographically weighted regression (GWR) is a relatively new process for modeling spatial data and processes (Brunsdon et al, 1996; Fortheringham et al 1996; 1992; 2002; Charlton, 2009). The basic concept of GWR is that estimate parameters are created for each observation. In essence a separate model is made at each sample location. This allows one to explore how the relationship between dependent and independent variables vary with geographic location. GWR uses a search window that is moved from one sample location to the other. A regression model is fitted to that subset of data, where nearby locations are given more weight than one further away. If a data set is composed of 1,000 observations, then GWR will fit 1,000 weighted regression models. The GWR equation is:

$$y_i(\mathbf{u}) = \beta_{0i}(\mathbf{u}) + \beta_{1i}(\mathbf{u})\chi_{1i} + \beta_{2i}(\mathbf{u})\chi_{2i} + \cdots + \beta_{mi}(\mathbf{u})\chi_{mi}$$

Where:

$\beta_{0i}(\mathbf{u})$ = the parameter that describes the relationship around location $\mathbf{u}$

The weights are derived from a method known as a kernel. There are many possible types of kernels, but the most common and the one used in this study has a Gaussian shape. Kernels other than Gaussian can be used, as long as they kernel is 'Gaussian-like' (Charlton, 2009). The formula for the Gaussian kernel is:

$$w_i(\mathbf{u}) = e^{-0.5\left(\frac{d_i(\mathbf{u})}{h}\right)^2}$$

Where:

$w_i(\mathbf{u})$ = the geographical weight of the i th observation in the dataset, relative to the location of $\mathbf{u}$

$d_i(\mathbf{u})$ = some measure of the distance between the i th observation and location $\mathbf{u}$. Distances are generally Euclidean.

h = is a quantity known as the bandwidth

The bandwidth is expressed in the same units as the dataset. As the bandwidth gets larger, the weights will approach unity and the GWR results will be similar to a global OLS model. Charlton (2009) states that the choice of bandwidth has more influence over the fit of the model than kernel shape. If sample points are evenly spaced, then a fixed bandwidth is usually preferred. If sample locations are clustered or unevenly distributed, then an adaptive bandwidth that allows for the same number of sample points for each estimation should be utilized.

For this study, all GWR calculations were performed in ArcGIS 10.3 with an adaptive kernel with a Gaussian spatial context. The bandwidth utilized the Akaike Information Criterion (AICc), which will be detailed below. The reason for the selection of kernel and bandwidth is based on recommendations from the ESRI help webpage and ESRI tutorials, and it was the most frequently referenced method found in the literature.

Several measures of goodness of fit can be utilized in GWR. As with OLS, sample and prediction points coincide, so r^2 and adjusted r^2 can be used. The measure of fit most commonly used in GWR is *corrected* Akaike Information Criterion (Hurvich et al, 1998). The formula for AICc is:

$$AIC_c = 2n\log_e(\hat{\sigma}) + n\log_e(2\pi) + n\left(\frac{n+tr(S)}{n-2-tr(S)}\right)$$

Where:

n = the number of observations
 $\hat{\sigma}$ = is the estimate of standard of deviation of residuals
tr(S) = the trace of the hat matrix.

AIC_c is a measure of the information distance between an unknown ‘true’ model and on that has been fitted. If two or more models are being compared, then an AIC_c value of less than three usually indicates that there is no difference in the models. Charlton (2009) stated that this is a rule of thumb, however more cautious individuals would be wise to consider four. AIC_c can be used to compare models with differing numbers of independent variables and between OLS and GWR models as well.

Outputs of GWR include parameter estimates and associated errors at each sample point, residuals, and standardized residuals. Additionally, each sample locations will have a regression coefficient for variable included in the model. These represent the type and scope of relationship the independent variables have on the dependent variable. The signs associated with the parameters just as important as the size of the parameter. These local regression coefficients can be, and should be, mapped out to reveal spatial patterns and relationships. Mapping the standard errors will provide insight on the presence or absence of spatial autocorrelation.

Examples of GWR

Holloway and Miller (2015) used GWR to explore the relationship between bird species richness and environmental factors, with an emphasis on how spatial scale impacts autocorrelation and relationships in New York State. The study utilized New York Girding Atlas (BBA) data, PRISM elevation and slope data, Enhanced Vegetation index (EVI), elevation and environmental data. OLS and GWR models were created to investigate relationships within the data. The authors selected a fixed kernel with and AIC_c bandwidth was used for all GWR procedures. GWR was found to produce better results based on R². They concluded that there are local differences in the relationships between EVI, elevation, temperature and bird species richness. Mapping the estimated coefficients allowed them to identify these patterns.

Xu et al. (2016) hypothesized that oak diversity is correlated with water availability and temperature. GWR was performed using oak diversity as a dependent variable and potential evapotranspiration (PET), temperature, precipitation, and an aridity index as dependent variables. A fixed kernel method with and AIC designated bandwidth was used for all GWR procedures. They

concluded that low water and cold temperatures had a negative effect on oak diversity.

GWR was used to study the relationship between tree canopy height and species richness of amphibians, birds and mammals on a global scale (Roll et al. 2015). Tree canopy height was derived from light detecting and ranging (LiDAR) at a 1-km resolution. Other dependent variables included mean annual temperature, annual temperature range, temperature of the hottest month, temperature of the coldest month, mean annual precipitation and a net primary productivity. GWR was performed only using single predictors as explanatory variables, as the authors thought that multicollinearity was severe. A fixed kernel with AIC designating bandwidth was used on all GWR models. Results revealed that tree canopy height was not as important in explaining species richness. It was significant, and very important in regional patterns. Net primary productivity explained more variation in the dataset than any other variable.

VITA

Andrew J. Hartsell was born in Dallas, Texas in 1967 and was raised in Grand Prairie, Texas. He graduated from Grand Prairie High School in 1985. He went on to attain a bachelor's degree in the science of forestry from Stephen F. Austin State University (SFASU) in Nacogdoches, Texas. While at SFASU, Mr. Hartsell grew both personally and academically. He recognized that he had an aptitude for understanding ecological processes and had the ability to apply this knowledge in a real time and on site. He developed academic and personal attributes that allowed him to evolve as a forester and academic. Forestry related employment at this time included working on a Christmas tree farm, a greenhouse technician for the School of Forestry, and cruising and marking timber for local forestry consultants.

After graduation in May 1990, Mr. Hartsell was hired by the Forest Inventory and Analysis (FIA) research unit of the U.S. Forest Service to collect field data across the south for the program. Mr. Hartsell moved to the main FIA office in Starkville, Mississippi in August 1991. His initial duties involved editing data and getting equipment and material ready for field crews. After a few months, Mr. Hartsell displayed ability and willingness to program computer using FORTRAN and manage databases with SQL. He moved to a data compilation/programmer position and began compiling FIA data. He progressed through many positions over the next five years, including: computer programmer, data compiler, data request liaison, and report writer. During this time he enrolled at Mississippi State University (MSU) to earn a Master's of Science in Forestry degree. He attained this degree in May 1999. The focus of his degree was forest economics, however, he gained a great deal of knowledge on GIS, silviculture, and statistics. Professionally he became a research forester for the FIA unit in 2000.

Mr. Hartsell moved to Knoxville, TN in 2003 to further advance his career with FIA and the Forest Service. In 2007 he was admitted to the graduate program at the University Of Tennessee. His initial research program centered on developing an online, real-time web-based tool that would facilitate the location of wood using mills and industries. A program and webpage (BioSAT) was created and earned several awards from the industry. As this program phased out, Mr. Hartsell chose to focus his studies on GIS and spatial statistics. This paper is the result of this research. During this time, Mr. Hartsell has maintained full time employment with FIA. Mr. Hartsell has produced over 60 publications and conducted over a hundred presentations and data requests over his career to this date. He currently lives in Maryville, TN with his wife and three children and maintains his full-time research position with FIA.