

## ABSTRACT

OTWELL, DWIGHT WOODARD. Conifer Discrimination in the Sandhills of North Carolina Using High Spectral Resolution Data. (Under the direction of Dr. Heather Cheshire).

We investigated techniques to discriminate long leaf pine (*Pinus palustris*) and loblolly pine (*Pinus taeda*) in 126 band HyMap imagery with a 4 meter spatial resolution. Field assessment provided stand composition information, and crowns of known species were selected in the imagery to represent species types for model construction. A Quadratic Discriminant Analysis used with a likelihood ratio test was able to identify southern yellow pine with a producer's accuracy of 98% and a user's accuracy of 96%. The same test identified loblolly pine with a producer's accuracy of 80% and a user's accuracy of 49%. Longleaf pine identification had a producer's accuracy of 60% and a user's accuracy of 76%. Price of image acquisition, the relatively low accuracy of discrimination between longleaf and loblolly pine crowns, and inherent bias in the approach make this particular method unreliable as an option for targeting potential sites for RCW habitat restoration.

Conifer Discrimination in the Sandhills of North Carolina  
Using High Spectral Resolution  
Data

by  
Dwight W. Otwell

A thesis submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the Degree of  
Master of Science

Natural Resources

Raleigh, North Carolina

2008

APPROVED BY:

---

Dr. Gary Blank

---

Dr. Stacy Nelson

---

Dr. Heather Cheshire  
Chair of Advisory Committee

## **DEDICATION**

This work is dedicated to:

My family, especially my mother, Laura Coble, and my fiancée, Karen Dowling. Without their support, material and mental, I would not have been able to even attempt this project.

Also to my committee members, whose guidance, patience, and encouragement has been invaluable.

Lastly, to my many friends who have helped me in ways large and small my entire school and professional career.

## **BIOGRAPHY**

Dwight Otwell earned a B.S. in Forest Management at North Carolina State University in the spring of 2002. As part of his education, he worked for Weyerhaeuser Inc. as a field forester as part of the University Co-Op program.

Dwight is currently employed by IAVO Research and Scientific as production manager.

## TABLE OF CONTENTS

<b>List of Tables .....</b>	<b>vi</b>
<b>List of Figures.....</b>	<b>vii</b>
<b>Background .....</b>	<b>1</b>
<b>Literature Review .....</b>	<b>5</b>
<u>Electromagnetic Spectrum and Interaction with Vegetation .....</u>	<u>5</u>
<u>Electromagnetic Imagery.....</u>	<u>7</u>
<u>Prior Research.....</u>	<u>9</u>
<b>Materials .....</b>	<b>13</b>
<u>Hymap Data .....</u>	<u>13</u>
<u>Study Area .....</u>	<u>15</u>
<b>Methods.....</b>	<b>18</b>
<u>Overview.....</u>	<u>18</u>
<u>Study Area .....</u>	<u>19</u>
<u>Image Classification.....</u>	<u>19</u>
<u>Field Data.....</u>	<u>20</u>
<u>Spectral Analysis Within Species .....</u>	<u>21</u>
<u>Modeling Cover Types .....</u>	<u>24</u>
<u>Accuracy Assessment .....</u>	<u>25</u>
<b>Results .....</b>	<b>26</b>
<u>Study Area .....</u>	<u>26</u>

<u>Image Classification</u> .....	26
<u>Spectral Analysis Within Species</u> .....	26
<u>Modeling Cover Types</u> .....	27
<u>Accuracy Assessment</u> .....	28
<b>Discussion</b> .....	37
<b>Suggestions for Further Research</b> .....	39
<b>References</b> .....	40
<b>Appendices</b> .....	44
<b>Appendix 1:</b> Eigen Values and Vectors Used to Construct Principal Components .....	45
<b>Appendix 2:</b> Means and Standard errors for Component Variables .....	50

## LIST OF TABLES

<b>Table 1.</b> Results of accuracy assessment with pine species separated .....	30
<b>Table 2.</b> Results of accuracy assessment with pine categories combined into a single category "yellow pine". .....	31

## LIST OF FIGURES

<b>Figure 1.</b> Known and historical range of RCW in the Southeastern United States. This map displays only known and historical range.....	3
<b>Figure 2.</b> Military installations and municipality borders in the NC Sandhills. A joint USFWS and US Army effort is underway to connect the RCW population in Ft. Bragg with the population in Camp MacCall .....	4
<b>Figure 3.</b> CIR 3-band combination of HyMap data used in the project. ....	16
<b>Figure 4.</b> Extent of study area imagery .....	17
<b>Figure 5.</b> Study area image overlaid with Gameland boundaries. We focused on public land that was easily accessible. ....	32
<b>Figure 6.</b> Study area image overlaid with stand boundaries delineated using heads-up digitizing.....	33
<b>Figure 7.</b> "Stacked" spectral data collected from loblolly pine crown within one stand. No outlying data points were found in this example. ....	34
<b>Figure 8.</b> Results from application of cover type models to the May image. ....	35
<b>Figure 9.</b> Results from application of cover type models to the October image.....	36



## Background

Identifying areas suitable for wildlife habitat rehabilitation is an expensive effort and it can be difficult to access some areas or to obtain the permission necessary for ground assessment in target areas. The purpose of this research is to investigate the feasibility of using high spectral resolution hyperspectral imagery to assist in identifying sites suitable for restoring red-cockaded woodpecker (*Picoides borealis*) habitat. Specifically, this research will attempt to differentiate loblolly pine (*Pinus taeda*) from longleaf (*Pinus palustris*) in a HyMap hyperspectral dataset collected in the southern sandhills of North Carolina (Conner, 2000).

The red-cockaded woodpecker, (RCW) is an endangered woodpecker found in mature, open southern yellow pine stands. They prefer mature longleaf pine stands with no midstory in the southeast United States. Nest cavities require one to three years to excavate (Conner, 2000). Historically, the RCW had a population distribution that ranged from Florida in the south, to New Jersey and Maryland in the north, inland as far as Missouri, and west as far as Texas and Oklahoma (Figure 1).

Several local and federal agencies including the US Army Environmental Command, the US Fish and Wildlife Service, and The Nature Conservancy are investigating possible techniques to identify privately owned stands of longleaf pine where fire exclusion has allowed a hardwood component to become established in the midstory. These areas are less suitable for RCW but have a potential to be restored to more desirable habitat. Once sites have been identified, land managers will work with local land owners to implement a management plan that will alter the forest composition

to be attractive to nesting RCW. Selective harvesting, periodic application of prescribed fire, and installation of artificial nest cavities are all techniques that have been successful in converting suitable target areas into attractive RCW habitat. Specific to this project, the US Fish and Wildlife Service, in cooperation with the United States Army is looking for ways to rehabilitate the RCW populations present on military land in North Carolina. Fort Bragg has one of the largest remaining RCW populations in the southeast. Camp MacKall, located to the southeast of Fort Bragg (Figure 2), also has a significant population. By creating suitable habitat between these populations, US Fish and Wildlife personnel from the office in Southern Pines hope to create a genetic link between these two disparate populations. If such a link could be proven, the population would be considered large enough to be stable. This would allow the Army more freedom to carry out training missions in the area, and improve the outlook for RCW in the area.

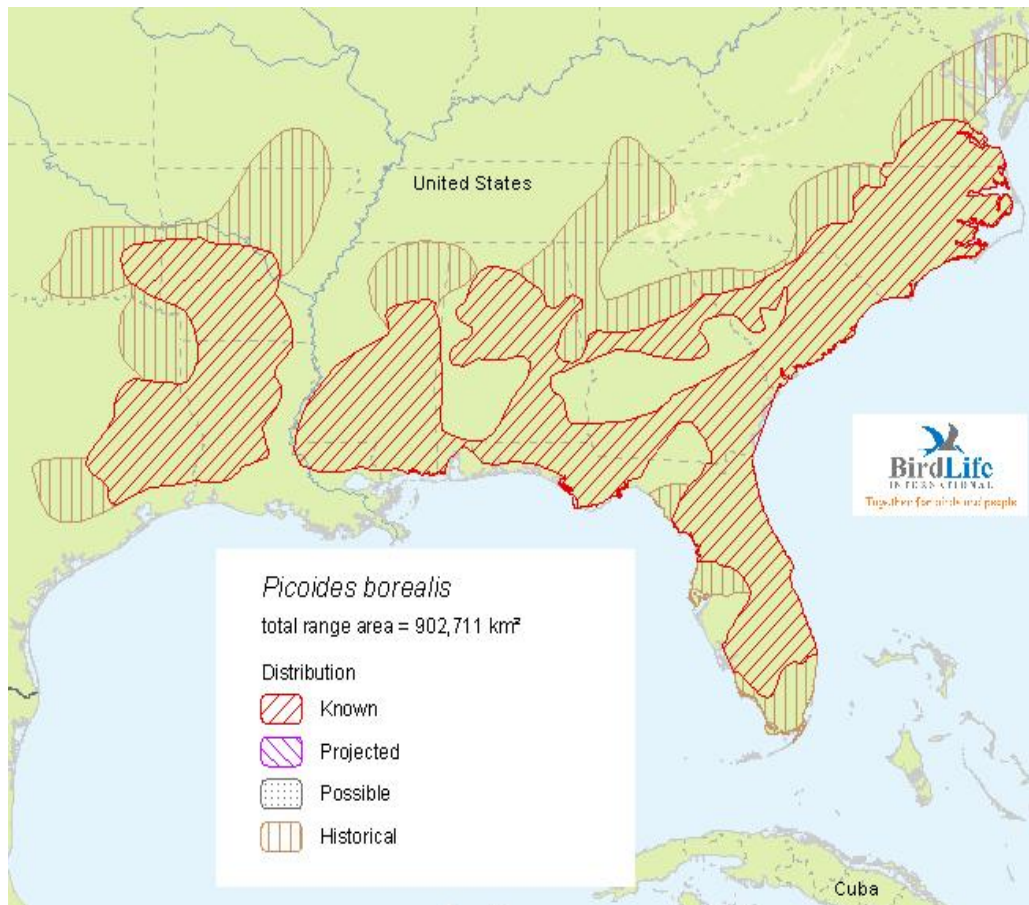


Figure 1. Known and historical range of RCW in the Southeastern United States. This map displays only known and historical range.

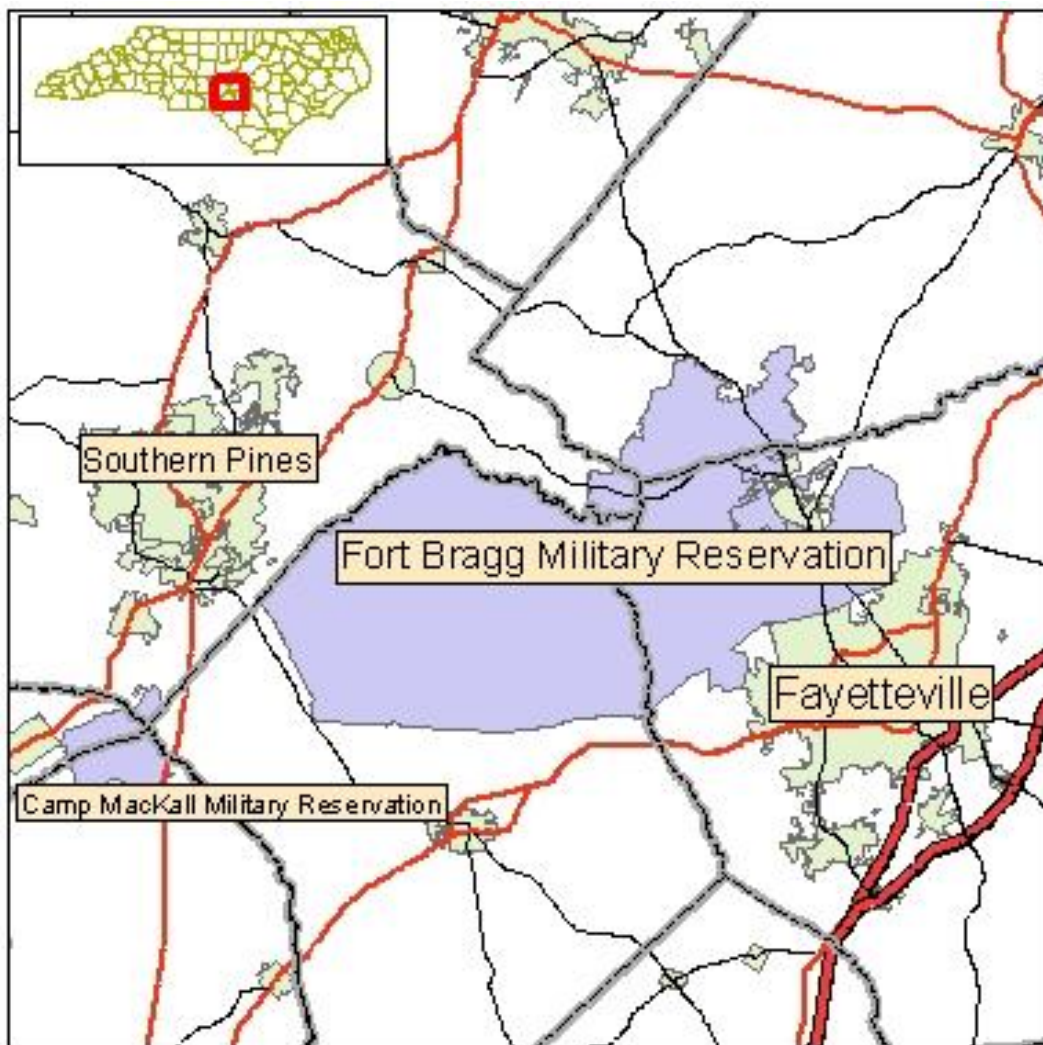


Figure 2. Military installations and municipality borders in the NC sandhills. A joint USFWS and US Army effort is underway to connect the RCW population in Ft. Bragg with the population in Camp MacCall.

## **Literature Review**

Forest stand identification and classification is an important part of any forest management effort. Stand identification traditionally has been accomplished by using field crews and performing a ground survey of the area in question. While this method is accurate, it is time consuming and expensive. Additionally, permission to access private land can be difficult, and sometimes impossible to obtain. It is more cost effective to know which areas might be managed to provide RCW habitat before an attempt is made to contact landowners. Remote sensing gives land managers the potential to study large, inaccessible areas quickly, and for lower cost when compared with field surveys (Govender, et al., 2007).

### Electromagnetic Spectrum and Interaction with Vegetation

To understand how photography and other forms of imagery can be used in species identification, it is necessary to have an understanding of the electromagnetic spectrum and how electromagnetic energy interacts with vegetation.

Electromagnetic energy, or radiation, is a self-propagating wave in space with electric and magnetic components. The electromagnetic spectrum (EMS) is the range of wavelengths of radiation, from gamma rays with a wavelength of 0.003 micrometers to oscillatory circuits with wavelengths up to 10,000 kilometers. The human eye can perceive radiation from 0.4 to 0.7 micrometers, and this is termed the visible portion of the EMS (Jensen, 2007).

The majority of the sun's energy is expressed in the ultra-violet (UV), visible, and infrared wavelengths. Ozone in the upper atmosphere of the earth absorbs most of the UV energy that would otherwise reach the earth's surface. Therefore, the bulk of the energy emitted by the sun that reaches the earth's surface is in the visible and the infrared wavelengths. When radiation comes in contact with an object, that energy wave is absorbed, transmitted, or reflected. Reflected energy waves are what allow us to see objects in the sunlight, in the form of reflected visible light, and reflected visible and infrared radiation can be recorded and displayed in the form of imagery (Crum, 2000).

Different types of objects interact differently with the EMS. Living vegetation appears green because the chlorophyll present in photosynthesizing plants absorbs much of the blue and red light in the visible portion of the EMS, and the green light reflects and hits the eye. The energy gained by the absorption of the other bands provides the energy for photosynthesis (Jensen, 2007). The interaction of radiation with the leaves of a plant is dependent on many factors, like cuticular composition and structure, cellular organization, intracellular air spaces, cytoplasmic inclusions, pigments, water content, emissivity characteristics, and temperature (Liesenberg, 2007).

Pigment molecules, especially chlorophylls, carotenes, and xanthophylls are the main leaf factors affecting reflectance and absorption of radiation in the visible part of the EMS. The main absorption occurs in the blue and red portion of the spectrum. Almost no absorption occurs between the wavelengths of those absorbed by pigments and those absorbed by water. This area is termed the near infrared (NIR) portion of the EMS and is loosely defined as occurring from 0.7 micrometers to 1.3 micrometers (Puritch, 1981).

Due to the high reflectivity in the region, it is of special interest when using spectral reflectance to determine species type. Reflection in this region depends mainly upon the refractive properties of leaf cells and leaf morphology. The greater the number of air pockets within the leaf, the greater the reflection of the leaf (Puritch, 1981).

Different types of vegetation reflect various wavelengths of radiation predictably. Thus, reflectance across a portion of the EMS creates a unique pattern for different kinds of vegetation. This pattern is referred to as a spectral signature, and defining and differentiating spectral signatures allows researchers to identify vegetation and other features of interest by their unique spectral signature (Lewis, 2003).

### Electromagnetic Imagery

Aeronautics and photography first began to converge in the 19<sup>th</sup> century. As with many technological advances, the first uses of these techniques were martial. Aerial observations, and probably photographic images, were made from balloons during the American Civil War, but aerial photo interpretation did not come into wide use in natural resource management until the 1940's, aided by many advancements developed for use in WWII (Crum, 2000).

While black and white photography is useful in stand delineation, color photography can help identify species type. The human eye can separate over 5,000,000 color combinations, but can only differentiate about 200 shades of gray (Puritch, 1981). Black and white film records radiation within the range of human sight, from about 0.36 to 0.72 nanometers, color infrared film (first developed to help identify camouflaged

installations in WWII), extends that range from 0.36 to 1.2 nanometers (Crum, 2000). Interpretation of film-based aerial photography of sufficient spatial resolution can result in accurate stand boundaries and allows for some species identification, but similar species are usually impossible to differentiate using film based photos.

Multi-spectral imaging systems, systems that capture a larger portion of the EMS than what the human eye can perceive, came into common usage through the 1970's and 80's (Crum, 2000). Multi-spectral scanners are sensitive to a specific subset of the electromagnetic spectrum, and detection is recorded in a number of discrete wavelength bands. Sensing range and band size vary from sensor to sensor. Using multi-spectral scanners, greater contrasts can be obtained within the images. Contrasts are usually different in different bands, improving discrimination. Multi-spectral imagery such as Landsat and SPOT provide more spectral separation but identification of specific tree species within the image is often still difficult. Unsuitably coarse spatial resolution in an image results in a large amount of mixing in each pixel. In other words, one pixel represents the mixture of spectral signatures from several different surfaces, and without an accurate representation of a pure pixel, identification is complicated.

Hyperspectral imagery provides higher spectral resolution data, often over a greater range of the EMS (Jensen, 2007). Improved sensing techniques, improved data storage capacity and processing power, and digital imaging technologies over the past three decades, especially since the mid 1990s, has led to an upsurge in the amount of hyperspectral imaging options available (Gong et al., 1997). Narrower bandwidths common in hyperspectral imagery, with each band representing a smaller portion of the



EMS, allow researchers to create an accurate representation of an object's spectral signature, which is not possible with multi-spectral imagers that capture data in much wider bandwidths. Thus differences between data points can be discriminated with much greater accuracy (Puritch, 1981).

### Prior Research

Hyperspectral imagery has proven useful in a large range of scientific endeavors. Bell (2007) used ultraviolet to near IR hyperspectral data to study the composition of minerals on the surface of Mars and confirmed prior research as to the presence and composition of several iron compounds that give Mars its reddish appearance. Vaughan (2005) used a combination of visual, near and far IR and thermal wavelengths from MASTER (Modis/Aster Simulator digital imaging system) and SEBASS (Spatially Enhanced Broadband Array Spectrograph System) datasets to map areas of active geothermal activity, and discriminate silica-rich versus clay-rich areas. Additionally, he identified the presence of several minerals, including opal, quartz, alunite, anorthite, albite, and kaolinite in the area around Steamboat Springs in Colorado. Maathuis, et al. (2004) proved that hyperspectral image analysis provides a feasible method for identifying mine fields and landmines.

Hyperspectral datasets have been useful in determining water quality and turbidity and in identifying submerged aquatic vegetation (Kallio, 2001).

One of the more useful applications of hyperspectral data analysis is the identification of invasive species on a landscape scale, invasive species are often easily

discriminated from established native vegetation in high spectral resolution datasets (Lewis, 2003). Several projects have explored using Compact Airborne Spectrographic Imager (CASI) data to delineate individual tree crowns (Bunting and Lucas, 2005, Leckie et al., 2004).

Puritch (1981) investigated, with varying levels of success, the use of hyperspectral datasets in identifying stress in forest stands before such stresses become visible to the naked eye. He had success in pre-visual identification of water stress and damage from invasive insects. His work relied imagery with very high spatial resolution.

There is a large body of work concerning species discrimination using hyperspectral datasets from environments ranging across the globe. Several studies show that hyperspectral data can be used to accurately discriminate grass species in a wide variety of conditions (Galvao, et al., 2005; Mathur, et al., 2002). Okin, et al. (2001) had less success (less than 30%) in discriminating species types in the southern Californian chaparral forests, though they were able to discriminate soil type with over 90% accuracy. Datt (1999) had over 93% accuracy in discriminating eucalyptus species in New South Wales, Australia, using linear discriminate analysis on hyperspectral imagery.

Several researchers have developed techniques used for vegetation classification using a combination of field acquired and *in situ* spectral measurements (Chen and Lee, 2004; Van Aardt, 2000; Gong, et al., 1997).

There has been less research into discrimination of the types of tree species common in the sandhills of North Carolina. Gong, et al. (1997) had a high accuracy rates

in using *in situ* hyperspectral data to discriminate between conifer species present in the American southwest.

Only four studies prior to 2005 explicitly mention conifer separation using hyperspectral imagery (Van Aardt, 2000). Van Aardt was the first to investigate the use of airborne hyperspectral imagery to differentiate commercial pine species in the southeastern USA. They had varying success (65-85% accuracy) in discriminating loblolly pine, shortleaf pine, and Virginia pine using the Advanced Visual and Infrared Imaging System (AVIRIS). AVIRIS records bands from 400-2500 nm with a 10nm spectral resolution and 3.4m spatial resolution.

Lewis (2003) summarizes several different attempts to use hyperspectral imagery to identify vegetation. Lewis found that earlier descriptive approaches are largely unsuccessful due to the inherent variability of biological material, and the spectral similarity of most plants. The spectral properties of all plants are controlled by the same set of pigments, biochemicals and structures. While Lewis found that descriptive techniques were not very successful, recent studies that have more quantitative approaches to analysis and comparison of spectra have suggested that there are indeed ecological and taxonomic differences in plant leaf spectra (Lewis, 2003). He determined that the most important regions of the spectra for discriminating vegetation types are wavelengths beyond 740nm, the chlorophyll absorption at 680 nm, and the green and other visible wavelengths. He notes that both principal components analysis and discriminant analysis point to broad spectral regions that are important to spectral variation in arid plants, rather than specific narrow bands.

Data sets associated with hyperspectral imagery are typically very large, and can tax the processing power of many computers. When many processes must be performed on imagery, processing time can become a limiting factor for practical uses. Petrie (1998) has studied band selection techniques for use with hyperspectral data sets. He suggests that the number of bands needed to solve a given problem can often be reduced to 10 or less. They looked at three overall methods of reducing the number of bands: expert judgment, mathematical methods, and compression. Expert judgment, while often quite effective, can be difficult to justify, is prone to human error, and is not often appropriate for use in unknown situations. Compression of data often negates the high spectral resolution that makes hyperspectral imagery useful in the first place; therefore, they concentrated on mathematical methods, both spatial and spectral.

One method of aggregating, Optimization with Distance Metrics, requires the user to pick a set of spectra for the target variable and the background, and an algorithm performs a search to determine the best bands. Basic function optimization asserts that “the hyperspectral data set can be represented accurately by a small set of  $n$  basic functions that, when multiplied by the appropriate scalar value and added together, will represent the original data well” (Petrie, 1998).

Van Aardt (2000) used a method of stepwise discriminant analysis that selects spectral bands that minimize within group variance while maximizing the between-group variance for a given alpha level.

Overall the literature supports the idea that species discrimination is possible with the use of hyperspectral imagery, even between species that have similar spectral

response curves. Studies that have concentrated merely on reflectance have not been very successful, as reflectance values are affected by many variables such as seasonality, solar angle, viewing geometry, and the spatial resolution of the sensors (Gong, et al., 1997). Statistical methods and first derivative techniques that examine the slope of the spectral response at critical bands have had much higher success.

The high spectral resolution of the image that we are using, the uniform illumination, and the uniformity of the viewing angle are good indicators that it will be possible to create a process that will satisfy the demands of this project. However, efforts that have shown the most success have had more homogenous study areas than the one in question here. The pixel size of our imagery is larger than the crowns of the trees in the area, and tree densities in longleaf pine stands in the sandhills of North Carolina are low. Complicating matters further, longleaf pine stands have high vegetative diversity in the understory due to numerous grass species and mixed hardwoods that are common in the area. A square meter can contain dozens of species of grasses and shrubs. Spectral signature mixing within pixels may be a major impediment to the success of this project.

## **Materials**

### HyMap Data

The imagery used in this project was acquired for use in Francisco Flores' dissertation; "Using Remote Sensing Data to Estimate Leaf Area Index in Foliar Nitrogen of Loblolly Pine Plantations" (Figure 3). The hyperspectral images were collected using a HyMap sensor, made by Integrated Spectronics Ltd. The spectral coverage ranged

from 0.47 to 2.5  $\mu\text{m}$ , with 126 bands and pixel size in the image was 4.5 m x 4.5 m. The image was provided by Science Applications International Corporation (SAIC), and was collected and calibrated to apparent reflectance and georectified by Analytical Imaging and Geophysics (AIG). The radiative transfer model ATREM (Atmosphere Removal Program) developed by the University of Colorado was used for atmospheric correction and to convert data to reflectance. ATREM is based on water vapor removal on a pixel-by-pixel basis using the 0.94 and 1.14 water vapor absorption bands and accounts for the effects of six gases ( $\text{CO}_2$ ,  $\text{O}_3$ ,  $\text{N}_2\text{O}$ ,  $\text{CO}$ ,  $\text{CH}_4$  and  $\text{O}_2$ ) by assuming a uniform distribution of gasses across the scene. After the data were transformed to reflectance, a spectral smoothing algorithm was performed (Flores, 2003). Data were georectified using a georeference file that was provided by SAIC and created as part of the HyMap data collection process.

Data consist of two data sets that cover the same area. Each image covers approximately ten square kilometers. One was captured on May 8<sup>th</sup>, 2000; the other was acquired on October 22<sup>nd</sup>, 2000. The timing of the image capture does not actually correspond to leaf-on and leaf-off conditions in this geographic area, and in fact, both images are taken with leaf-on conditions on the ground. While both images were taken with leaf-on conditions, Francisco Flores (2003) found that the Leaf Area Index measures are markedly different in the two images.

## Study Area

The area covered by the imagery is located in the southeastern region of NC, in the Sandhills, on the border between Scotland and Richmond counties, just to the east of Rockingham. The Sandhills region lies between the Coastal Plain and the Piedmont in North and South Carolina (Figure 4).

The Sandhills is a region dominated by ancient sand dunes that were on the shore of the Atlantic Ocean when sea levels were at their highest point in recent geological history (Conner, 2000). Soils are sandy, well drained and support vegetation that is tolerant of dry conditions. Species native to the Sandhills of the Carolinas are longleaf pine, loblolly pine, turkey oak and blackjack oak. Historically, frequent (2-3 years) lightning fires suppressed oak and other hardwood species in the uplands, and longleaf pine dominated the region. In the past century, fire exclusion practices have allowed oaks to become dominant in the midstory and overstory.

Most of the land covered by the imagery is private property, making access for field surveys difficult. However, approximately 10% of the area is designated NC Gamelands, which makes it available to the public, and it is suitable for non-destructive field surveys when not in hunting season.

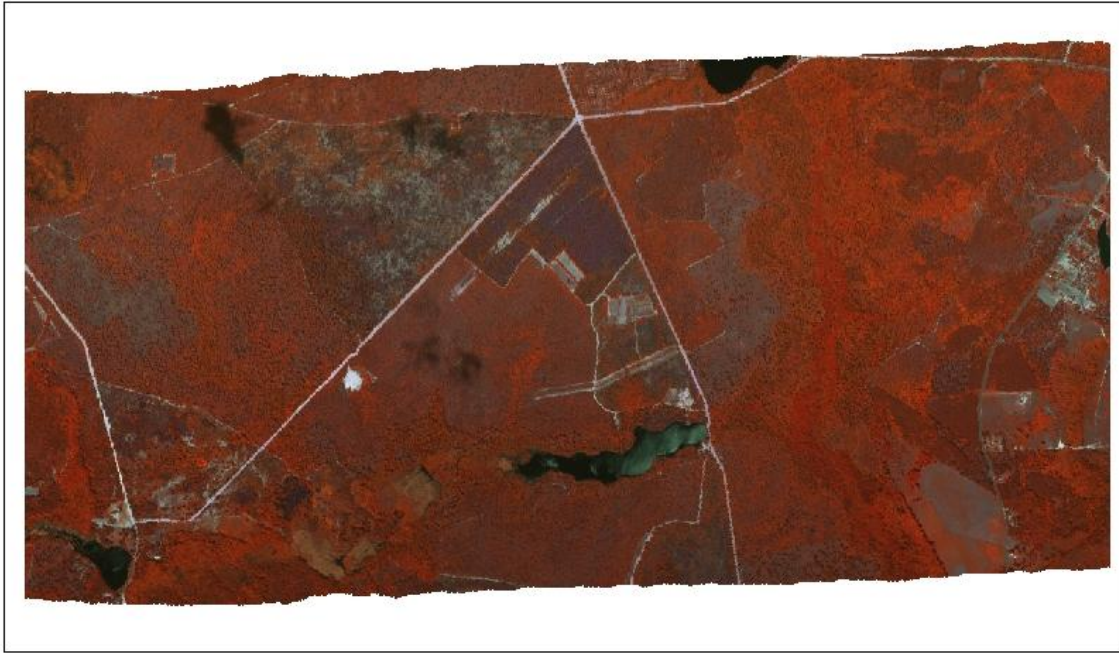


Figure 3. CIR 3-band combination of HyMap data used in the project



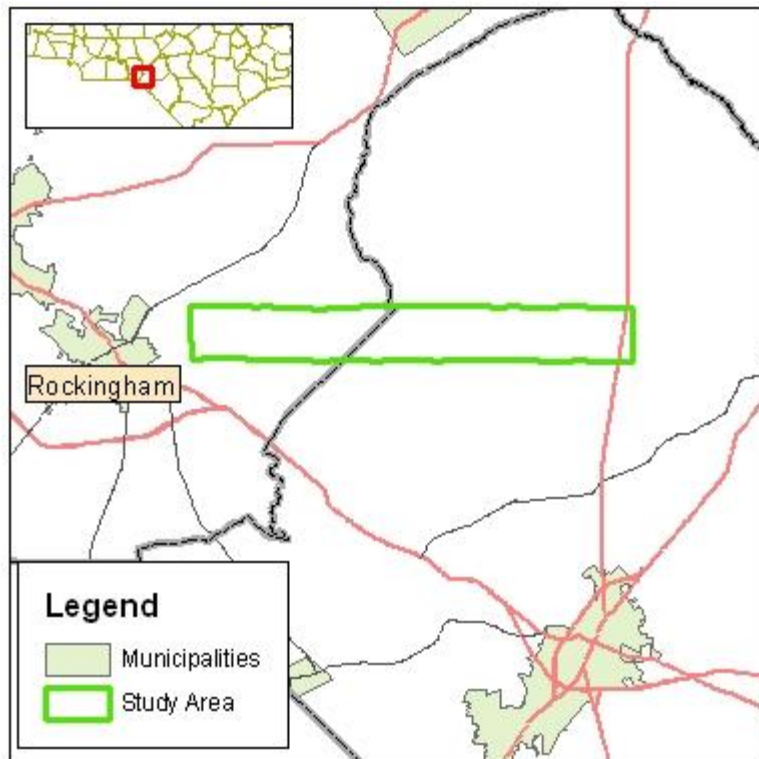


Figure 4. Extent of study area imagery

## **Methods**

### Overview

All image analysis was performed using ENVI (v 4.4, 2007). The images were inspected and displayed in ENVI image analysis software and a study area of appropriate size was selected, based on the availability of ground surveys in the region. Within the study area, we used heads-up digitizing to delineate visually discernible stands greater than 1000 square meters. Field surveys provided information used to classify digitized stands by overstory, midstory, and ground cover composition.

From stands with uniform overstories, pixels representing overstory crowns were randomly selected to provide data to build spectral signature models for five cover types- longleaf pine, loblolly pine, hardwoods, bare earth, and grass. First, we compiled data from pixels within stands, then compared data between stands with the same overstory species. The final step compared different species type against each other.

We used a quadratic discriminant analysis (QDA) to create models representing the spectral signatures of the five overstory cover types of interest. A likelihood ratio test compared the output from these spectral models for each pixel in the dataset, and returned the probability that the pixel was represented by each model. Each pixel was assigned to the type of model with that returned the highest probability.

### Study Area

After reviewing the image datasets, we determined that a test area of approximately 8 square km would be used to collect field data and develop techniques (Figure 4). The primary study area is approximately 4000 meters by 2500 meters, and is located in the center of the imagery. The primary study area has a good mix of forest types and densities, is uninhabited and is public property, making it ideal for field work.

ENVI image analysis software was used to determine the overlap between the May and October HyMap images. This polygon was imported into ESRI ArcMap GIS mapping software (v 9.2, 2007), and compared against published NC State Game Lands boundaries. The overlap between NC Gamelands and the two image dates was used to define the extent of the primary study area.

### Image Classification

The primary study area boundary was then imported back in to ENVI and overlaid on the imagery. Using several different band-display combinations, and both the May and October data, color, texture, and homogeneity were used for heads-up digitization of stand boundaries, with a minimum mapping unit of 1000 square meters. Stand polygons were delineated for portions of the image that fell within the border of NC Gamelands, where interior portions of the stand would be accessible. Stands were delineated into broad categories of cover types: conifer, hardwood, grass, and bare earth.

These classifications were preliminary, to be revised and characterized more specifically when verified in the field.

### Field Data

Field data were collected during the summer of 2004, approximately four years after the imagery was captured. Special care was taken to note any signs of recent disturbance, or any stand-altering events. Forest characterization was achieved using a traversing method. Traverse lines with a spacing of 300 meters, running due east and west, were walked. We used a Trimble GEO III handheld GPS unit (Trimble Systems, Inc.) to record our position in relation to stand boundaries. Along each traverse, we recorded overstory and understory composition and state. Composition included species present in the overstory, understory, and ground cover. State is defined as the approximate age and density of overstory. Overstory density for each stand was measured with a densitometer for 10 representative locations in each stand. Each forest stand was then classified based on average crown closure into one of three groups: low density (less than 50% canopy closure), medium density (between 50 and 75 percent canopy closure), and high density (greater than 75% canopy closure). Each stand was either designated as mature or immature. Mature stands were defined as having an average total height of greater than 20 feet, allowing for the lapse of four years from the time the image was acquired to when the fieldwork was carried out. We assumed there had been growth in stands since 2000. A note was made of the average crown diameter for each stand based on visual estimation. A brief description of the ground cover for

each stand was recorded, including species (identified to genera). Percent dead vegetation and percent open ground were also recorded for each stand, based on visual estimation.

### Spectral Analysis Within Species

Hyperspectral data sets often exhibit excessive noise across a limited range of the EMS. This error in measurement can be attributed to atmospheric conditions or equipment imprecision, and is often found near water absorption bands (Gong, et al., 1997). One of the methods for data reduction revealed in a search of prior research is a visual appraisal of the dataset, and a corresponding elimination of bands with excessive noise (Petrie, 1998). This technique was used to exclude all bands for wavelengths between 1.3038 and 1.4707 micrometers.

Extraction of data from the HyMap imagery was performed using ENVI image analysis software. At the inception of this project, ENVI was the best available software for handling datasets with large numbers of bands.

Our first intention was to accurately describe the spectra for individual overstory species. To do this, spectral data for each band were extracted from individual crowns within stands with homogeneous overstory composition. ENVI allows the user to extract a spectral profile consisting of a reflectance value for each band for each individual pixel. These data could be exported in a number of different formats for use in statistical analysis programs. Spectral profiles may be “stacked” in a process designed to collect large test collections. In this procedure, the user selects a series of pixels, each

representing the crown of an overstory species of interest. Each pixel is then stored as a record in a database, with each record including a value for each band represented in the pixel. ENVI stored the spectral profile for each pixel selected and maintained this list until it was actively cleared. This data set could also be exported as a text file.

Using the spectral profile stacking method, spectra were collected for crowns from every delineated stand with homogenous overstory in the primary study area. The size of the sample window can be set as a function of the number of pixels in the x and y direction, with a minimum of 1 x 1, with no set maximum. Due to the low density of trees in stands under consideration, and the relatively small crowns of the trees, we determined that a single pixel window would suit our study the best. A larger capture window would increase the amount of spectral mixing in the extracted data. A minimum of 70 spectral signatures were collected for each stand, with a maximum of 170, depending on the total size of the stand in question.

The highest canopy closure of all stands were question was 70%, so a random selection of coordinates from such a stand could have approximately 30% of the pixels representing understory and ground return. Spectra collection with EVNI was problematic in that once a spectra was added to the “stack” it could not be removed without clearing the entire stack. Erroneous pixel selections were hard to avoid when compiling over 100 points, and it took a great deal of training to reach a point where datasets could be produced containing spectra from only the desired pixels.

In order to eliminate pixels that represented a significant amount of material besides the target tree species, an outlier identification technique was used to cull

individual spectra that did not appear representative of the dominant species. This process removed spectra from pixels that appeared to have been selected correctly, but in actuality had such a strong mixing from non-target surfaces that it would adversely affect results. We used the Mahalanobis Distance test to identify these outliers. Mahalanobis distance is based on correlations between variables in which different patterns can be identified and analyzed. It is commonly used to determine similarity of an unknown sample set to a known one (Hadi, 1992). As implemented, the Mahalanobis value from a group of values is:

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}.$$

With means  $\mu = (\mu_1, \mu_2, \mu_3, \dots, \mu_p)^T$ , a covariance matrix  $\Sigma$  for a multivariate vector  $x = (x_1, x_2, x_3, \dots, x_p)^T$  for all bands.

A script was written in C++ to implement the Mahalanobis test directly on the stand data exported from ENVI. The Mahalanobis test gives a score to each sample in a dataset; the higher the score for each sample, the greater likelihood that the sample in question is an outlier.

After determining the similarity of within-stand samples, we attempted to verify that species overstories from different stands within a cover type were similar enough to aggregate for use in a comparison of means between species. To accomplish this, we used Hotelling's T-square distribution. Hotelling's T-square is a generalization of Student's t-test that is used for multivariate hypothesis testing (Murtagh and Heck, 1987). Our null hypothesis was that spectra collected from the same species would be equal, i.e.,

the means of bands between different stands with the same overstory species should be equal.

### Modeling Cover Types

The modeling approach that we used was a quadratic discriminate analysis (QDA), a general version of linear discriminant analysis (LDA). Unlike an LDA, a QDA does not assume covariance matrices to be equal. First, a principal components analysis was performed on the dataset. The number of variables was too high to run a QDA on the full dataset, and it was necessary to perform a data reduction. The eigen values and vectors used to produce the principal components are detailed in appendix 1. A total of 8 principal components were used to develop likelihood ratios. This number of variables can be easily handled by STATA, captures 99.93 percent of the variability in the dataset, and is consistent with the number of principal components used in prior studies (Van Aardt, 2000 and Becker and Lusch, 2005).

We used the following likelihood ratio test (Murtagh and Heck, 1987):

Likelihood ratio =

$$\frac{\sqrt{2\pi|\Sigma_{y=1}|}^{-1} \exp\left(-\frac{1}{2}(x - \mu_{y=1})^T \Sigma_{y=1}^{-1} (x - \mu_{y=1})\right)}{\sqrt{2\pi|\Sigma_{y=0}|}^{-1} \exp\left(-\frac{1}{2}(x - \mu_{y=0})^T \Sigma_{y=0}^{-1} (x - \mu_{y=0})\right)} < t$$

Where  $\mu_{y=0}, \mu_{y=1}$  are the means of the components of each cover type, and  $\Sigma_{y=0}, \Sigma_{y=1}$  are the covariances, and  $x$  is the spectral signature for a specific pixel. For every pixel in the study area, the QDA likelihood ratio test determined the probability that the



pixel in question matched each cover type. The algorithm then assigned that pixel to the stand type with the highest probability. As a preliminary step, the models were used to identify the pixels used for training. The analysis correctly identified 99% of the pixels used to develop the model for each type, except loblolly pine. The analysis correctly identified 85% of the pixels used to develop the model for loblolly pine. The QDA likelihood ratio test was then used to assign all pixels in the study area to a cover or forest stand type.

### Accuracy Assessment

A minimum of 50 points in each class were selected from stands with known composition. Using the stand descriptions obtained during the field survey, each pixel used in the accuracy assessment was identified using aerial photography interpretation. Pixels representing classes not targeted by the algorithm (e.g. water) were not used in the assessment.

For each point, we compared photo-interpreted type to the results of classification modeling. Error matrices were constructed for both producer's and user's accuracy.

## **Results**

### Study Area

We selected a study area approximately 2500 by 4000 meters. This area was chosen due to the presence of target species, determined by a cursory field assessment, and the large portion of the area that is managed as State Game Lands (Figure 5). The area is dominated by sandy soils and dry conditions. Pine stands (both loblolly and longleaf) appear to be fire maintained, and midstory is sparse to non-existent in most stands.

Hardwoods dominate the overstory in this area only in drainage areas, where fire has been excluded. These drainage stands provide the majority of stands used to define hardwood cover type.

### Image Classification

Stands were delineated using a heads-up digitizing technique. This gave us a total of 42 distinct stands (Figure 6). Stands were identified using data obtained from a field survey, employing traverse techniques. Of the 42 stands in question, 6 had a pure loblolly pine overstory and 13 had a pure longleaf pine overstory.

### Spectral Analysis Within Species

A minimum of 70 data points were collected from each stand. Very few bands were determined to be outliers by using the Mahalanobis Distance test, and datasets for most stands were accepted in their entirety (Figure 7).

Hotelling's T-square test was used in an attempt to show that stands with the same overstory species are spectrally indistinguishable from each other. The null hypothesis was that stands with overstories composed of the same species would have equal band means. Each stand was tested against every other stand with the same cover species. Only one pairing showed a probability greater than 0.03 of the null hypothesis being true, at 22.46%, still a low enough probability to reject the null hypothesis in that case. Based on our statistical analysis, we were unable to justify aggregating within-species data for a reflectance analysis. Instead, we adopted an analytical approach that concentrates on relationships between variables, instead of total reflectance of individual variables. This allowed us to aggregate same-species datasets to create unique models based on species type, and is a common technique used in hyperspectral species discrimination (Van Aardt, 2000).

### Modeling Cover Types

Using a QDA likelihood ratio test, each pixel was assigned to one of the five model types that had the highest probability of a match. Output from STATA was in the form of a comma-delimited text file, arranged in x and y columns corresponding to the image x and y coordinates, effectively stripping the data of all geographic data.

A C++ script was written to convert the comma-delimited text file into a tagged image file format (TIF) file. The TIF world file associated with the input study area was then associated with the output TIF file, allowing the classification image to be viewed in GIS software for accuracy assessment (Figures 8 and 9).

One of the most striking differences between the images is the apparent increase in amount of hardwood classified from May to October. When the May image was taken, hardwoods were early in their growth cycle for the year, and ground cover hardwood had yet to fully flush out. In October, the hardwoods had completed their growth for the year, but had not yet begun to go dormant, while the pines had begun to shed their needles, resulting in a greater representation of understory hardwood in the later image.

### Accuracy Assessment

Our technique proved very accurate at distinguishing hardwood, grass, bare earth, and yellow pine within the study area. However, it was much less accurate in differentiating longleaf pine from loblolly pine. Our results show good identification of bare earth, hardwood, grass, and pine, and lower accuracy at differentiating between longleaf and loblolly pine. Overall accuracy for the set was 78% for October and 77% for May when pine species are considered separately (Table 1). With pines considered separately, the kappa statistic is 0.72 for May and 0.70 for October. A kappa statistic value between 0.4 and .08 indicates moderate agreement between the reference and classified data (Congalton, 1999).

When pine categories are combined and considered one cover type, modeling results improve with overall accuracies 92% and 93% for May and October, respectively. Kappa value for the classification also increases with pines combined, to 0.89 for both

time periods when longleaf and loblolly are combined (Table 2). In both time periods, grass and bare earth had high classification accuracies.

Table 1. Results of accuracy assessment with pine species separated.

october							
	loblolly	longleaf	hardwood	grass	bare earth	User's	
loblolly	22	29	2	0	0	53	<b>41.51%</b>
longleaf	15	49	1	2	0	67	<b>73.13%</b>
hardwood	3	3	53	8	2	69	<b>76.81%</b>
grass	0	0	1	58	0	59	<b>98.31%</b>
bare earth	0	0	3	0	58	61	<b>95.08%</b>
	40	81	60	68	60	.	
Producer's	<b>55.00%</b>	<b>60.49%</b>	<b>88.33%</b>	<b>85.29%</b>	<b>96.67%</b>		

May							
	loblolly	longleaf	hardwood	grass	bare earth	User's	
loblolly	47	42	4	2	0	95	<b>49.47%</b>
longleaf	11	62	9	0	0	82	<b>75.61%</b>
hardwood	0	1	59	0	0	60	<b>98.33%</b>
grass	1	1	3	44	1	50	<b>88.00%</b>
bare earth	0	0	0	3	51	54	<b>94.44%</b>
	59	106	75	49	52		
Producer's	<b>79.66%</b>	<b>58.49%</b>	<b>78.67%</b>	<b>89.80%</b>	<b>98.08%</b>		

Table 2. Results of accuracy assessment with pine categories combined into a single category “yellow pine”.

october						
	yellow pine	hardwood	grass	bare earth	User's	
yellow pine	115	3	2	0	120	<b>95.83%</b>
hardwood	6	53	8	2	69	<b>76.81%</b>
grass	0	1	58	0	59	<b>98.31%</b>
bare earth	0	3	0	58	61	<b>95.08%</b>
Producer's	121	60	68	60		
	<b>95.04%</b>	<b>88.33%</b>	<b>85.29%</b>	<b>96.67%</b>		

May						
	yellow pine	hardwood	grass	bare earth	User's	
yellow pine	162	13	2	0	177	<b>91.53%</b>
hardwood	1	59	0	0	60	<b>98.33%</b>
grass	2	3	44	1	50	<b>88.00%</b>
bare earth	0	0	3	51	54	<b>94.44%</b>
Producer's	165	75	49	52		
	<b>98.18%</b>	<b>78.67%</b>	<b>89.80%</b>	<b>98.08%</b>		

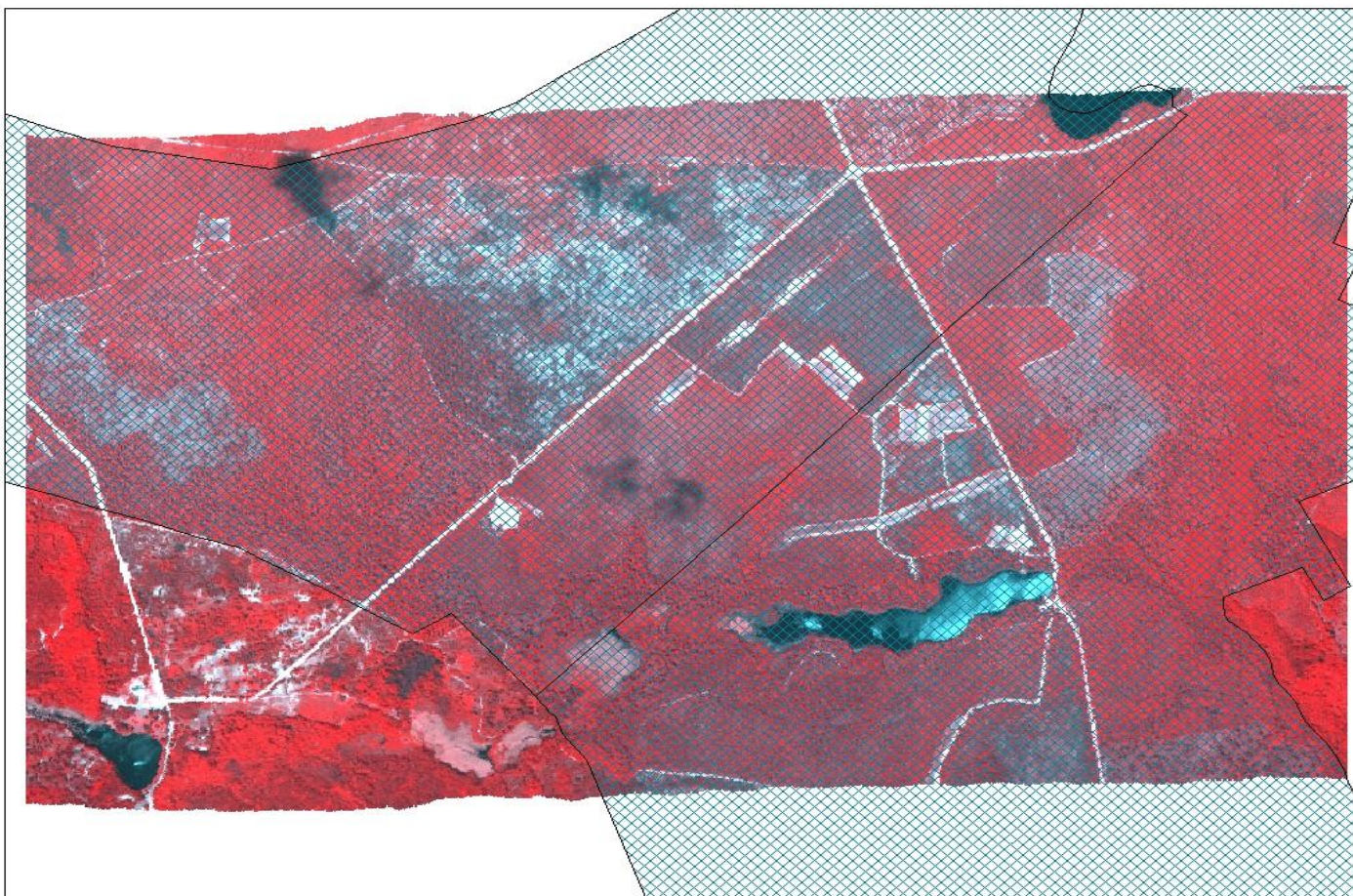


Figure 5. Study area image overlaid with gameland boundaries. We focused on public land that was easily accessible.



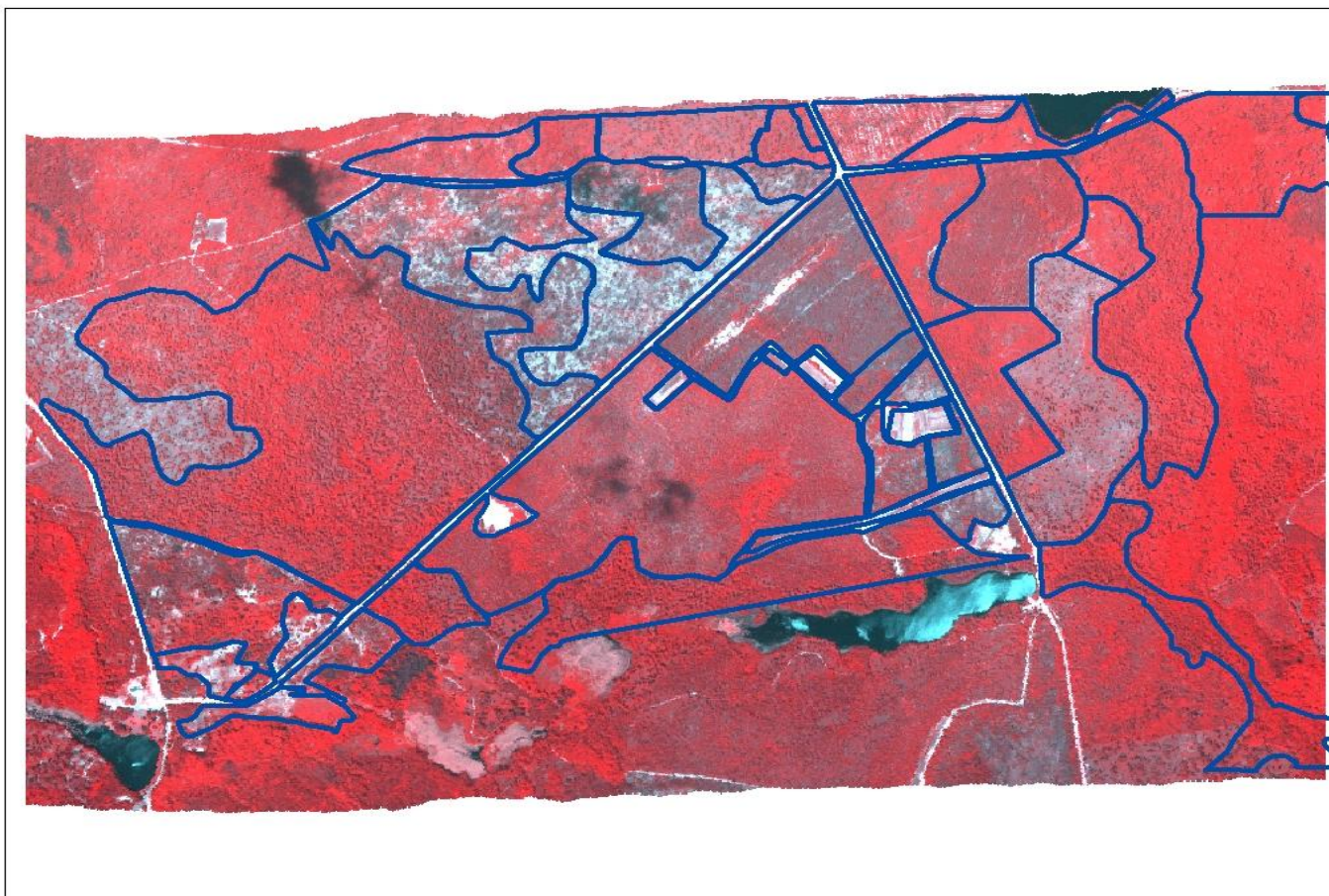


Figure 6. Study area image overlaid with stand boundaries delineated using heads-up digitizing.

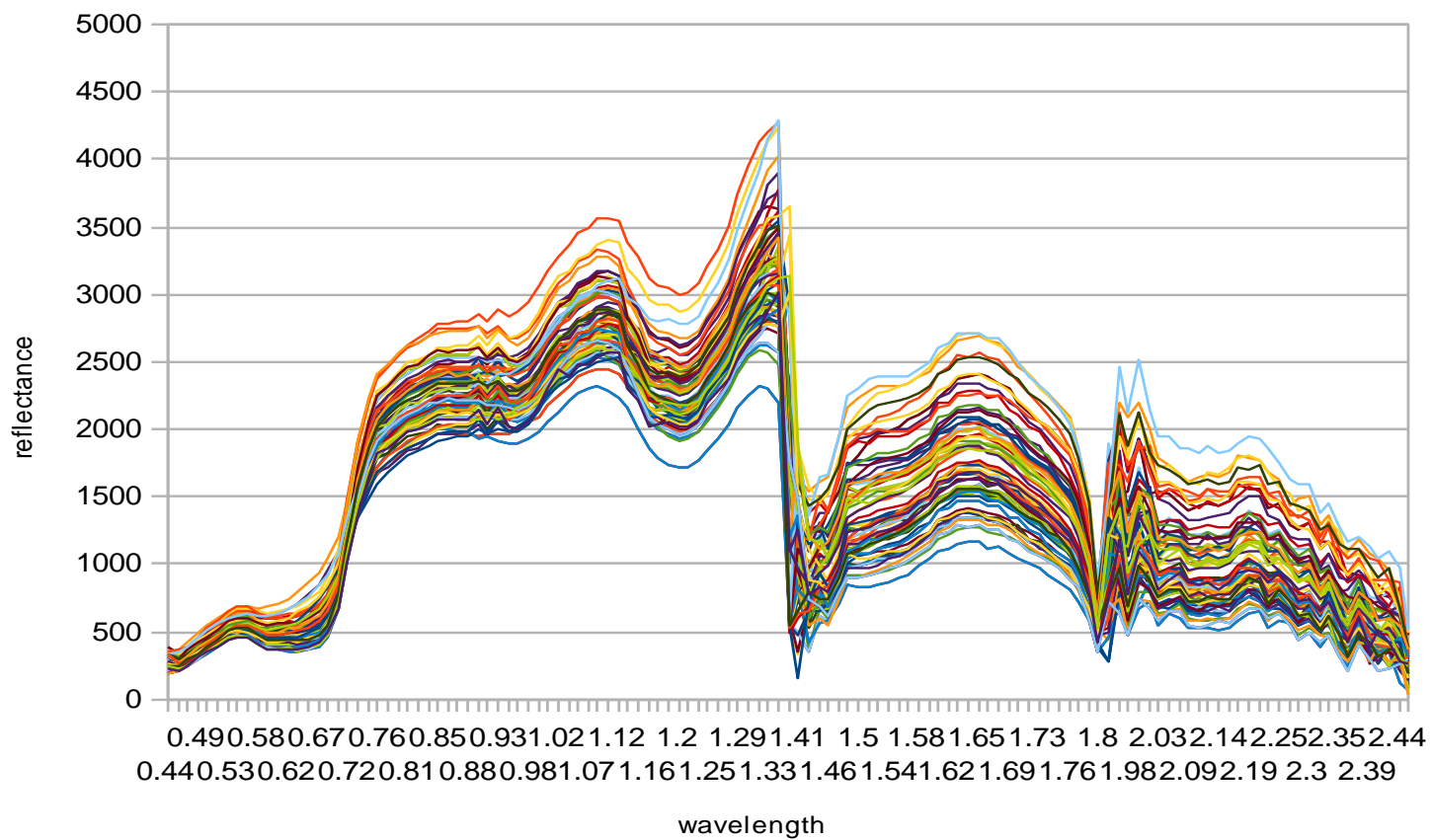


Figure 7. “Stacked” spectral data collected from loblolly pine crowns within one stand. No outlying data points were found in this example.



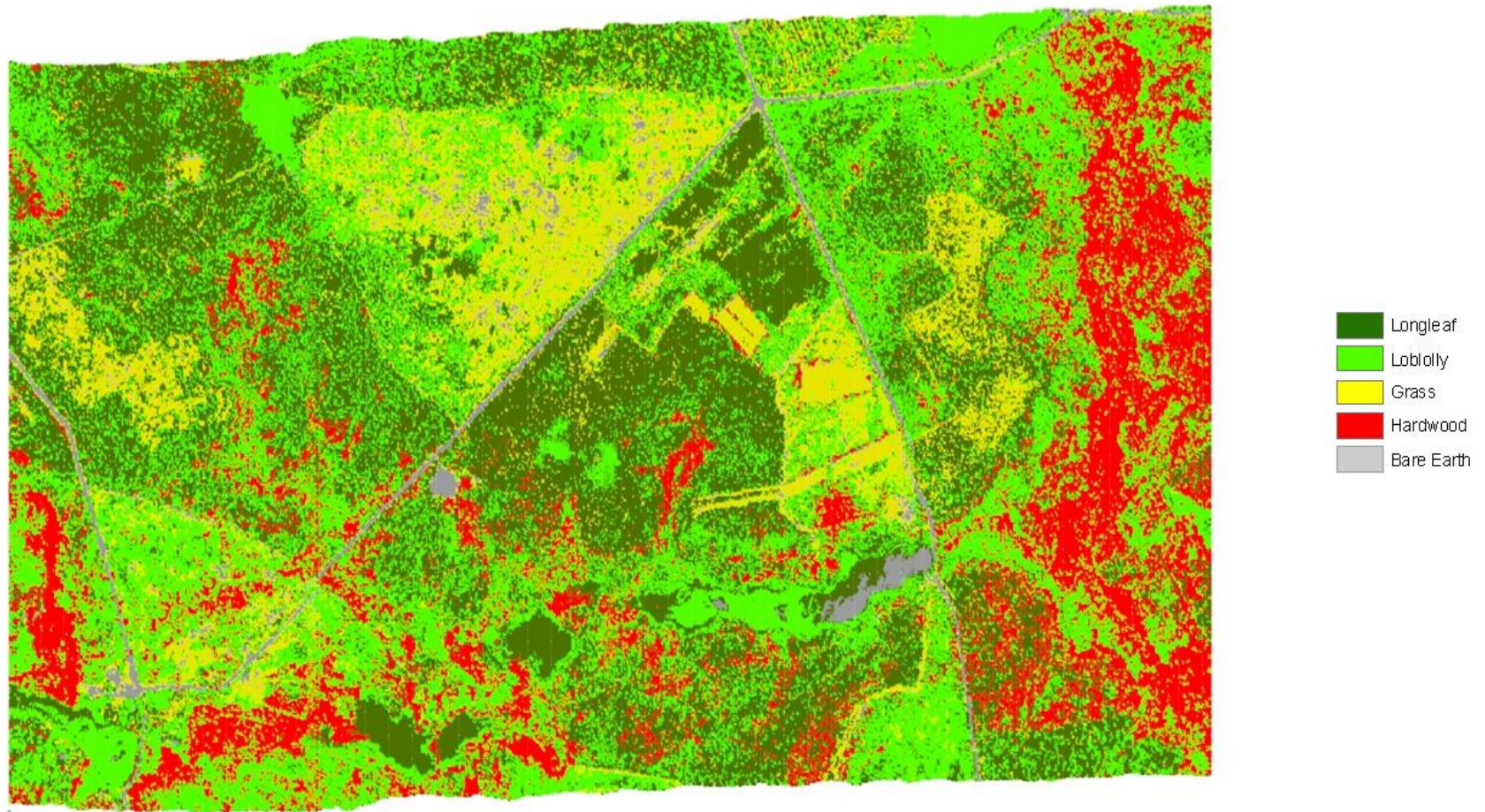


Figure 8. Results from application of cover type models o the May image.



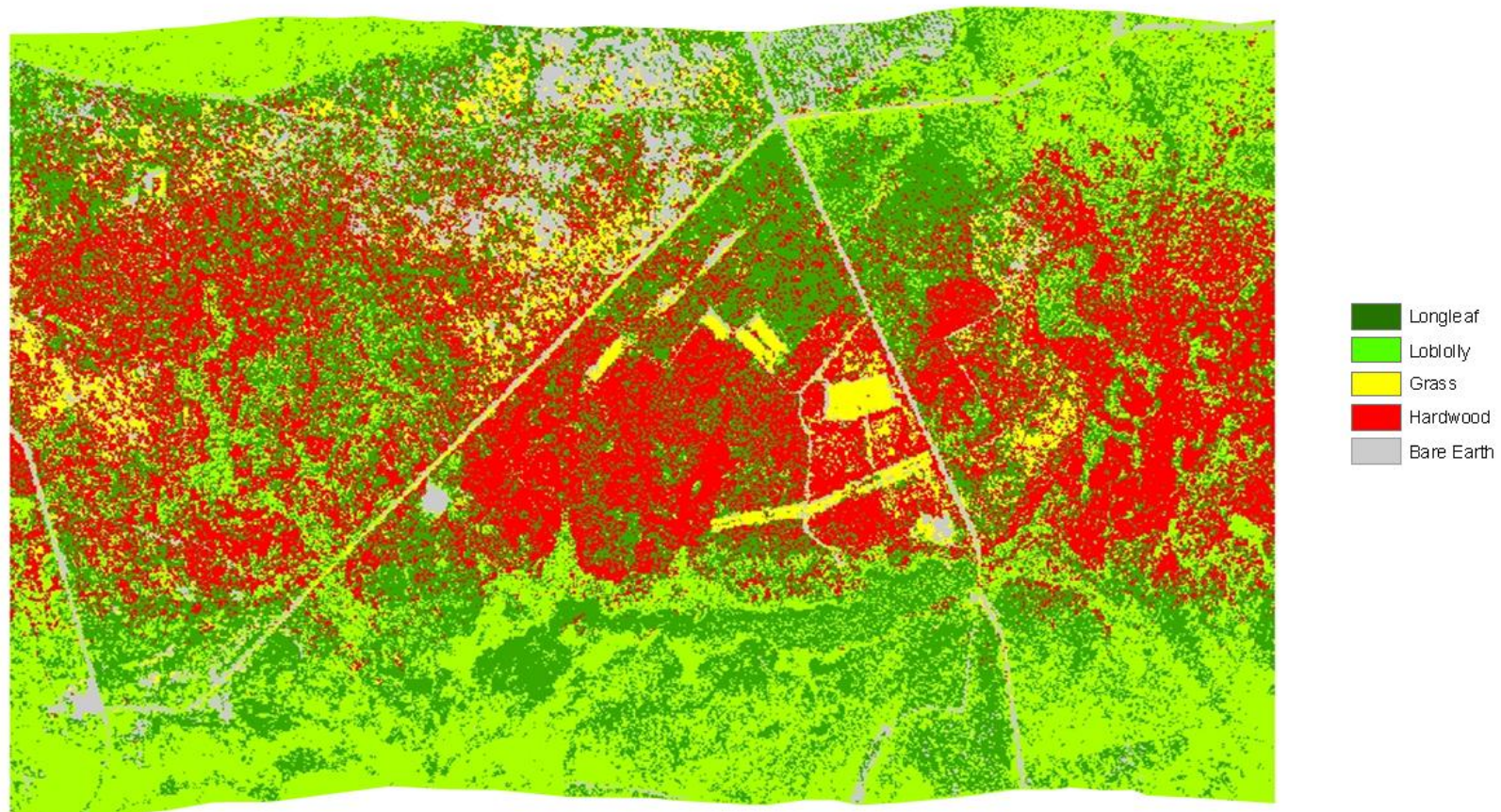


Figure 9. Results from application of cover type models to the October image.

## Discussion

The superior spectral resolution of hyperspectral imagery does not guarantee easy species differentiation. According to Price (1994), variation within a species can be great enough that variation between species can become impossible to quantify. We had several challenges to overcome with this dataset.

Our field data were collected a full four years after the collection of the imagery. This prevented us from taking detailed measurement *in situ* concurrent with the image capture, which has been an advantage shared by researchers that have been successful in similar studies. Van Aardt, (2000) found that using spectral signatures acquired *in situ*, high levels of accuracy were achievable in classifying remotely sensed images. There is ample evidence of active prescribed fire management in the area that could conceivably have caused significant vegetative change between image capture and field data collection.

Low density canopy closure is a significant challenge. Okin (2001) found that multiple endmember spectral mixture analysis applied to high quality field spectra and AVIRIS data did not provide reliable vegetation type when cover is less than 30%, and Okin's research was attempting to discriminate between vegetation types with much more pronounced phenological differences. The highest canopy closure in our study was barely 70%, with the majority of stands in question falling well below 50%. Not only was canopy cover very low in our study area, but leaf area index was extremely low as well. As research at SETRES has shown, the extremely well-drained soils of the sandhills do not provide adequate moisture for robust growth of longleaf, and especially

loblolly pine. Thus, the reflecting light represented by a pixel situated perfectly over a target tree will often represent less than 50% actual green vegetation. The remainder would be a mix of branches, understory, ground cover and sand.

Hotelling's T test compares reflectance at specific wavelengths across the entire variable set, so differences in environmental conditions that affect reflectance – like time of day, angle of the sun and moisture content – make it difficult to prove similarity between stands. Spectral signatures of individual species have similar shape, but total reflectance is influenced more by factors such as moisture content, atmospheric conditions and angle of light reflectance. Because of this, total reflectance is a poor identification tool (Puritch 1981). Hotelling's T-square is also a particularly stringent test – a large variation in between one or two variables can result in the rejection of the null hypothesis (Murtagh and Heck 1987).

There are alternate promising paths for future research into this question. A dataset collected with this research in mind could more adequately anticipate the needs of the research. Higher spatial resolution would help identify target trees, and choice of location to target stands with higher density would also be of some benefit. This would also allow field data to be collected concurrently with the image capture, eliminating the uncertainty of any disturbance occurring between the two time frames.

This research shows that even with highly mixed pixels, it is possible to discriminate longleaf pine from loblolly pine given ample training data.

## **Suggestions for Further Research**

There are several steps needed to take the results of this research to an applicable stage. Our method was successful at determining overstory pine species in pure stands, and had much lower accuracy in mixed stands. The mixing of the two types in a single pixel created a signature that is difficult to predict.

Perhaps the greatest failing of this method, with regards to the project for which it was conceived, is its inability to determine stand structure. This method is meant to help identify pine stands with hardwood midstory. This process can identify pine and hardwood pixels fairly accurately, but tells nothing about the structure of the overstory or understory composition, age, or height.

A process such as this, while accurate at describing an individual pixel, is not designed to classify an entire stand. A fairly substantial amount of *a priori* knowledge will still be needed to classify large areas into discrete stand types. A model that groups and classifies entire stands is beyond the scope of this project.

## References

- Becker, B. L., D. P. Lusch. "Identifying Optimal Spectral Bands From *In Situ* Measurements of Great Lakes Coastal Wetland Using Second-Derivative Analysis." Remote Sensing of the Environment 97 (2005) 238-48.
- Bell, J. F. "High Spectral Resolution UV to Near-IR Observations of Mars Using HST/STIS" ICARUS 191 (2007) 581-602.
- Bunting, P., R. Lucas. "The Delineation of Tree Crowns in Australian Mixed Species Forests Using Hyperspectral Compact Airborne Spectrographic Imager (CASI) Data" Remote Sensing of Environment 101 (2006) 230-48.
- Chen, X., Lee V. "Spectral Mixture Analyses of Hyperspectral Data Acquired Using a Tethered Balloon" Remote Sensing of Environment 103 (2006) 338-50.
- Congalton, R. G., K. Green. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices Ann Arbor, MI: Lweis Publishers, 1999.
- Conner, R. N. D., C. Rudolph, J. R. Walters. The Red-Cockaded Woodpecker: Surviving in a Fire Maintained Ecosystem. Austin, Texas: University of Texas Press, 2000.
- Crum, S. Aerial Photography and Remote Sensing 2000. Department of Geography, University of Texas at Austin. 06 Aug. 2008.  
[http://www.colorado.edu/geography/gcraft/notes/remote/remote\\_f.html](http://www.colorado.edu/geography/gcraft/notes/remote/remote_f.html).
- Datt, B. "Recognition of Eucalyptus Forest Species Using Hyperspectral Reflectance Data" International Journal of Remote Sensing 20 (1999) 2741-59.



- Flores, F. J. "Using Remote Sensing Data to Estimate Leaf Area Index and Foliar Nitrogen of Loblolly Pine Plantations" Ph.D. dissertation, University of North Carolina, Raleigh, NC, USA, 2003.
- Galvao, L. S., A .R. Formaggio, D. A. Tisot. "Discrimination of Sugarcane Varieties in Southeastern Brazil With EO-1 Hyperion Data" Remote Sensing of Environment 94 (2005) 523-34.
- Gong, P., R. Pu, B. Yu. "Conifer Recognition: An Exploratory Analysis of *In Situ* Hyperspectral Data" Remote Sensing of Environment 62 (1997) 189-200.
- Govender, M., K. Chetty, H. Bulcock. "A Review of Hyperspectral Remote Sensing and its Application in Vegetation and Water Resource Studies" Water SA 33 (2007) 145-52.
- Hadi, A. S. "Identifying Multiple Outliers in Multivariate Data" Journal of the Royal Statistical Society 54 (1992) 761-77.
- Jensen, J. R. Remote Sensing of the Environment: An Earth Resource Perspective Upper Saddle River, NJ: Prentice Hall, 2007.
- Kallio, K. "Retrieval of Water Quality From Airborne Imaging Spectrometry of Various Lake Types in Different Seasons" The Science of the Total Environment 268 (2001) 59-77.
- Leckie, D. G., C. Burnett, F.A. Gougeon, T. Nelson, D. Paradine, S. Tinnis. "Automated Tree Recognition in Old Growth Conifer Stands With High Resolution Digital Imagery" Remote Sensing of Environment 91 (2005) 311-26.

- Lewis, M.M. “Discriminating Vegetation with Hyperspectral Imagery – What is Possible?” presented at EPA Conference Spectral Remote Sensing of Vegetation, Las Vegas, Nevada, 2003.
- Liesenberg, V., S. G. Lenio, P. F. Jorge. “Variations in Reflectance With Seasonality and Viewing Geometry: Implications for Classification of Brazilian Savanna Physiognomies with MISR/TYA Data” Remote Sensing of Environment 107 (2007) 273-86.
- Maathuis, B. H. P., J. L. Van Denderen. “A Review of Satellite and Airborne Sensors for Remote Sensing Based Detection of Minefields and Landmines” International Journal of Remote Sensing 25 (2004) 5210-45.
- Mathur, A., L. M. Burce, J. Byrd. “Discrimination of Subtly Different Vegetative Species via Hyperspectral Data” IEEE International Geoscience and Remote Sensing Symposium 2 (2002) 805-7.
- Murtagh, F., A. Heck. Multivariate Data Analysis Dordrecht: Astrophysics and Space Science Library, 1987.
- Okin, G. S., D. A. Roberts, B. Murray, W. J. Okin. “Practical Limits on Hyperspectral Vegetation Discrimination in Arid and Semiarid Environment” Remote Sensing of Environment 77 (2001) 212-5.
- Petrie, G.M. “Optimal Band Selection Strategies for Hyperspectral Data Sets” IEEE International Geoscience and Remote Sensing Symposium 3 (1998) 1582-4.
- Price, J.C. “How Unique are Spectral Signatures?” Remote Sensing of Environment 49 (1994) 181-6.

Puritch, G. S. Nonvisual Remote Sensing of Trees Affected by Stress – a Review

Canada: Minister of Supply and Services, 1981.

Vaughn, G. R. “Surface Mineral Mapping at Steamboat Springs, Nevada, USA, With Multi-Wavelength Thermal Infrared Images” Remote Sensing of Environment 99 (2005) 140-58.

Van Aardt, J. A. N. “Spectral Sperability Among Six Southern Tree Species” Master’s thesis, Virginia Polytechnic Institute, Blacksburg, VA, USA, 2000.

## Appendix

## Appendix 1: Eigen Values and Vectors Used to Construct Principal Components

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	106.581	95.7808	0.8882	0.8882
Comp2	10.8002	8.81913	0.0900	0.9782
Comp3	1.98102	1.61856	0.0165	0.9947
Comp4	.362459	.281076	0.0030	0.9977
Comp5	.0813823	.0291123	0.0007	0.9984
Comp6	.0522701	.0139851	0.0004	0.9988
Comp7	.038285	.0200121	0.0003	0.9991
Comp8	.0182729	.00657811	0.0002	0.9993

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8
v1	0.0922	-0.0160	-0.1962	0.0481	0.3349	0.0685	-0.0009	0.1191
v2	0.0924	-0.0216	-0.1957	0.0430	0.3189	0.0707	-0.0140	0.1070
v3	0.0925	-0.0232	-0.1928	0.0401	0.3074	0.0748	-0.0094	0.0947
v4	0.0927	-0.0246	-0.1873	0.0432	0.2862	0.0650	-0.0095	0.0779
v5	0.0931	-0.0240	-0.1798	0.0527	0.2510	0.0444	-0.0229	0.0354
v6	0.0934	-0.0184	-0.1754	0.0624	0.2222	0.0319	-0.0311	-0.0365
v7	0.0936	-0.0070	-0.1736	0.0712	0.1917	0.0145	-0.0416	-0.1369
v8	0.0936	0.0005	-0.1757	0.0747	0.1272	-0.0087	-0.0539	-0.1722
v9	0.0935	0.0022	-0.1804	0.0715	0.0463	-0.0300	-0.0624	-0.1407
v10	0.0934	-0.0005	-0.1857	0.0639	-0.0359	-0.0601	-0.0703	-0.0837
v11	0.0933	-0.0023	-0.1883	0.0586	-0.0929	-0.0785	-0.0766	-0.0453
v12	0.0933	-0.0044	-0.1876	0.0540	-0.1233	-0.0808	-0.0804	-0.0144
v13	0.0933	-0.0076	-0.1843	0.0515	-0.1398	-0.0794	-0.0832	0.0159
v14	0.0934	-0.0115	-0.1802	0.0484	-0.1510	-0.1022	-0.0992	0.0192
v15	0.0935	-0.0157	-0.1751	0.0474	-0.1622	-0.0950	-0.1003	0.0599

### Eigen Values and Vectors Used to Construct Principal Components

v16	0.0936	-0.0192	-0.1693	0.0457	-0.1734	-0.0926	-0.0996	0.0991
v17	0.0937	-0.0198	-0.1659	0.0413	-0.1781	-0.0870	-0.0995	0.0856
v18	0.0941	-0.0051	-0.1624	0.0342	-0.1628	-0.0950	-0.1045	-0.1013
v19	0.0940	0.0323	-0.1466	0.0439	-0.1501	-0.0748	-0.0781	-0.2820
v20	0.0902	0.0973	-0.1099	0.0855	-0.1352	-0.0381	0.0015	-0.3059
v21	0.0834	0.1499	-0.0654	0.1233	-0.1080	0.0138	0.1211	-0.1414
v22	0.0798	0.1692	-0.0389	0.1323	-0.0901	0.0404	0.1811	-0.0134
v23	0.0788	0.1742	-0.0259	0.1345	-0.0837	0.0320	0.1754	-0.0016
v24	0.0789	0.1742	-0.0158	0.1280	-0.0736	0.0309	0.1676	0.0078
v25	0.0792	0.1733	-0.0059	0.1206	-0.0627	0.0274	0.1546	0.0076
v26	0.0797	0.1715	0.0048	0.1127	-0.0507	0.0214	0.1355	-0.0000
v27	0.0796	0.1719	0.0167	0.1052	-0.0383	0.0272	0.1213	0.0093
v28	0.0795	0.1724	0.0271	0.0985	-0.0226	0.0299	0.1040	0.0206
v29	0.0796	0.1719	0.0368	0.0917	-0.0099	0.0262	0.0872	0.0121
v30	0.0802	0.1688	0.0433	0.0812	0.0081	0.0322	0.0764	0.0174
v31	0.0801	0.1693	0.0422	0.0829	0.0055	0.0324	0.0791	0.0167
v32	0.0809	0.1651	0.0526	0.0665	0.0254	0.0380	0.0535	0.0619
v33	0.0820	0.1596	0.0571	0.0567	0.0328	0.0201	0.0341	0.0053
v34	0.0830	0.1536	0.0646	0.0417	0.0423	0.0367	0.0306	0.0442
v35	0.0858	0.1384	0.0616	0.0237	0.0352	0.0185	0.0245	-0.0238
v36	0.0876	0.1272	0.0587	0.0102	0.0274	0.0026	-0.0001	-0.0428
v37	0.0881	0.1236	0.0574	0.0021	0.0301	-0.0044	-0.0181	-0.0368
v38	0.0879	0.1246	0.0627	-0.0054	0.0358	0.0112	-0.0269	0.0151
v39	0.0874	0.1274	0.0695	-0.0089	0.0421	0.0180	-0.0403	0.0597
v40	0.0868	0.1306	0.0767	-0.0116	0.0458	0.0266	-0.0519	0.1091
v41	0.0860	0.1355	0.0802	-0.0120	0.0477	-0.0013	-0.0680	0.0609
v42	0.0857	0.1370	0.0846	-0.0150	0.0494	-0.0023	-0.0897	0.0924
v43	0.0857	0.1368	0.0853	-0.0182	0.0504	-0.0179	-0.1162	0.0920
v44	0.0861	0.1339	0.0866	-0.0274	0.0498	-0.0107	-0.1184	0.1148
v45	0.0871	0.1273	0.0852	-0.0400	0.0504	-0.0034	-0.1220	0.1233
v46	0.0885	0.1175	0.0816	-0.0515	0.0518	0.0041	-0.1228	0.1130
v47	0.0910	0.0988	0.0655	-0.0615	0.0208	-0.0351	-0.0939	-0.0112

### Eigen Values and Vectors Used to Construct Principal Components

v48	0.0933	0.0771	0.0516	-0.0798	0.0040	-0.0385	-0.0855	-0.0311
v49	0.0946	0.0592	0.0351	-0.0907	-0.0005	-0.0599	-0.0907	-0.0931
v50	0.0949	0.0550	0.0344	-0.0961	0.0002	-0.0478	-0.0924	-0.0593
v51	0.0951	0.0514	0.0365	-0.1026	0.0014	-0.0238	-0.0764	-0.0311
v52	0.0952	0.0498	0.0365	-0.1056	0.0031	-0.0280	-0.0800	-0.0357
v53	0.0951	0.0506	0.0397	-0.1100	0.0058	-0.0313	-0.0903	-0.0355
v54	0.0949	0.0529	0.0456	-0.1148	0.0074	-0.0318	-0.1089	-0.0238
v55	0.0947	0.0549	0.0501	-0.1218	0.0094	-0.0260	-0.1240	-0.0049
v56	0.0946	0.0556	0.0511	-0.1283	0.0094	-0.0316	-0.1401	-0.0009
v57	0.0947	0.0546	0.0517	-0.1344	0.0121	-0.0435	-0.1530	-0.0206
v58	0.0948	0.0502	0.0543	-0.1396	0.0143	-0.0447	-0.1674	-0.0361
v59	0.0952	0.0418	0.0560	-0.1433	0.0175	-0.0456	-0.1744	-0.0628
v60	0.0957	0.0298	0.0516	-0.1466	0.0148	-0.0466	-0.1671	-0.0946
v61	0.0961	0.0158	0.0434	-0.1484	0.0103	-0.0166	-0.1429	-0.0848
v67	0.0948	-0.0573	-0.0411	-0.0442	-0.0795	0.0008	0.0487	0.1345
v68	0.0949	-0.0567	-0.0394	-0.0540	-0.0751	0.0198	0.0476	0.1280
v69	0.0949	-0.0559	-0.0365	-0.0630	-0.0700	0.0364	0.0534	0.1296
v70	0.0951	-0.0545	-0.0331	-0.0725	-0.0599	0.0195	0.0571	0.1048
v71	0.0952	-0.0527	-0.0297	-0.0816	-0.0498	0.0049	0.0581	0.0752
v72	0.0953	-0.0509	-0.0259	-0.0885	-0.0413	0.0035	0.0664	0.0583
v73	0.0954	-0.0493	-0.0216	-0.0939	-0.0304	0.0056	0.0772	0.0408
v74	0.0955	-0.0473	-0.0181	-0.0982	-0.0246	0.0050	0.0837	0.0269
v75	0.0955	-0.0457	-0.0149	-0.1035	-0.0168	-0.0009	0.0883	0.0053
v76	0.0956	-0.0446	-0.0114	-0.1076	-0.0127	0.0076	0.0968	-0.0007
v77	0.0956	-0.0435	-0.0091	-0.1137	-0.0075	0.0088	0.1007	-0.0137
v78	0.0956	-0.0432	-0.0062	-0.1198	-0.0050	0.0200	0.1037	-0.0206
v79	0.0956	-0.0427	-0.0044	-0.1247	-0.0008	0.0109	0.1015	-0.0402
v80	0.0956	-0.0421	-0.0026	-0.1286	-0.0008	0.0108	0.1023	-0.0506
v81	0.0956	-0.0426	-0.0014	-0.1275	0.0043	0.0132	0.1064	-0.0603
v82	0.0956	-0.0421	-0.0005	-0.1242	0.0049	0.0120	0.1129	-0.0596
v83	0.0956	-0.0419	0.0013	-0.1202	0.0116	0.0127	0.1228	-0.0562
v84	0.0957	-0.0421	0.0020	-0.1161	0.0078	0.0217	0.1283	-0.0439

### Eigen Values and Vectors Used to Construct Principal Components

v85	0.0956	-0.0429	0.0019	-0.1098	0.0116	0.0128	0.1281	-0.0478
v86	0.0956	-0.0439	0.0014	-0.1032	0.0068	0.0255	0.1372	-0.0228
v87	0.0956	-0.0450	0.0007	-0.0984	0.0062	0.0191	0.1367	-0.0296
v88	0.0955	-0.0465	-0.0001	-0.0974	0.0031	0.0219	0.1325	-0.0193
v89	0.0954	-0.0486	0.0000	-0.0979	0.0012	0.0167	0.1325	-0.0199
v90	0.0953	-0.0502	-0.0015	-0.0962	-0.0001	0.0068	0.1294	-0.0197
v91	0.0952	-0.0522	-0.0013	-0.0953	-0.0016	-0.0030	0.1270	-0.0275
v92	0.0951	-0.0543	-0.0004	-0.0958	-0.0018	0.0036	0.1291	-0.0266
v93	0.0950	-0.0548	-0.0018	-0.0966	-0.0029	-0.0127	0.1214	-0.0433
v94	0.0945	-0.0556	-0.0020	-0.0984	0.0211	0.0293	0.1351	-0.1282
v95	0.0908	-0.0809	-0.0182	0.0167	-0.2228	0.9168	-0.2529	-0.1125
v96	0.0931	-0.0822	-0.0082	0.0011	-0.0894	0.0410	-0.0039	0.1927
v97	0.0933	-0.0810	-0.0083	-0.0044	-0.0816	-0.0206	-0.0332	0.1348
v98	0.0934	-0.0794	-0.0131	-0.0108	-0.0788	-0.0221	-0.0205	0.0955
v99	0.0935	-0.0780	-0.0027	0.0084	-0.0790	0.0240	-0.0090	0.1885
v100	0.0939	-0.0743	-0.0035	0.0207	-0.0696	-0.0305	-0.0185	0.1512
v101	0.0940	-0.0725	0.0017	0.0345	-0.0702	-0.0129	-0.0037	0.1649
v102	0.0940	-0.0725	0.0101	0.0491	-0.0615	-0.0096	0.0016	0.1646
v103	0.0939	-0.0732	0.0199	0.0590	-0.0534	-0.0145	0.0041	0.1614
v104	0.0936	-0.0752	0.0306	0.0671	-0.0468	-0.0219	0.0025	0.1262
v105	0.0931	-0.0802	0.0460	0.0726	-0.0298	-0.0193	0.0077	0.0725
v106	0.0911	-0.0942	0.0915	0.0947	0.0096	-0.0356	-0.0032	-0.0501
v107	0.0892	-0.1047	0.1210	0.0952	0.0491	-0.0364	-0.0052	-0.1493
v108	0.0888	-0.1071	0.1251	0.0859	0.0662	-0.0371	0.0017	-0.1895
v109	0.0864	-0.1177	0.1557	0.0858	0.1042	-0.0199	0.0122	-0.2889
v110	0.0899	-0.1023	0.1087	0.0650	0.0600	-0.0138	0.0203	-0.1440
v111	0.0919	-0.0901	0.0745	0.0590	0.0321	-0.0066	0.0218	-0.0128
v112	0.0923	-0.0868	0.0661	0.0690	0.0303	-0.0046	0.0253	0.0089
v113	0.0922	-0.0867	0.0714	0.0857	0.0256	-0.0052	0.0191	0.0528
v114	0.0917	-0.0898	0.0806	0.1022	0.0115	-0.0178	0.0054	0.0278
v115	0.0911	-0.0936	0.0903	0.1153	0.0027	-0.0121	0.0001	0.0234
v116	0.0904	-0.0974	0.1007	0.1256	0.0119	-0.0234	-0.0133	0.0149



### **Eigen Values and Vectors Used to Construct Principal Components**

v117		0.0904	-0.0985	0.0971	0.1206	0.0088	-0.0144	-0.0079	-0.0160
v118		0.0895	-0.1027	0.1110	0.1318	0.0212	-0.0199	-0.0273	0.0165
v119		0.0890	-0.1060	0.1169	0.1299	0.0237	-0.0419	-0.0346	-0.0107
v120		0.0886	-0.1094	0.1167	0.1321	0.0216	-0.0088	-0.0283	-0.0441
v121		0.0890	-0.1074	0.1104	0.1259	0.0156	-0.0425	-0.0385	-0.0160
v122		0.0886	-0.1087	0.1138	0.1362	0.0095	-0.0218	-0.0444	0.0241
v123		0.0881	-0.1115	0.1194	0.1453	0.0084	-0.0311	-0.0595	-0.0274
v124		0.0872	-0.1136	0.1323	0.1669	-0.0055	-0.0127	-0.0513	-0.0228

## Appendix 2: Means and Standard errors for Component Variables

Estimation sample discrim qda  
Summarized by group

	group	1	2	3	4	5	6	7
8								
--								
c1								
Mean		-8.820697	-11.38913	-9.266449	-7.747677	-12.25717	-7.617916	-9.125302
6.065312								
SE mean		.1884281	.2044501	.1629782	.2043311	.0398522	.2038417	.204733
.329353								
--								
c2								
Mean		-2.412469	-3.23906	1.242198	1.218697	-1.076852	-4.647342	-.0231434
.8192862								
SE mean		.250769	.128654	.111483	.1479805	.0484987	.1677644	.1341343
.2350531								
--								
c3								
Mean		-.6763825	-1.225185	-.3316707	-.0587775	-.9875267	-1.003985	-.6025413
.4244523								
SE mean		.073025	.0473937	.048803	.0561192	.0137916	.0603479	.064054
.1020753								
--								
c4								
Mean		-.0421337	.2903304	.1667075	-.1280269	.3690419	-.6756592	.5987431
.8200823								

## Means for Component Variables

SE mean		.0433607	.0479572	.0441062	.0361026	.0122846	.0478114	.0431207
---------	--	----------	----------	----------	----------	----------	----------	----------

-----+-----

c5								
Mean		-.1125063	-.244598	.1548455	.0285001	-.0836586	-.3204833	-.0963342 -
SE mean		.0129361	.0098181	.0078804	.0068003	.0057768	.0157148	.0133917

-----+-----

c6								
Mean		-.0717479	.07483	-.030755	-.0612754	-.09501	.0336565	.1453435 -
SE mean		.0281049	.0202814	.0177925	.0184182	.0164329	.024566	.0337171

-----+-----

c7								
Mean		.1768116	.054975	-.1250968	-.1274366	-.1116845	.1327075	.0874708 -
SE mean		.019794	.0189018	.0155561	.0105153	.0074453	.0214039	.0202835

-----+-----

c8								
Mean		-.0061494	.0078807	.0482103	-.033201	.1669329	.0220615	-.1117499 -
SE mean		.0135678	.0072659	.0084767	.0059735	.0039786	.0077016	.0089448

## Means for Component Variables

		9	10	11	12	16	17	18
21								
-----+-----								
--								
c1								
	Mean	-8.443721	-8.864097	13.07828	17.01734	6.964982	5.112456	3.692223
6.996878								-
	SE mean	.2148585	.1312007	.1305504	.25334	.3857112	.2819166	.1624993
.2567996								
-----+-----								
--								
c2								
	Mean	-.0613924	-.1450162	-1.691854	3.191397	-2.300481	-3.324724	-3.221232
2.219111								
	SE mean	.1262823	.0958773	.0828064	.0893225	.1147101	.0814447	.0527334
.3911662								
-----+-----								
--								
c3								
	Mean	-.1100944	-.6481551	1.714703	-2.432356	.5874182	.6444531	.3897336
.2282963								
	SE mean	.0611795	.0327587	.0217857	.0915679	.0381798	.0202904	.0379009
.098664								
-----+-----								
--								
c4								
	Mean	-.4130754	-.0686277	.417575	-.1656303	-.1123998	-.0099292	-.2464072
.3091772								
	SE mean	.0641422	.0419696	.0259047	.0499011	.0568794	.0471144	.0504247
.0338742								

## Means for Component Variables

-----+-----								
--								
c5								
Mean		-.1237459	.0057708	-.3047567	-.0025629	.3089071	.3350376	.1842081 -
.0552184								
SE mean		.0090143	.0094173	.0252776	.024689	.0176147	.0143611	.0090394
.015585								
-----+-----								
--								
c6								
Mean		-.0329416	-.0227139	-.1035839	-.0010441	.1034273	.0853044	.0142419 -
.04626								
SE mean		.0228405	.0285685	.0143937	.0173362	.0237498	.0174106	.0160604
.0379832								
-----+-----								
--								
c7								
Mean		-.1508856	-.1498106	-.1018326	.0108082	.028098	.1349337	.0544949
.2456407								
SE mean		.0164561	.0163301	.0093044	.0109501	.0202677	.0117479	.0120115
.0202754								
-----+-----								
--								
c8								
Mean		-.0091615	-.0511646	.0492705	.004346	-.11918	-.0078967	-.0446832 -
.1466521								
SE mean		.0076122	.009049	.0092835	.0127532	.0061724	.0065186	.0065127
.0149707								

### Means for Component Variables

c1				
	Mean	-6.251152	-5.860384	7.39e-09
	SE mean	.1293877	.158116	.2294179
c2				
	Mean	6.364834	5.065092	-1.32e-09
	SE mean	.1284793	.1860547	.0730302
c3				
	Mean	1.653562	1.285691	0
	SE mean	.0427765	.0536472	.0312775
c4				
	Mean	.2319323	-.0629256	-2.29e-10
	SE mean	.0169828	.0407671	.0133788
c5				
	Mean	.0932094	-.0306838	6.87e-11
	SE mean	.0094293	.0072281	.0063395
c6				
	Mean	.10482	.0527071	-1.72e-10
	SE mean	.0159226	.0181058	.0050806
c7				
	Mean	.2128822	.0793873	-1.16e-10
	SE mean	.0113741	.0165286	.0043481