8-2008

Comparing Linear Discriminant Analysis with Classification Trees Using Forest Landowner Survey Data as a Case Study with Considerations for Optimal Biorefinery Siting

Yingjin Wang
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Yingjin Wang entitled "Comparing Linear Discriminant Analysis with Classification Trees Using Forest Landowner Survey Data as a Case Study with Considerations for Optimal Biorefinery Siting." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Statistics.

Timothy M. Young, Major Professor

We have read this thesis and recommend its acceptance:

Frank M. Guess, Russell Zaretzki

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Yingjin Wang entitled "Comparing Linear Discriminant Analysis with Classification Trees Using Forest Landowner Survey Data as a Case Study with Considerations for Optimal Biorefinery Siting". I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Statistics.

Dr. Timothy M. Young, Major Professor

We have read this thesis

and recommend its acceptance:

Dr. Frank M. Guess

Dr. Russell Zaretzki

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Comparing Linear Discriminant Analysis with Classification Trees Using Forest Landowner Survey Data as a Case Study with Considerations for Optimal Biorefinery Siting

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Yingjin Wang

August, 2008

DEDICATION

This thesis is dedicated to my husband. You are the most generous person I ever met in my life. Thank you for your love, support and patience since we have been together. I am eternally grateful.

Also, I dedicate this thesis to my parents for their unconditional love and support since the day I was born. Even though they are on the other side of the earth, I feel their love and keep them in my heart every day.

ACKNOWLEDGEMENT

This research was partially supported by The University of Tennessee Agricultural Experiment Station McIntire-Stennis E112215 (MS-75); USDOT Sun Grant Center Grant R11-0515-016 and USDA Forest Service Grant R11-0515-019. Funding was also provided by University of Tennessee, Department of Statistics, Operations, and Management Science. Special thanks to Dr. Timothy G. Rials, Professor and Directors University of Tennessee, Sun Grant and Forest Products Centers; and Mr. James H. Perdue, USDA Forest Service Director's Representative to the University of Tennessee.

I would like to express my gratitude to the Committee, Dr. Timothy M. Young, Dr. Frank M. Guess, and Dr. Russell Zaretzki for their guidance throughout this research project. These scholars have provided wonderful guidance. It has been an honor to work with them. They have helped to provide wonderful growth in my graduate work. Other professors who have had a profound influence on my education include Dr. Mary Sue Younger, Dr. Hamparsum Bozdogan, Dr. Robert Mee, Dr. William Seaver, and Dr. Mary Leitnaker.

I especially want to thank Dr. Guess for being such a wonderful mentor and providing me with support, guidance, and encouragement throughout this research. He is one of the nicest persons I ever met in the world. I have learned so much from him that will help me not only in my professional career as a statistician but also in my personal life.

Also, I would like to thank Graduate Research Assistant Nancy Liu and Kerri Norris for their helpful discussion in my research. In addition, I thank Amanda Silk for all of her professional and personal support and kindness.

ABSTRACT

Bioenergy is reemerging as an important topic in energy-related research. The rapid increase in costs of petroleum products has led to a renewed interest in alternative sources of energy such as biofuels. World-wide energy consumption has increased 17 times in the last century and the demand for energy in emerging markets such as China and India is projected to increase in the future at unprecedented rates. A review of the current bioenergy literature is presented in this thesis. Also, comments on the economics of bioethanol are discussed.

The primary part of the thesis focuses on statistical classification methods related to factors that influence landowner attitudes towards harvesting timber. A comparison of linear discriminant analysis (LDA) and classification tree (CT) methods is presented using the results of a forest landowner survey as a case study. Several CT techniques are discussed with an emphasis on the CRUISE classification tree program. The LDA procedure in SPSS is used to construct linear discriminant functions of the survey results. CRUISE is also used to construct classification trees of the survey results. Survey results showed that 73.3 percent of farmer forest landowners harvested timber, and 69.6 percent of non-farmers who had a length of residency beyond 36.5 years harvested timber. For landowners who conducted commercial timber harvests, the importance level of income from the harvest was the overriding factor relative to all other factors. Discriminant analysis results supported the results of CTs. However, the linear discrimination functions and corresponding coefficients did not provide the level of two-dimensional detail of CTs, which also detected hidden interactions.

The other component of the thesis assesses optimal cellulosic biorefinery sites that use mill residues in Texas and Louisiana based on the cost of trucking transportation costs. A transportation model for trucking costs is developed and is used to estimate supply curves or marginal cost curves for mill residues. Mill residue data are provided by the U.S. Forest Service. The resolution of the data is by zip code. The top five sites in Texas and Louisiana are presented.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1

1. Introduction

The twentieth century was marked by rapid growth and increased prosperity in the world. The rapid increase in costs of fossil fuels has led to a renewed interest in new sources of energy (Goldemberg 2000). Also, the increasingly serious problem of greenhouse gases in atmosphere from the combustion of fossil fuels has accelerated the emergence toward the development of alternative energy sources such as bioenergy and biofuels (Ture et al. 1997). Liquid fuels and value-added chemicals from biomass will be required to meet the greater energy demand and represent low-risk solutions to providing a renewable, sustainable and secure domestic energy supply (Demirbas 2000, Balat 2005). In this thesis, economic and environmental issues related to biofuels are presented. The key topic of finding optimal biofuel plant locations in Texas and Louisiana are also discussed. Statistical classification methods are used to study the factors that influence landowner attitudes towards harvesting timber, a key consideration for bioenergy plants interested in procuring cellulosic fiber from standing trees for conversion to biofuels or biochemicals. A comparison of linear discriminant analysis (LDA) (Balakrishnama and Ganapathiraju 1998) and classification tree (CT) methods (Brieman et al. 1984) is presented using the results of a landowner survey as a case study (Longmire et al. 2007). These methods provide procurement foresters and land managers with objective and scientific-based tools to assess characteristics of forest landowners likely to harvest timber commercially and also characteristics of forest landowners that are not likely to harvest timber.

Chapter 2 provides a brief introduction of the main concepts and issues related to biomass and biofuels. The literature review begins with a brief history and status quo of global energy supply and consumptions, and the importance and urgency of developing renewable energy sources. Next, the biomass and biofuels are defined; and the economic and environmental issues of using bioenergy and biofuels are discussed. Transportation costs associated with supplying mill residues to biorefineries are estimated. Cellulosic ethanol, a biofuel produced from the woody parts of trees/plants, perennial grasses, or their residues, is contrasted with grain-based ethanol produced mainly from corn. The chapter concludes with a discussion of long-term potential of biofuels.

In Chapter 3 statistical classification methods are presented with an emphasis on linear discriminant analysis (LDA) and classification trees (CT). The history, theory, limitations, validation, and applications of these two methods are compared. Several CT techniques are studied and compared. CRUISE, which is an unbiased classification tree program, is specially examined. CRUISE has three split methods, three variable selection methods, and three pruning methods; it is unique among classification tree algorithms with high prediction accuracy and fast computation speed. CRUISE is free of selection bias, sensitive to local interaction between variables, and robust to missing values in the learning sample (Kim and Loh 2001, Kim and Loh 2003).

Chapter 4 presents a case study of using LDA and CT methods for the results of a forest landowner survey. The private woodland owner survey was conducted in Tennessee in seven counties located in the region known as the Northern Cumberland Plateau in 2005 (Longmire et

al. 2007). One objective of the survey was to characterize differences between forest landowners that harvest trees with forest landowners that do not harvest trees. The LDA procedure in SPSS is used to construct linear discriminant functions. CRUISE is used to construct classification trees: 13 combinations of variable selection methods and split-point selection methods are examined, and optimal classification trees are presented.

Chapter 5 discusses preliminary research work supporting a larger research study on developing a real-time, web-based optimal siting system for biomass energy producing facilities for the 33 Eastern United States. Methods are presented in this chapter that can be used to select optimal biorefinery locations based on mill residue quantities and trucking transportation costs. The smallest geographic resolution of the data for selecting a biorefinery location is a zip code. Mill residues from primary wood manufactures are considered the cellulosic feedstocks, and supply curves for transportation costs by truck are constructed (Hodges et al. 2007). As a case study, the five lowest cost zip codes are selected in Texas and Louisiana (Liu et al. 2008).

Finally, chapter 6 summarizes the overall intent in this thesis. Possibilities for future research are also discussed.

Chapter 2

2. Literature Review of Biomass and Biofuels

2.1. Introduction

A plethora of literature are available in the field of biomass and biofuels. Six hundred references that focus on the economics and availability of biomass and biofuels, and the process of siting for a biofuel refinery are reviewed in this chapter. Many citations are from the "*Journal of Biomass & Bioenergy*", which was established in 1991 (*http://www.elsevier.com/wps/find/journaldescription.cws_home/986/description#description*, referenced 5/08/2008).

The 20th century has been marked by rapid industrial growth and increased prosperity in the world. Energy consumption world-wide has increased 17-fold in the last century (Goldemberg 2000). A frequently documented problem related to this issue is the increasing emission into the atmosphere of greenhouse gases resulting from the combustion of fossil fuels (Ture et al. 1997). The rapid increase in real prices of fossil fuels in the world market during the last decade has accelerated the interest of both scientists and politicians toward the development of new and alternative sources of energy. The oil reserves of the world are estimated to be depleted in less than 50 years at the present rate of consumption (Gommers et al. 2000, Sheehan et al. 2000). As exploration for new petroleum resources becomes increasingly expensive, there will be more economic advantages in substituting conventional petroleum-based fuels and products with alternative renewable sources of energy, such as

cellulosic-based fuels that use agricultural feedstock and woody residues (Ozcimen and Karaosmanoglu 2004, Jefferson 2006).

There are many factors, such as sustainability, energy security, economic stability, and environmental and other socioeconomic issues that must be considered for bioenergy and biofuels. Modern innovations and technology have made some biofuels cost-competitive with respect to fossil fuels (Demirbas 2000). This cost-competitiveness may develop into a cost advantage if the prices of fossil fuels continue to increase at the same rate (Cadenas and Cabezudo 1998).

The transition to a renewable energy-supply source, which has lower levels of pollutants, may prove to be more vital, given the potential threat of global climatic change, primarily caused by the combustion of fossil fuels. The agreement at Kyoto in December 1997 (*http://en.wikipedia.org/wiki/Kyoto_Protocol*, referenced 5/08/2008) signaled the political acceptance of threats of global pollution and need for an alternative-fuel strategy among the industrialized countries around the globe. As of November 2007, 175 countries, with the exception of the United States, have ratified the protocol, which insists that emission of carbon dioxide and other greenhouse gases must be reduced. Therefore, energy from biomass is thought to be an environmentally friendly alternative.

In the following parts of this chapter, certain important issues related to biofuels are discussed. The second part introduces the concept of biomass and biofuels and discusses the relative economic and environmental issues. The third section discusses the transportation costs of biofuels, which is a critical issue for determining the sites to locate biofuel facilities. The

fourth section contains discussions regarding cellulosic ethanol, which is a biofuel produced from the woody parts of trees/plants, perennial grasses, or their residues. This technology has been commercialized and has great long-term potential. The fifth section comments on a sensitive topic: grain-based ethanol, which is mainly produced from corn. Grain-based ethanol has a long history and its technology and application are quite established. However, it has been increasingly criticized as economically inefficient and of questionable social benefit; the pros and cons of this issue are objectively discussed in this section. Some novel research studies are also listed herein, e.g., biofuels based on life residues or animal fat. Summarizing remarks are provided in the last section.

2.2. Biomass and Biofuels

2.2.1. Concept of Biomass and Biofuels

Biomass is a generic term and it refers to living and recently dead biological material that can be used as fuel or for industrial production of other byproducts. Most commonly, biomass refers to plant material grown for use as biofuel, but it also includes plants or animals used for production of fibers, chemicals, or heat. Biomass may also include biodegradable wastes that can be burnt as fuel (*http://en.wikipedia.org/wiki/Biomass*, referenced 5/08/2008).

Biomass has been used for energy purposes for a long time. The most common and traditional use of biomass is as firewood for cooking and heating, which is not sustainable because it may contribute to land degradation, sometimes leading to desertification. The modern use of biomass is different from the traditional use. It focuses on the conversion of

feedstock into high-quality energy carriers, for example, electricity and liquid biomass fuels used for transportation (Wyman 1996). Milbrandt (2005) has commented that compared with other renewable resources, biomass is very flexible: it can be used as fuel for direct combustion, gasified, and used in combined heat and power technologies or biochemical conversions. Biomass is receiving increasing attention as scientists, policy makers, and biomass growers search for clean, renewable energy alternatives. Main biomass conversion process is illustrated in Figure 2.1.

Biofuels are biomass-based transportation fuels. Broadly, biofuels are defined as solid, liquid, or gas fuels, consisting of or derived from biomass. However, biofuels are strictly defined as liquid or gas fuels used for transportation that are derived from biomass (*http://en.wikipedia.org/wiki/Biofuels*, referenced 5/08/2008). Biofuel is an alternative to fossil fuel that reduces greenhouse-gas emission and increases long-term energy security.

Biofuels are useful globally and biofuel industries are expanding rapidly in Europe, Asia, and the Americas (Byrne et al. 1996, Puhan et al. 2005, Soccol et al. 2005). The most common use for biofuels is in automotive transport (Fulton et al. 2004). Biofuels can be produced from any carbon source that can be replenished rapidly. Biofuels have many advantages, including sustainability, reduction of greenhouse-gas emission, and guaranteed supply of raw material (Reijnders 2006). Sources of the main liquid biofuels are illustrated in Figure 2.2.

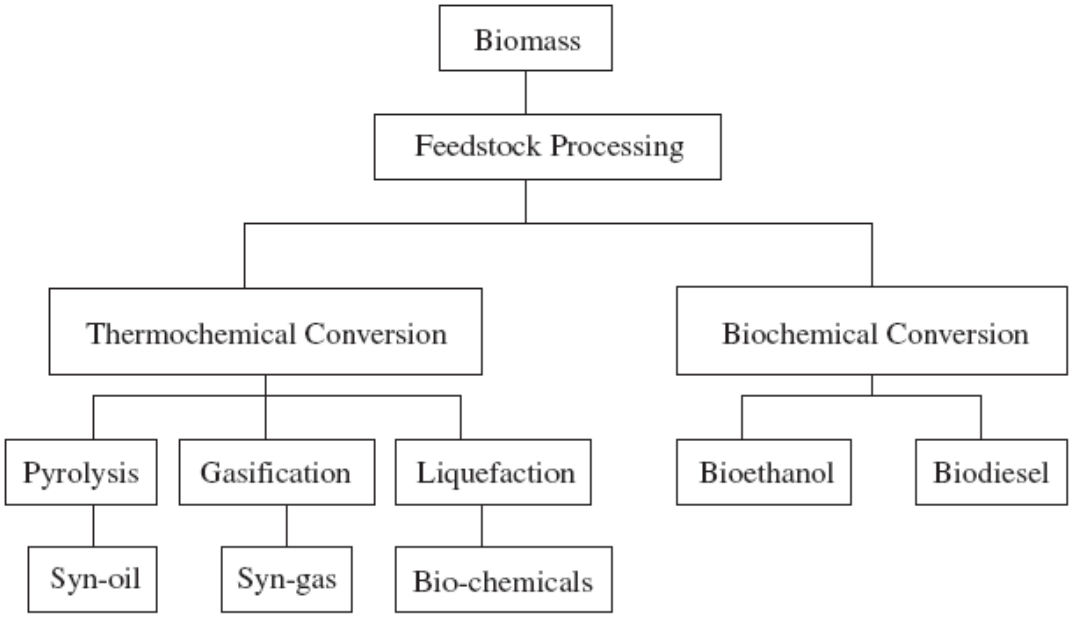

Figure 2.1 Main biomass conversion process. Source: (Demirbas 2007)

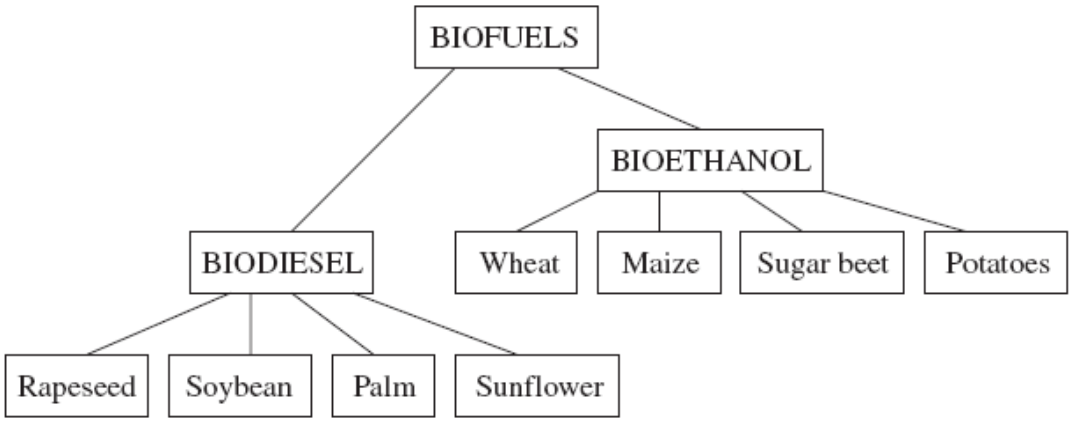

Figure 2.2 Sources of the main liquid biofuels. Source: (Demirbas, 2007)

2.2.2. Overview of the Production and Consumption of Biofuels

Bioethanol and biodiesel are the two most modern biofuels (Demirbas 2002). Bioethanol is a fuel derived from renewable sources of feedstock such as corn, wheat, sugar beet, straw, and wood. It is an additive to or substitute for gasoline. Wood, straw, and household wastes can be converted to bioethanol. Bioethanol is derived from the alcoholic fermentation of sucrose or simple sugars, which are produced from biomass by hydrolysis (Demirbas 2007). Biodiesel, which is derived from vegetable oil, is an alternative to diesel fuel. Producing biodiesel from vegetable oil yields less pollution than petroleum-diesel fuel, but its price is more than double the cost of diesel because of the substantial increase in the price of feedstock (Ma and Hanna 1999). Although biodiesel fuels are currently mainly prepared from soybean oil, there are large amounts of low-cost oils and fats that can be converted into biodiesel, such as restaurant wastes and animal fats including beef tallow and pork lard (Canakci and Van Gerpen 2001, Zhang et al. 2003).

Recent advances related to the production of biofuels are based on the development of new technologies in the fields of chemistry and engineering. Ptasinski et al. (2007) have compared different types of biofuels with reference to their gasification efficiency and have used this as a benchmark against the gasification efficiency of coal. Poole et al. (2007) have analyzed the emission of various elements from a biofuel gasifier. Some insightful reports are listed herein, which deal with the latest technologies in biofuel production: Demirbas (2003), Bothast (2005), Demirbas (2006), Chen and Dixon (2007), and Brothier et al. (2007).

2.2.3. Economic and Environmental Advantages of Biofuels

Economic and environmental advantages promote the use of biofuels as an attractive alternative to use of fossil fuels. Two important aspects lead to biofuels being considered competitive in price: innovative technology substantially decreases the cost of producing biofuels; the dramatically increasing price of oil facilitates the use of biofuels for economic advantages. More important, the use of biofuels is considered to be environmentally friendly. Biofuel is a potential substitute for conventional fuels that reduces pollution and supports sustainable agriculture.

Matthews (2001) has described the development and application of a standard methodology for evaluating the energy and carbon budgets of a biofuel production system. Puppan (2002) discusses life-cycle assessment, which is a scientific evaluation method to investigate the net environmental impacts of biofuels. He positively concludes that the impacts of biodiesel and bioethanol on the environment are more favorable than those brought about by conventional fuels. Gan and Smith (2006) have investigated the cost-competitiveness of woody biomass for electricity production, by considering the reduction in greenhouse-gas emission and taxes payable, which might enhance the economic potential of production and use of bioenergy.

Further details about the economic efficiency of biofuels are available in studies by Weber (1993), Bender (1999), Taleghani and Kia (2005), and Hill et al. (2006). More articles about the evaluation of the environmental impact of biofuel production are available in Demirbas (2004), Balat (2005), Demirbas and Balat (2005), Wierzbicka et al. (2005), Bies (2006),

Gorski (2006), Lal (2006), Eriksson and Johansson (2006), Arellano et al. (2006), Jain et al. (2006), Wu et al. (2006), and Adler et al. (2007).

2.2.4. Prospective Vision of Biofuels

The International Energy Agency has published a report "*Biofuels for Transport: An International Perspective"* in 2004. This report contains a global perspective and assesses the strides made and developments achieved in the field of bioenergy and the direction toward which the bioenergy field appears to be heading. It reviews recent research efforts and experiments in various aspects of bioenergy: process technology, impacts of reduction in greenhouse-gas emission, biofuel costs, market impacts, feedstock availability, and policy incentives (Fulton et al. 2004).

Biofuel production throughout the world has increased very rapidly, from 28 to 44 billion liters in 2006: production of bioethanol and biodiesel increased by 22 percent and 80 percent, respectively. Although biofuels comprise less than one percent of the global liquid fuel supply, the increase in the production of biofuels in 2006 led to a 17 percent increase in worldwide liquid fuel supply (Hunt 2006). Modern biomass energy is expected to make up a significant share of the future energy market. Adoption of bioenergy is expected to be the result of: 1) decreasing costs related to the production and conversion of biomass energy, 2) abundant feedstock resources, and 3) environmentally-friendly characteristics of biomass.

Biofuel is considered to be a strategic resource for energy supply in the future, i.e., which will fulfill the Kyoto agreement to replace fossil fuels and to mitigate greenhouse-gas

emissions. Many industrialized countries already use a significant quantity of biofuels for their energy supply (Hillring 2006). Many influential organizations anticipate biomass to play a major role in the future, because it is a more sustainable, renewable and globally applicable energy supply. Meanwhile, the environmental and social pros and cons of biomass must be considered based on its life-cycle analysis and evaluation of the external ramifications. Further discussions about the future of biofuels are available in publications by Hall and Scrase (1998), Parikka (2004), Hoogwijk et al. (2005), and Kaltschmitt and Weber (2006).

2.3. Transportation Costs and Biofuel Refinery Siting Models

2.3.1. Transportation Costs

Transportation cost, the cost of moving feedstock or products, is an important component of the overall costs for recovering energy from biomass. Transportation cost typically represents a substantial portion of the total costs of woody biomass, due to the low value of the wood, long distance traveled for delivery, and increasing costs of diesel fuel. Evaluating the economic feasibility of biomass resources as sources of energy therefore requires a comprehensive study of transportation costs.

Transportation cost is determined significantly by the geographical location of forests and energy plants. Searcy et al. (2007) have analyzed the transportation cost for projects of small and large sizes. They have calculated the relative costs of transportation by truck, rail, ship, and pipeline for three biomass feedstocks, which were straw, corn stover, and woodchips. They suggest that the transportation costs should be differentiated into distance-fixed costs

(loading and unloading) and distance-variable costs (transport, including power losses during transmission). Generally, transport of biomass feedstocks is more expensive than that of energy products. The difference between the cost for the transport of biomass and that for the transport of energy is significantly higher than the incremental cost of building and operating a power plant situated a long distance away from a transmission grid. Moller and Nielsen (2007) have presented a method based on continuous cost-surface mapping using raster-based geographical information systems (GISs). After mapping the wood-chip resources, the model can be built using cost-distance functions, supply curves, and sensitivity analysis. This model can be thereafter used to evaluate the transportation costs for selected plants yielding bioenergy. Matthew (2006) has built a system that allows users to identify the least costly path from the sources of wood to the mill, on the basis of road quality, speed limits, terrain, and other transportation variables. Nilsson (1999) presents a dynamic simulation model for the analysis of various delivery alternatives to improve and optimize the performance of the system and to reduce the costs related to the harvest, transport, and storage of raw materials. This model is based on other sub-models that associate the infrastructural and geographical aspects with the weather conditions. Walsh (1998) has summarized the production cost for a bioenergy-yielding crop, its supply curve, and its transportation costs to arrive at an estimate of the end-cost of the biomass. The economics of biomass feedstocks in the United States and the characteristics of several biomass feedstocks (such as agricultural residues, forestry residues, solid municipal waste, and crops dedicated to bioenergy) are examined. Switchgrass and short-rotation wood are the focuses in this report because of their large yield potential, wide

geographical distribution, and broad commercial and research use. For the end-user of bioenergy crops, the costs include both the price of crop production and the cost of transporting the bioenergy crops from the site of production to the site of utilization.

Biomass resources in the southeast of the United States, such as urban wood waste and forest residues, can be used to generate renewable energy and provide economic benefits to rural communities. However, the feasibility of any biomass project depends on the availability of woody biomass resources. Therefore, it is important to comprehensively consider the type of available biomass material, the distance of its transport, and the available transportation infrastructure (Langholtz 2006). Only after examining all these major issues, does it become possible to evaluate the economic feasibility of using biomass resources as energy sources.

2.3.2. Biorefinery Siting Modeling

Biorefinery siting modeling is a key topic in the research related to biofuel production. Considering the environmental impacts, economic influences, political incentives, and availability of labor, biorefinery siting modeling is a complex procedure, and the interest in modeling optimal sites for biorefineries has increased immensely among researchers.

Sperling (1984) presents a generalized framework for analyzing the relative attractiveness of investments made in biomass fuel production at a disaggregate level, using a system-approach to integrate site-specific considerations. On the basis of this approach, a model has been built to site and size the prospective biomass-fuel manufacturing plants and to identify and specify the critical factors for their operations. He points out that there are five

crucial issues that most strongly influence the site-selection process and determination of the size of biomass-fuel manufacturing plants. They are feedstock supply, fuel distribution, fuel demand, co-product demand, and feedstock processing. The relationship between these five factors is used to estimate cost functions. Generalized feedstock-supply curves, processing-cost functions, and model transportation-cost functions have been formulated to specify the feedstock production and fuel-transport subsystems. After calibrating the model for each local area with local costs and conditions and fuel-demand patterns, this framework may be applied in any area with abundant biomass.

Young et al. (1991) have developed a model to estimate the average total cost of producing whole-tree chips from woody biomass for energy production. The total cost consists of the costs related to the harvest, transportation, and stumpage of the woody biomass. This model has been used to estimate the total costs of 62 potential locations for biomass-manufacturing plants in Southeast United States. This model has applied a spatial analytical component and uses a GIS to locate potential sites for biorefineries. This model also measures the impact of market and nonmarket conditions on the economic availability of woody biomass. The authors concluded that Northeast Florida, Southern Georgia, Southern Alabama, and the Coastal Plain of South Carolina were low-cost regions for the production of woody biomass for bioenergy. The South Delta of Louisiana, state of Kentucky, state of West Virginia, and the mountains regions of Tennessee and Virginia were higher-cost regions.

Noon (1996) constructed a regional integrated biomass assessment (RIBA) system, which consisted of two phases: 1) the descriptive phase that characterizes the farmgate cost

and supply surface for biomass production over a given state; and 2) the analytical phase that uses a transportation model to compute the marginal cost of supplying raw material to an energy-producing plant at a prescribed level of demand. The model generates a marginal cost-surface that illustrates the most promising regions for locating a bioethanol plant. A sequential location model simulated the commercial development of ethanol-production facilities. This model considers every road-network node as a potential site and generates a sequence of probable plant locations. Furthermore, Noon and Daly (1996) designed a system to estimate the total purchase and transportation costs of three types of wood-fuel under various levels of demand. This system includes information on all possible wood-fuel supply points, demand points, and product-movement costs.

Graham (1996) developed a GIS-based modeling system for analyzing the geographic variation in potential bioenergy-feedstock supplies and optimal locations for locating bioenergy facilities. The modeling system was designed for analyzing individual U.S. states, but it can be adapted to any geographic region. The modeling system has four basic components: mapping crop-land availability, calculating the expected yields and farmgate price, mapping the cost of the delivered energy-crop feedstocks, and mapping the probable sites for the co-location of bioenergy facilities.

Graham et al. (2000) constructed a regional-scale, GIS-based modeling system for estimating the potential biomass supplies from energy crops. The system considers the regions where energy crops could be grown, the spatial variability in their yield, and the transportation costs associated with acquiring the feedstock needed for an energy facility. The potential costs

and supplies of switchgrass in 11 U.S. states are estimated by this system. They concluded that transportation costs are the lowest in Iowa, North Dakota, and South Dakota; and are the highest in South Carolina, Missouri, Georgia, and Alabama. They additionally estimated across 11 states, the costs of delivered feedstocks which ranged from $33 to $55 per dry ton for supplying a facility that requires 100,000 ton/yr. Delivered feedstock costs for a 630,000 ton/yr facility range from $36 to $58 per dry ton. Graham et al. (1997) and Husain et al. (1998) also conducted insightful research on modeling for optimal biofuel sites.

2.4. Biofuels based on Cellulosic Ethanol

Cellulose is the most abundant organic compound on earth. Cellulosic materials are considered to be the most potential raw materials for the production of ethanol. Although there are multiple technical and economic challenges associated with the large-scale production of ethanol from cellulosic biomass, including the collection and transportation of the biomass raw material and the preprocessing or pretreatment associated with it, many advances have been achieved in each of these areas during recent years.

The basic design for cellulosic ethanol is to extract the cellulose locked up in cell walls of plants, break it down into its component sugars, and ferment these sugars into ethanol. People have long ago established the method of making alcohol from grains, and now, a similar process is used to convert farm products into ethanol for fuel (Wu 2007). Cellulosic-alcohol fuels are produced from the woody parts of trees and plants, perennial grasses, or farm residues. This technology is now being commercialized and has great long-term potential.

"Cellulosic ethanol is projected to be much more cost-effective, environmentally beneficial and to have a greater output-to-input ratio of energy than grain ethanol" (Solomon et al. 2007). The environmental costs and benefits of biofuel production from lignocellulose-based energy crops exceed those from food-based crops (Hill 2007). The new generation of biofuels derived from lignocellulosic sources offers greatly reduced environmental impacts while simultaneously avoiding potential conflicts between food and energy production (Dietter et al. 1997).

Solomon et al. (2007) investigated pilot plants and demonstration plants set up for the manufacture of cellulosic ethanol in the United States, Canada, Japan, Sweden, and Denmark. In the United States, the plants are grown in Alabama, California, Mississippi, Colorado, Arizona, Arkansas, Hawaii, Louisiana, and Nebraska. They discussed many related factors such as the economic investments needed for erecting the plants and the cost of biomass. More recently, cellulosic biomass has been studied by many individuals and organizations. The topics deal with a wide range of issues, from technological improvements in the field to the environmental impacts and economic availabilities. The following citations constitute a brief list of many articles written on cellulosic biofuels: Lynd (1996), Scurlock et al. (2000), Murray et al. (2003), McLaughlin and Kszos (2005), Stephanopoulos (2007), Polagye et al. (2007), Granda et al. (2007), Foyle et al. (2007), and Nakagawa et al. (2007).

2.5. Grain-based Ethanol and Biofuels Based on Other Materials

2.5.1. Biofuels Based on Food stock: Grain-based Ethanol

Grain-based ethanol has a long history, and the technology and application involving its production are quite mature. Biofuel production in the United States is currently dominated by ethanol obtained from corn and biodiesel obtained from soybeans (Hamelinck and Faaij 2006a). Currently, American corn-based ethanol is expensive, although it can help cut oil imports and provides modest reductions in greenhouse-gas emissions compared to conventional gasoline. Corn-ethanol has some risks. Because corn is a foodstuff for people and animals, diverting large amounts of corn into ethanol production could push up prices and even cause shortages. In the last decade, criticism aimed at the subsidization of grain-based ethanol has increased (Hamelinck and Faaij 2006b). The economical efficiency and social benefits of this manufacturing process are doubted (Charles et. al. 2008).

Currently, the total production of ethanol worldwide is approximately ten billion gallons. In 2006, 18 percent of the corn crop in the United States was used for ethanol production. Ethanol production from corn grain is predicted to grow in volume to 12 to 15 billion gallons per year, and any additional growth in ethanol production will come from other raw materials (Gray 2007). Other grain-based fuel studies are by Jorapur and Rajvanshi (1997), Pordesimo et al. (2005), Sims et al. (2006), and Torney et al. (2007).

2.5.2. Biofuels Based on Other Materials

Because of the existence of a wide variety of biomass, studies which examine other types of biomass are ongoing. Gigler et al. (1999) discussed strategies related to the supply of the willow plant for energy-producing purposes and develop minimum cost-supply strategies. Gercel (2002) has evaluated the production of biofuel by the pyrolysis of sunflower-oil cake. Das et al. (2004) detailed a type of biofuel obtained from the pyrolysis of the cashew nut shell. Brumbley et al. (2007) have focused on sugarcane that has the potential to be a major crop in the developing field of biomass production. It is the second fastest-growing tropical grass and can be harvested multiple times before replanting. There are many studies that have been published about new materials and technology in relation to biofuel production, see Allen and Bennetto (1993), Angenent et al. (2004), Ieropoulos et al. (2005), Bullen et al. (2006), Davis and Higson (2007), and Du et al. (2007).

2.6. Conclusions

The world is now in the age of biomass and biofuel technology development. Biomass technology is far reaching from modern medicine and the pharmaceutical industry, to the fields of energy and agriculture, to the manufacturing biofuel. Political instability in the Middle East and the increasing demand from developing countries for crude oil are accelerating the interest in bioenergy and biofuel technology. The combination of effects of new biomass technology, the economic impact of increasing cost of crude oil, and government investment is creating a new “bio-economy”.

Transportation cost is a critical issue in low-cost biofuel production. Because transportation costs account for a significant portion of the total biofuel cost, it is important to appropriately locate new biofuel facilities to minimize transportation costs and maximize the economic advantages. Site-selection for biofuel facilities is complex. A comprehensive understanding of the mechanisms of feedstock supply, fuel distribution, fuel demand, co-product demand, feedstock processing, local economic availability, and policy incentives is needed.

The long-term potential for biofuels is in nonfood feedstock, such as agricultural and forestry wastes, and fast-growing, cellulose-rich energy crops such as perennial grasses and trees. Cellulosic ethanol can reduce greenhouse-gas pollution that results from present-day biofuel crops. “The question is not whether biofuels will play a major part in the global transportation-fuel market, but when and at what price” (Hunt 2006). Many countries are actively encouraging the use of biomass for energy and prioritizing the development of the necessary knowledge and technology for modern biomass-energy systems. This new renewable energy source must displace the use of the traditional fossil-fuels.

Chapter 3

3. Methods of Statistical Classification

3.1 Introduction

Statistical classification is a procedure in which individual cases are sorted into groups based on one or more quantitative and/or qualitative characteristics in the cases. There are numerous techniques and algorithms for classification problems (Agresti 2002), such as logistic regression, linear discriminant analysis, cluster analysis, and classification trees (CTs); most software packages already include these functions. Each technique has its own limitations and advantages (Press and Wilson 1978, Dudoit et al. 2002). Logistic regression classifies observations by building a regression function to estimate the probability of occurrence of an event; therefore, it can only be used to classify the response variable with two levels, i.e., success or failure of the event (Menard 2002). Linear discriminant analysis (LDA) develops classifications by determining linear combinations of the predictor variables to split the observations, but the application of LDA has some important assumptions, i.e., multi-normality of each group and homogeneity of covariance matrices (Balakrishnama and Ganapathiraju 1998). Cluster analysis separates observations into two or more unknown groups based on combinations of predictor variables; results of the cluster analysis may differ significantly when using different algorithms at different initial settings (Aldenderfer and Blashfield 1984). Classification trees determine a set of statistical-based "if-then" conditions (instead of linear equations) for predicting and classifying cases; the major advantage of CTs is the direct and intuitive way by which they can be interpreted (Buntine 1992, Kim and Loh 2003).

LDA was originally developed in 1936 and has wide applications in various fields (Fisher 1936, Altman 1968, Lachenbruch and Goldstein 1979, Klecka 1980, McLachlan 1992, Altman et al. 1994, Zhao et al. 1999). LDA functions are based on two assumptions: multi-normality in each group and common covariance between groups. LDA has two limitations: (a) there must be statistical significant difference in the mean vectors between groups, and (b) the number of observations in each group must be greater than the number of predictors. If any one of these assumptions is not met, the results may be unreliable (Eisenbeis 1977).

CT methods are an element of decision tree theory (Quinlan and Rivest 1989). A decision tree (DT) is a decision support tool, which generally uses a graph or model of decisions and their possible consequences. Although DT is a relatively new method, it is encompassed within the larger body of data mining theory (Fayyad et al. 1996, Cherkassky and Mulier 1998). DTs are popular in statistics as a predictive modeling technique. CTs are used for modeling and predicting categorical response variables (e.g., gender, day of week, class, or group membership) from a set of continuous predictors and/or categorical predictors. CTs are promising classification methods because of their simple interpretation, high classification accuracy, and ability to characterize complex interactions among variables. A popular reference for the background of CTs from both historical and mathematical perspectives is the monograph "*Classification and Regression Trees*" (CART) by Brieman et al. (1984). More detailed algorithms and numerous applications of CTs are available (Frydman et al. 1985, Buntine 1992, Kuhn and De Mori 1995, Loh and Shih 1997, Poon et al. 2001, Kim and Loh 2001, Kim and Loh 2003).

3.2 Linear Discriminant Analysis

Discriminant analysis, originally developed in 1936 by R.A. Fisher, is a multivariate method of classification (Fisher 1936). Discriminant analysis is similar to regression analysis except that the dependent variable is categorical rather than continuous (Draper and Smith 1981). In discriminant analysis, the intent is to predict class membership of individual observations based on a set of predictor variables. LDA generally attempts to find linear combinations of predictor variables that best separate the groups of observations. These combinations are called discriminant functions (Mika et al. 1999).

Suppose there are *K* different groups, each assumed to have a multivariate normal distribution with mean vectors $\boldsymbol{\mu}_k$ (*k=1,…,K*) and common covariance matrix $\boldsymbol{\Sigma}$. The actual mean vectors and covariance matrices are almost always unknown; the maximum likelihood estimates are used to estimate these parameters.

The idea of LDA is to classify observations $\boldsymbol{x}_i$ to the group *k*, which minimize the within-group variance, i.e.,

$$k = \text{argmin}_k(\boldsymbol{x_i} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) \tag{3.1}$$

Under multivariate normal assumptions, this is equivalent to finding the group that maximizes the likelihood of the observation. Generally, we can estimate prior probability using the proportion of the number of observations in each group to the total. For example, let $\pi_k = \frac{n_k}{n}$ be the proportion of group *k*, such that$\pi_1 + \cdots + \pi_K = 1$. Then, instead of maximizing

the likelihood, the posterior probability is maximized; the observation belongs to a particular group,

$$k = \text{argmax}_k \left[-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) + log\pi_k \right] \tag{3.2}$$

Simplifying (2), the *k* LDA functions are,

$$d_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k{}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + log\pi_k \tag{3.3}$$

When the assumption of common covariance matrix is not satisfied, an individual covariance matrix for each group is used. This leads to quadratic discriminant analysis (QDA) as the discriminating boundaries are quadratic curves instead of straight lines. Box's M test is used to test the homogeneity of variance (Box 1949, Geisser and Greenhouse 1958, Pintrich and De Groot 1990). When the test is significant, QDA is used. QDA does not guarantee an improved classification rate (Bouveyron 2007).

In the binary case, two linear discriminant functions are built as follows:

$$d_1(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \frac{1}{2} \boldsymbol{\mu}_1{}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + log\pi_1 \tag{3.4}$$

$$d_2(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 - \frac{1}{2} \boldsymbol{\mu}_2{}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + log\pi_2 \tag{3.5}$$

If $d_1(\boldsymbol{x}) > d_2(\boldsymbol{x})$, the observation $\boldsymbol{x}$ will be assigned to group one, otherwise to group two. The two discriminant functions can also be combined, i.e.,

$$\begin{aligned} d(\boldsymbol{x}) &= d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) \\ &= \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + log \frac{\pi_1}{\pi_2} \end{aligned} \tag{3.6}$$

If $d(\boldsymbol{x}) > 0$, the observation $\boldsymbol{x}$ will be assigned to group one, otherwise to group two. The last two parts in the equation (3.6) are constant given a data set; the discriminant function coefficients are $\boldsymbol{D} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. The coefficients reflect the joint contribution of the variables to the function, thereby showing the influence of each variable in the presence of the others. The standardized coefficients $\boldsymbol{D}^* = \mathrm{diag}(\boldsymbol{\Sigma})\boldsymbol{D}$ are computed by multiplying each coefficient by the standard deviation of the corresponding variables. When the variable scales differ substantially, the standardized coefficient vector provides better information about the relative contribution of each variable to the canonical discriminant function (Rencher 1992).

Suppose there are two groups of *p* predictor variables, which allow for construction of LDA functions using all predictors. A practical process is to choose significant variables using stepwise procedure, which uses the Wilks' Lambda statistics to identify significant independent variables of the discriminant functions (Siotani et al. 1985, Rencher 1993). The Wilks' Lambda criterion maximally discriminates between groups by maximizing the multivariate *F* ratio in the tests of differences between the group means.

Discriminant functions are built based on two assumptions, i.e., multi-normality in each group and homogeneity of covariance between groups. If there are many categorical predictor variables, these two assumptions are often violated, which may influence the quality of the models and predictions. Other limitations with discriminant analysis are that the mean vectors of the groups must be distinguishable and that the number of observations in each group must be greater than the dimension of the variables. If the mean vectors are not different enough, it

is hard for LDA to yield decent classification rates. If the observations in some groups are limited, a stepwise procedure is required to select potential important variables before LDA can be used.

While performing classification problem, the classification rate needs to be estimated (Koehler and Erenguc 1990). One simple method is called re-substitution, which applies the discriminant model to the original training data set to observe the frequency of correctly classified observations. Re-substitution generally overestimates the correct classification rate (Braga-Neto et al. 2004).

Another method for measuring the probability of correct classification is *q*-fold cross-validation (Geisser 1974). For *q*-fold cross-validation, the original sample is partitioned into *q* subsamples. A single subsample is retained for validation of the model built from the other *q-1* subsamples each time. The process is repeated *q* times, with each of the *q* subsamples used exactly once for validation. The results are combined to produce a single classification rate estimate. A specific application is the "leave-one-out" cross-validation, where *q* equals the number of observations in the original data set (Verbyla and Litvaitis 1989, Kohavi 1995).

In this thesis, linear discriminant functions are estimated with SPSS v16.0 (SPSS Inc. 2007) using a stepwise procedure where the variable that minimizes the overall Wilks' Lambda is entered at each step. The minimum partial *F* to enter is 3.84 (α = 0.05), and maximum partial *F* to remove is 2.71 (α = 0.10). SPSS provides the classification rate from re-substitution and from "leave-one-out" cross-validation (SPSS Inc. 2007).

3.3 Classification Tree

The machine-learning technique for inducing a DT from data is called decision tree learning or (colloquially) "decision trees" (Young 2007). DT methods can be applied either to continuous data, which are called regression trees (Chaudhuri et al. 1994), or to categorical data, which are called classification trees (Kim et al. 2007). DT models have grown into a powerful class of methods for examining complex relationships with various types of data. Researchers and practitioners find great explanatory value in DT models. The main advantage of a tree over the other models is the ease with which the model can be explicitly interpreted (Loh 2002).

The *Automatic Interaction Detection* (AID) algorithm by Morgan and Sonquist (1963), Kass (1975), and Fielding (1977) is the first implementations of the DT idea. The CART algorithm followed AID and is a popular DT method (Brieman et al. 1984). Since CTs were first introduced, many adaptations and extensions have been proposed, e.g., CART, C4.5, FACT, CHAID, FIRM, GUIDE, QUEST, and CRUISE (*Classification Rule with Unbiased Interaction Selection and Estimation*), see Wilkinson (1992).

In general, CTs build the rules by recursive binary or multiway partitioning of the data space into subspaces that are increasingly homogeneous with respect to the class variable. The homogeneous regions are called nodes. At each step in fitting a CT, an optimization is carried out to select a node, a predictor variable, and a split-point for numeric variables or group of codes for categorical variables that result in the most homogeneous subgroups for the data

(Brieman et al. 1984). The splitting process continues until further subdivision no longer increases the homogeneity of the nodes. When this occurs, a CT is said to be fully grown, and the final regions are called terminal nodes. But the lower branches of a fully grown CT are actually developed based on sampling error, which is called "over-fit," and these lower branches are typically pruned (Chou et al. 1989, Gelfand et al. 1991). Interpretation of CTs increases in complexity as the number of terminal nodes increases.

In classification trees methods, selection bias is one of the most important issues. Selection bias is a phenomenon that favors certain types of variables over others as the split variable (Strobl et al. 2007). The definition of selection bias is that if the predictor variables are independent of the class variable, they will not have the same chance of being selected for splitting. There are two sources of bias: one is when variables differ significantly in their numbers of splits, and the other is when variables differ in their proportions of missing values. Selection bias is a serious problem in most classification tree techniques. For example, greedy search algorithms, such as CART, have selection bias due to joint selection of the predictor variable and the split point. CART has a preference in selecting variables with more splits. When such bias is not prevented, the inferences for those variables can be flawed and misleading. Loh and Shih (1997) showed that separation of the variable selection from the split-point selection can avoid such selection bias. QUEST and CRUISE use this method to minimize or to prevent selection bias. There are some other problems with other algorithms, e.g., CART and QUEST only have binary splits; FACT, C4.5, CHAID, and FIRM have multiway splits but have selection bias; FACT and FIRM do not prune; and C4.5 has multiway splits but only for categorical

variables. CRUISE was selected in this research because of its exemption from the aforementioned limitations.

3.4 CRUISE

CRUISE is a program for tree-structured classification (Kim and Loh 2001). It contains several algorithms for the construction of CTs. CRUISE has three split methods, i.e., univariate split, linear combination split, and univariate split with bivariate node models (Kim and Loh 2003). It also has three variable selection methods and three pruning methods. CRUISE has four ways to handle missing values, which is an inherent problem with survey data. CRUISE is unique among classification tree algorithms with the following properties:

- Fast computation speed by using multi-way splits,
- Practically free of selection bias,
- Sensitive to local interaction between variables,
- Robust to missing values in the learning sample.

The first variable selection method in CRUISE is the univariate split method, where each split involves only one variable. Recall that the key to avoiding selection bias is separation of variable selection from split-point selection. The univariate split method uses the following steps to construct its split rules:

Step 1. Selection of split variable

- One-dimensional method: Computes *p*-values from analysis of variance *F*-tests for numerical variables and from contingency table χ^2 -tests for categorical variables, and then selects the variable with the smallest *p*-value.

- Two-dimensional method: Focuses on interactions between variables; tests the most significant numerical variable, the most significant categorical variable, and the most significant pair of numerical variables, the most significant pair of categorical variables, and the most significant pair of numerical and categorical variables, respectively; and then chooses the most significant variable.

Step 2. Selection of split point for the variable

- Because LDA is most effective when the data are normally distributed with the same covariance matrix, if the selected variable *X* is a numerical variable, perform a Box-Cox transformation on the *X* value first, and then apply LDA to the *X* values to find the split points.

- If *X* is a categorical variable, it is first converted to a *0-1* vector, and then follows the previous step as a new numerical variable.

Another split method of CRUISE is linear combination splits, which have greater flexibility, prediction accuracy, and fewer terminal nodes, although this does not necessarily translate to improved interpretation. Linear combination splits contain the following three steps:

Step 1. Each categorical variable is transformed to a dummy vector and then projected onto the largest discriminant coordinate; this maps each categorical variable into a numerical variable.

Step 2. Perform a principal component analysis of the correlation matrix of the variables; principal components with small eigenvalues are dropped to reduce the influence of noise variables.

Step 3. LDA is applied to the remaining principal components to find the split.

The most novel split method in CRUISE is univariate splits with node models, which retains univariate splits but fits a linear discriminant model to the best two-variable plot at each node. The common goal in classification tree methods is to obtain a tree such that the learning sample in each terminal node is quite homogeneous. When this cannot be achieved with a small number of univariate splits, a large tree or an extremely simple one (due to over-pruning) will be constructed. A possible solution is to use linear combination splits, but such splits are usually difficult to interpret if they involve more than two variables. By building a tree with node models, better classification performance without losing lucid interpretation is achieved.

To provide useful information, the tree structure must be easy to understand and free of bias in the split selections. CRUISE uses two techniques to improve the interpretability of its trees (Kim et al. 2007). First, it splits each node into multiple subnodes, with one for each class; this reduces tree depth. Second, it selects variables based on both one-factor and two-factor effects; therefore, CRUISE can immediately identify a variable with a significant two-factor

interaction even when it does not have a significant one-factor effect. More importantly, CRUISE uses a two-step approach to free itself from selection bias. First, it uses the *p*-values from significance tests to select variables, which avoids the bias of the greedy search approach caused by variables with unequal numbers of splits. It also automatically accounts for unequal numbers of missing values using the degrees of freedom. CRUISE then uses bootstrap bias correction to further reduce the bias due to differences between numerical and categorical variables (Friedman 2001, Schapire 2002). The bootstrap correction is critical because the amount of bias is dependent on many aspects of the data, such as sample size, number and type of variables, missing value pattern, and configuration of the data points.

Chapter 4

4. A Comparison of Linear Discriminant Analysis with Classification Trees for a Forest Landowner Survey as a Case Study

4.1 Introduction

In 2005, a survey of 495 private forest landowners was conducted in seven counties of Tennessee located in the Northern Cumberland Plateau region. The aim of the survey was to differentiate between forest landowners who harvest trees from forest landowners who do not harvest trees. Linear discriminant analysis (LDA) methods (Balakrishnama and Ganapathiraju 1998) are compared with classification tree (CT) methods (Brieman et. al. 1984, Kim and Loh 2001, Kim and Loh 2003) where the strengths and weaknesses of each method are noted for the analysis of the forest landowner survey (Agresti 2002). The forest landowner survey was used as a case study for the comparison of methods.

In the construction of the classification trees in this study, 13 combinations of variable selection methods and split-point selection methods were used; optimal classification trees are presented. Survey results showed that 73.3 percent of farmer forest landowners harvested timber and 69.6 percent of non-farmers who had a length of residency beyond 36.5 years harvested timber. For landowners who conducted commercial timber harvests, the importance level of income from the harvest was the overriding factor relative to all other factors. Discriminant analysis results supported the results of classification trees. However, the linear discrimination functions and corresponding coefficients do not provide the level of two-

dimensional detail of classification trees and split-points of predictors, which also detected hidden interactions. The graphical representation of classification trees may provide useful and easy-to-interpret information to foresters and other business professionals interested in identifying characteristics of forest landowners likely to harvest timber.

4.2 The Survey Dataset

A private woodland owner survey was conducted in Tennessee in seven counties located in the region known as the Northern Cumberland Plateau in 2005. The objective of the survey was to characterize differences between forest landowners who harvest trees from those forest landowners who do not harvest trees. The survey response rate was 55 percent based on 1,012 verifiable forest landowners in the region (Longmire et al. 2007). Twenty-four independent variables were examined, which related to years of residency, management planning, importance of timber harvest income, areas of the forest land, number of tracts of land owned, reasons for conducting timber harvests, demographic characteristics, etc. The categorical dependent variables were "whether the forest landowners conducted timber harvests", and "whether the timber harvests were commercial or non-commercial harvests".

Of the 495 usable survey responses, 243 respondents harvested timber. Of the 243 respondents who harvested timber, 111 conducted commercial timber harvests. LDA functions and CT models were developed for three pairs of responses: 1) landowner who harvested timber versus landowners who did not harvest timber, 2) forest landowners who distinguished

between commercial and non-commercial timber harvests, and 3) forest landowners who conducted commercial timber harvests versus all other forest landowners.

4.3 The Results of Linear Discriminant Analysis

4.3.1 Comparing "Timber Harvest" with "No Timber Harvest" Responses

Five significant variables (α = 0.05) selected by the stepwise procedure are given in Table 4.1. The standardized canonical discriminant function coefficients, which measure the relative importance of the selected variables (i.e., the larger absolute value of the coefficient corresponds to greater discriminating ability) indicate that the independent variable "multi-year management plan" was the most powerful discriminating variable, followed by "farmer identification."

Given the intention to compare these two groups, two classification functions were used to assign cases into each group. For each observation, two classification scores were computed for each function. The cases were assigned to the group whose function obtained the higher score. Given the magnitude and signs of the coefficients (Table 4.2), it is evident that "multi-year management plan" and "farmer identification" increased the likelihood of timber harvest, followed by "years of residency" and "area of land in acres." If the landowners owned any land outside the study area, their probability of conducting a timber harvest declined. The decrease in harvesting for other land ownership is surprising but may reflect second home or non-resident landownership, which was not directly assessed by the survey.

Table 4.1 Standardized canonical discriminant function coefficients distinguishing "timber harvest" from "no timber harvest" respondents.

Survey Responses Designed as Variables	Coefficients
Multi-year plan or not? (categorical, "1" for "Yes", "0" for "No")	0.513
Farmer or not? (categorical, "1" for "Yes", "0" for "No")	0.418
Own any land outside the study area? (categorical, "1" for "Yes", "0" for "No")	-0.417
Years of residency (numeric)	0.414
Area of land in acres (numeric)	-0.382

Table 4.2 Fisher's linear discriminant function coefficients distinguishing "timber harvest" from "no timber harvest" respondents.

Variables	LDA functions	
	1 (Timber harvest)	2 (No timber harvest)
Multi-year plan or not?	3.016	2.158
Farmer or not?	1.350	0.451
Own any land outside the study area?	0.840	1.626
Years of residency	0.106	0.083
Area of land in acres	0.001	0.000
Constant	-3.162	-2.222

Table 4.3 Classification results of discriminant function, distinguishing "timber harvest" from "no timber harvest" respondents.

Actual		Predicted		
		1 (Timber harvest)	2 (No timber harvest)	Total
Re-substitution	1 (Timber harvest)	165	78	243
	2 (No timber harvest)	87	165	252
Cross-validation	1 (Timber harvest)	158	85	243
	2 (No timber harvest)	89	163	252

One hundred and sixty-five of 243 respondents from the timber harvest group were correctly classified, and 165 of 252 respondents from the group who did not harvest timber were correctly classified (Table 4.3). The overall correct classification rate was 66.7 percent. The "leave-one-out" cross-validation classification rate (Braga-Neto et al. 2004) was slightly lower at 64.8 percent. Given these classification rates, it seems plausible using LDA to identify factors that distinguish forest landowners who harvest timber from those who do not harvest timber.

Because the survey data set contained several categorical variables, the assumption of multi-normality was not satisfied. Box's M test (Geisser and Greenhouse 1958) was statistically significant at $\alpha = 0.05$, which meant there were no homogeneous covariance matrices. QDA was used, but the classification rate did not improve. Therefore, the common covariance matrix to build LDA functions was assumed in the analysis.

4.3.2 Comparing "Commercial Timber Harvest" with "Non-commercial Timber Harvest" Responses

Three variables were identified as being statistically significant ($\alpha = 0.05$) using the stepwise procedure. The "importance of income from timber harvest" was the most significant variable (Table 4.4 and

Table 4.5). The "the number of tracts" owned by the landowners and "years of residency" were also significant predictors.

Approximately 62.2 percent of the 111 commercial timber harvest respondents were classified correctly. Nearly 81.7 percent of the 126 non-commercial timber harvest respondents

Table 4.4 Standardized canonical discriminant function coefficients distinguishing "commercial timber harvest" from "non-commercial timber harvest" respondents.

Survey Responses Designed as Variables	Coefficients
Importance of income from timber harvest (categorical, from 1 "not important" to 5 "very important")	0.663
The number of tracts (numeric)	0.550
Years of residency (numeric)	0.349

Table 4.5 Fisher's linear discriminant function coefficients distinguishing "commercial timber harvest" from "non-commercial timber harvest" respondents.

Variables	LDA functions	
	1 (Commercial)	2 (Non-commercial)
Importance of income from timber harvest	1.750	1.092
The number of tracts	1.154	0.734
Years of residency	0.078	0.055
constant	-5.223	-2.506

Table 4.6 Classification results of discriminant function distinguishing "commercial timber harvest" from "non-commercial timber harvest" respondents.

Actual		Predicted		
		1 (Commercial)	2 (Non-commercial)	Total
Re-substitution	1 (Commercial)	69	42	111
	2 (Non-commercial)	23	103	126
Cross-validation	1 (Commercial)	69	42	111
	2 (Non-commercial)	23	103	126

were correctly assigned to the group. The overall "leave-one-out" cross-validation classification rate was 72.6 percent (

Table 4.6).

4.3.3 Comparing "Commercial Harvest" with All Other Responses

Five variables were chosen as being statistically significant ($\alpha = 0.05$) using the stepwise procedure to build the linear discriminant function to separate respondents conducting a commercial timber harvest from all other respondents (Table 4.7). Similar significant variables occurred for this model as in previous models, e.g., "importance of income from the timber harvest," "years of residency," "farmer identification," "multi-year plan," and "area of land." Linear discriminant function coefficients further support these results (

Table 4.8). The correct classification rate was 82.0 percent from re-substitution and 81.6 percent from cross-validation (

Table 4.9).

4.4 The Result of Classification Trees

4.4.1 Comparing "Timber Harvest" with "No Timber Harvest" Responses

Thirteen CTs were initially developed using different combinations of split methods, variable selection methods, and split-point selection methods (Table 4.10). The split method of linear combination splits yielded the largest correct classification rate of 62.0 percent from

cross-validation, but the improvement was not distinct compared with 60.4 percent using univariate split (Kim and Loh 2001, Kim and Loh 2003). The split method of univariate splits with

Table 4.7 Standardized canonical discriminant function coefficients distinguishing "commercial timber harvest" from all other respondents.

Survey Responses Designed as Variables	Coefficients
Importance of income from timber harvest (categorical, from 1 "not important" to 5 "very important")	0.532
Years of residency (numeric)	0.436
Farmer or not? (categorical, "1" for "Yes", "0" for "No")	0.371
Multi-year plan or not? (categorical, "1" for "Yes", "0" for "No")	0.301
Area of land in acres (numeric)	0.263

Table 4.8 Fisher's linear discriminant function coefficients distinguishing "commercial timber harvest" from all other respondents.

Variables	LDA functions	
	1 (Commercial harvest)	2 (All others)
Importance of income from timber harvest	1.960	1.328
Years of residency	0.121	0.082
Farmer or not?	1.428	0.327
Multi-year plan or not?	3.269	2.316
Area of land in acres	0.001	-0.001
constant	-6.830	-2.596

Table 4.9 Classification results of discriminant function distinguishing "commercial timber harvest" from all other respondents.

Actual		Predicted		
		1 (Commercial harvest)	2 (All others)	Total

Re-substitution	1 (Commercial harvest)	45	66	111
	2 (All others)	22	356	378
Cross-validation	1 (Commercial harvest)	44	67	111
	2 (All others)	23	355	378

Table 4.10 Comparison of the performance of 13 classification trees of comparing "timber harvest" with "no timber harvest" respondents[1].

Split type	Variable selection method	Split method/ Pairwise variable selection method	# of terminal nodes of 1-SE tree/0-SE tree	Important variables	Re-substitution misclassification rate	Cross-validation misclassification rate
Univariate splits	1D	Exhaustive search	5/5	q44_farmer;q41_yrres; q14_plan;q36_lndout; q14_plan;	0.3333	0.3960
		Linear discriminant analysis	10/10	q44_farmer; q41_yrres; q14_plan; q36_lndout; q38_live; q54_income;	0.3030	0.4000
	2D	Exhaustive search	6/6	q39_arealife; q1_acres; q2_pctwd; q50_yrborn	0.3010	0.4000
		Linear discriminant analysis	4/8	q39_arealife; q1_acres	0.3535	0.3919
Linear combination splits			2/2		0.2909	0.3798
Univariate splits with node models	1D	Exhaustive search / MANOVA	1/1	q39_arealife; q45_impwdinc	0.3859	0.3939
		Exhaustive search / LDF	5/6	q44_farmer; q41_yrres; q14_plan; q36_lndout	0.2485	0.4040
		Linear discriminant analysis / MANOVA	1/1	q39_arealife q45_impwdinc	0.3859	0.3899
		Linear discriminant analysis / LDF	2/2	q44_farmer; q3_tracts q54_income; q30_prptxfair q45_impwdinc	0.3273	0.4141
	2D	Exhaustive search / MANOVA	1/2	q39_arealife q45_impwdinc	0.3859	0.3939
		Exhaustive search / LDF	2/4	q39_arealife; q43_employ q54_income; q14_plan q54_income	0.3111	0.4182
		Linear discriminant analysis / MANOVA	1/1	q39_arealife q45_impwdinc	0.3859	0.3899
		Linear discriminant analysis / LDF	3/3	q44_farmer; q14_plan; q3_tracts q54_income; q45_impwdinc q48_educ	0.2909	0.4323

[1] "0-SE tree" is the full tree with the smallest cross-validation (CV) estimate of error; "1-SE tree" is the smallest subtree with CV estimate of error within one standard error of the minimum.

node models did not improve prediction performance in this case. Therefore, the split method of univariate splits was selected for the final CT model.

The final CT model presented for comparing "timber harvest" with "no timber harvest" responses came from the following CT methods:

- Split type: Univariate splits
- Variable Selection method: One dimensional
- Split-point selection method: Exhaustive search

The important variables identified in the CT model (Figure 4.1) are the same as those in the corresponding LDA model (Table 4.1):

- Farmer or not? (categorical, "1" for "Yes", "0" for "No")
- Years of residency (numeric)
- Multi-year plan or not? (categorical, "1" for "Yes", "0" for "No")
- Own any land outside the study area? (categorical, "1" for "Yes", "0" for "No")

Of all 495 responses, 243 landowners harvested timber and 252 did not harvest timber. The most significant variable of the CT was "farmer identification." Sixty-six of the 90 farmers (73.3 percent) responding to the survey harvested timber. This result does not conflict with the LDA results discussed in the previous section. However, CTs provide much more detailed information on interactions within the group of "non-farmers" than does LDA. Within the group of non-farmers conducting timber harvests, the most significant characteristic (α = 0.05) was "years of residency." For non-farmers who lived at their residence for less than or equal to 36.5

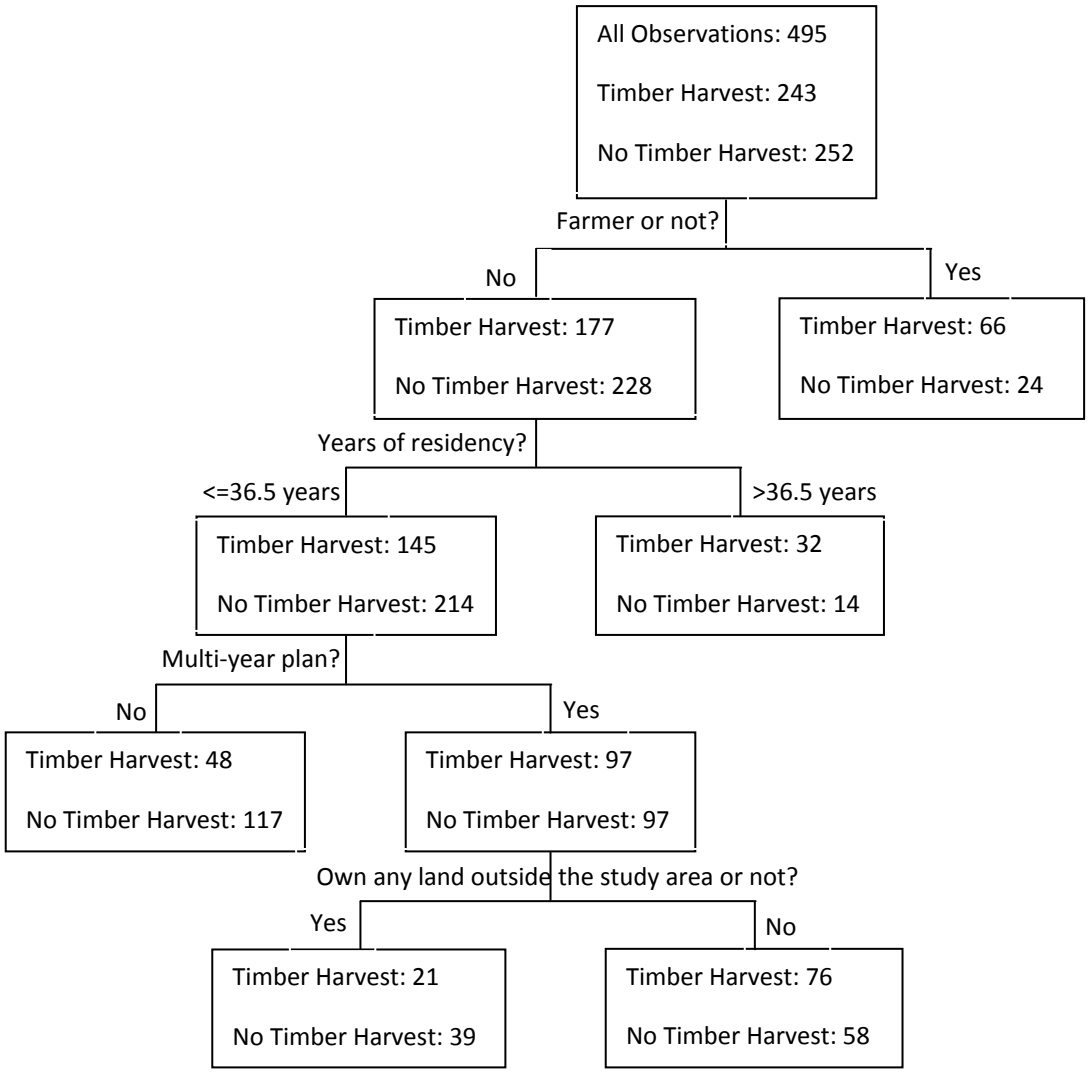

Figure 4.1 The 1-SE classification tree was built using CRUISE; comparing “timber harvest” with “no timber harvest” respondents.

years, only 145 of 359 (40.4 percent) respondents in this subgroup had harvested timber. For non-farmers who lived at their residence for less than or equal to 36.5 years and had a multi-year management plan, 97 of the 194 respondents (50.0 percent) in this subgroup had harvested timber. If the non-farmer lived at their residence for less than or equal to 36.5 years, had a multi-year management plan, and did not own any other land outside the study area, 76 of 134 of this respondent subgroup (about 56.7 percent) had harvested timber. The optimal CT correctly classified 330 of the 495 respondents. The re-substitution correct classification rate of 66.7 percent was the same as that of discriminant analysis.

4.4.2 Comparing "Commercial Timber Harvest" with "Non-commercial Timber Harvest" Responses

Of the 243 respondents who harvested timber, 111 conducted a "commercial timber harvest," and 126 conducted a "non-commercial timber harvest." Six landowners did not respond to the type of timber harvests. Several combinations of split method and variable selection methods were compared before developing an optimal CT (Table 4.11). The optimal CT came from univariate split and the one dimensional variable selection method.

The only significant variable ($\alpha = 0.05$) in this CT (Figure 4.2) was "importance level of income" expected from the timber harvest. Recall that the respondents were asked to rank importance based on an ordinal scale from one to five (i.e., 1 = "not important," 2 = "of little importance," 3 = "somewhat important," 4 = "important," and 5 = "very important"). The split point for this tree occurred at the ordinal rank of "not important" and all the other levels were

Table 4.11 Comparison of the performance of potential classification trees based on 243 observations who harvested timber before[2].

Split type	Variable selection method	Split method		# of terminal nodes of 1-SE tree/0-SE tree	Important variables	Re-substitution misclassification rate	Cross-validation misclassification rate
Univariate splits	1D	Exhaustive search		2/19	q45_impwdinc	0.2911	0.3249
	2D	Exhaustive search		2/2	Q1_acres	0.3038	0.3333
Linear combination splits				2/2		0.2152	0.3038
Univariate splits with node models	1D	Exhaustive search	MANOVA	1	q3_tracts; q45_impwdinc	0.2827	0.2869
	2D	Exhaustive search	MANOVA	1	q3_tracts; q45_impwdinc	0.2827	0.2911

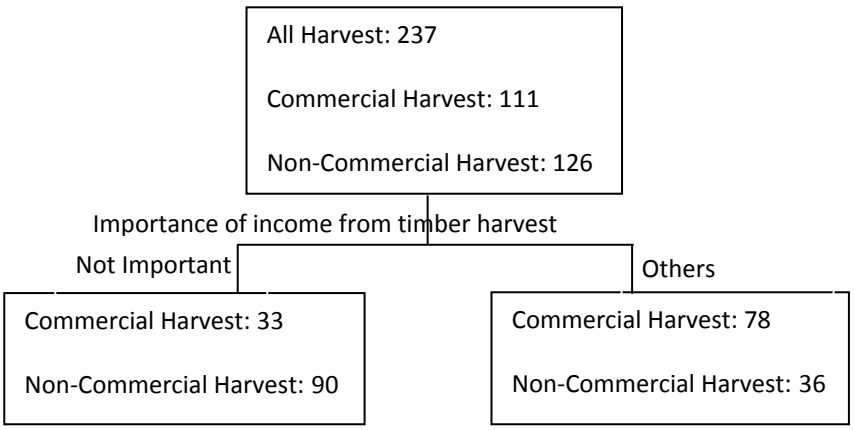

Figure 4.2 The 1-SE classification tree was built using CRUISE; comparing "commercial timber harvest" with "non-commercial timber harvest" respondents.

[2] Some combinations of variable selection methods and splits methods were ignored in Table 2 because from Table 1 we knew those combinations would not make any improvement.

combined in one subgroup. For those who considered that income from a timber harvest was at least "somewhat important," more than 68.4 percent conducted a commercial timber harvest. The CT correctly classified 70.9 percent of respondents who had conducted a timber harvest. This classification rate was 72.6 percent using LDA. In case of trees with only one split for "importance of income," there does not seem to be any additional information provided from the CT relative to the LDA model.

4.4.3 Comparing "Commercial Timber Harvest" with All Other Reponses

The usefulness of CTs is further illustrated by the third tree constructed by comparing "commercial timber harvest" respondents with all other respondents (Figure 4.3). The "importance level of income" from the timber harvest is still the most significant variable (α = 0.05), influencing the decision to conduct a commercial timber harvest (22.4 percent of all respondents). Seventy-two percent of respondents who conducted a commercial timber harvest considered that "importance of income" was at least "somewhat important." For the subgroup where "importance of income" was at least "somewhat important" and respondents had a "multi-year management plan," the "years of residency" was influential on the decision to conduct a commercial timber harvest. For the subgroup of the respondents who have lived in their current residence for more than 16.5 years, 40 of 62 in this subgroup (64.5 percent) had conducted a commercial timber harvest. Surprisingly, "farmer identification" is less important (i.e., declines to lower branches of the tree) as a split variable, when respondents conducting a commercial timber harvest are compared with all other respondents. The CT correctly classified 52.3 percent of respondents conducting a commercial timber harvests and 92.6 percent of all

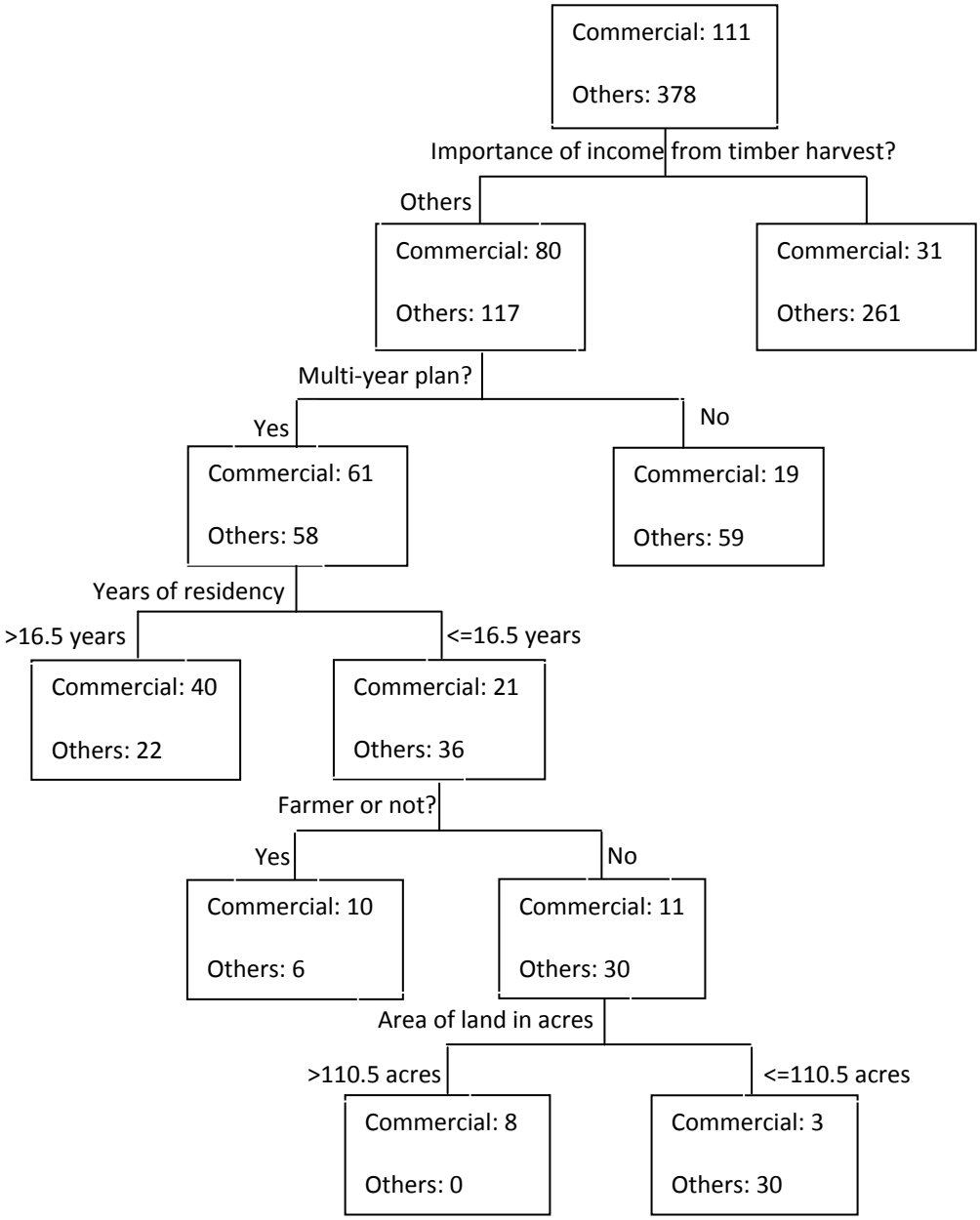

Figure 4.3 The 1-SE classification tree was built using CRUISE; comparing "commercial timber harvest" with all other respondents.

other respondents. This CT model discriminated the “non-commercial harvest” and “non-harvest” landowners quite well. The overall classification rate was 83.4 percent, which was slightly higher than the 82.0 percent of LDA model.

The CTs discussed above identify detailed subgroups and interactions, which are not given by LDA models alone. These tree structures may provide practical information for foresters and land managers in identifying characteristics of forest landowners who are likely to conduct commercial timber harvests and also in discriminating forest landowners who are not likely to conduct commercial timber harvests.

4.5 Conclusion

In this chapter, LDA and CT methods were compared using a survey of forest landowners as a case study. The survey consisted of 495 private forest landowners located in the Northern Cumberland Plateau region of Tennessee. Three pairs of LDA models and CTs were constructed and compared. LDA models and CTs identified approximately the same significant variables and had similar classification rates.

Survey results showed that 73.3 percent of farmer forest landowners harvested timber, and 69.6 percent of non-farmers who had a length of residency beyond 36.5 years harvested timber. For forest landowners who conducted commercial timber harvests, the importance level of income from the harvest was the overriding factor relative to all other factors. Discriminant analysis results supported the results of CTs. However, the linear discrimination

functions and corresponding coefficients did not provide the level of two-dimensional detail of CTs, which also detected hidden interactions.

Given the stringent assumptions of LDA, the CT methods may offer some advantage in robustness for data structure, in capability for detecting interactions between independent variables, in two-dimensional representation of the discriminating rules, and in straightforward interpretation. These methods provide foresters and land managers with objective, and scientific-based tools to assess characteristics of forest landowners likely to harvest tree commercially. These methods also characterize forest landowners not likely to harvest timber.

Chapter 5

5. Optimal Biorefinery Sites in TX and LA Based on Trucking Transportation Costs

5.1 Introduction

Energy Information Agency (2006) predicted that by 2020, the world's energy consumption will be 40 percent higher than it is today. Oil is located in high-cost, complex geopolitical environments and as noted in the U.S. Forest Service Research and Development Strategic Plan for the next several years (2006), decreasing economic dependence on conventional energy supplies will necessitate the need for improved energy efficiency and conservation. Energy efficiency and conservation are only part of the solution to meet the demand for more energy. Liquid fuels and value-added chemicals from biomass will be required to meet the greater energy demand and represent low-risk solutions to providing a renewable, sustainable and secure domestic energy supply.

U.S. Forest Service Southern Research Station and University of Tennessee Forest Product Center are supporting a large research project on developing a real-time, web-based optimal siting system for biomass energy producing and distribution facilities for the 33 Eastern United States (Hodges et al. 2007). The focus of this proposal is on identifying and projecting spatial comparative advantage for delivered wood and agricultural fiber costs based on resource costs, logging costs and transportation costs. While several prior studies have examined the availability of wood for biomass or the transportation costs associated with

cellulosic biomass (Young et al. 1991, Noon and Day 1997, Jensen et al. 2002, Perlak et al. 2005, Langholtz et al. 2006), this project will be the first to incorporate all costs associated with providing woody biomass to biorefineries and to evaluate the economic supply from all sources (e.g., standing trees, logging residues, mill residues, etc.).

This project will develop geo-referenced estimates of logging and transportation costs and incorporate these costs into an existing timber supply model to develop cost curves for woody biomass delivered to biorefineries at different locations throughout the southern United States. This project consists of three modules that will be aggregated to provide spatially explicit estimates of standing volume and residues available for bioenergy: 1) Sub-Regional Timber Supply (SRTS) model (Abt et al. 2000) to utilize USDA Forest Service inventory data and an economic supply and demand framework to project timber inventory, supply, and price into the future; 2) a GIS-based logging cost model to estimate the costs associated with harvesting standing trees; and 3) a transportation model to calculate the optimal paths and costs of transporting biomass to potential biorefinery sites.

The strategy of ensuring long-term sustainable bioenergy is the assessment of the economic availability of woody and agricultural-derived biomass. The optimal siting strategy must demonstrate the sustainability of a feedstock that will support profitability of the facility (Young et al. 1991). The goal of the project is to develop a real-time, web-based siting model for optimally siting cellulosic biofuel facilities in the eastern U.S. The objectives of this project are:

- Develop a web-based biorefinery economic siting model for forest and agricultural resources that exists in public domain.
- Develop a Microsoft SQL© database of resource data (forest and agricultural feedstocks).
- Develop resource costs for database.
- Develop a transportation cost model for database.
- Develop a harvesting cost model for database.
- Develop a web-base software tool to search for optimal biorefinery sites (www.BioSAT.net).
- Update BioSAT real-time as new data becomes available.

The scope of this research covers 33 eastern United States. The smallest resolution is zip code.

In this chapter, mill residues are considered as the cellulosic feedstock and the sum of primary wood manufacturing mill residues in a 40 mile radius area is defined as "total mill residue". Supply curves using trucking transportation costs are constructed for each zip code based on the "Million Ton Supply" categories for woody feedstocks, where we assumes an annual wood consumption for a biorefinery to be one million tons mill residues. Transportation cost by truck are calculated and used to choose optimal biorefinery locations. The method is used to decide the optimal biorefinery locations in Texas (TX) and Louisiana (LA); the top five sites are presented. In the complete study, the methods will be generalized to the 33 eastern states. The flow chart of the supply curve method based on trucking costs is illustrated in Figure 5.1.

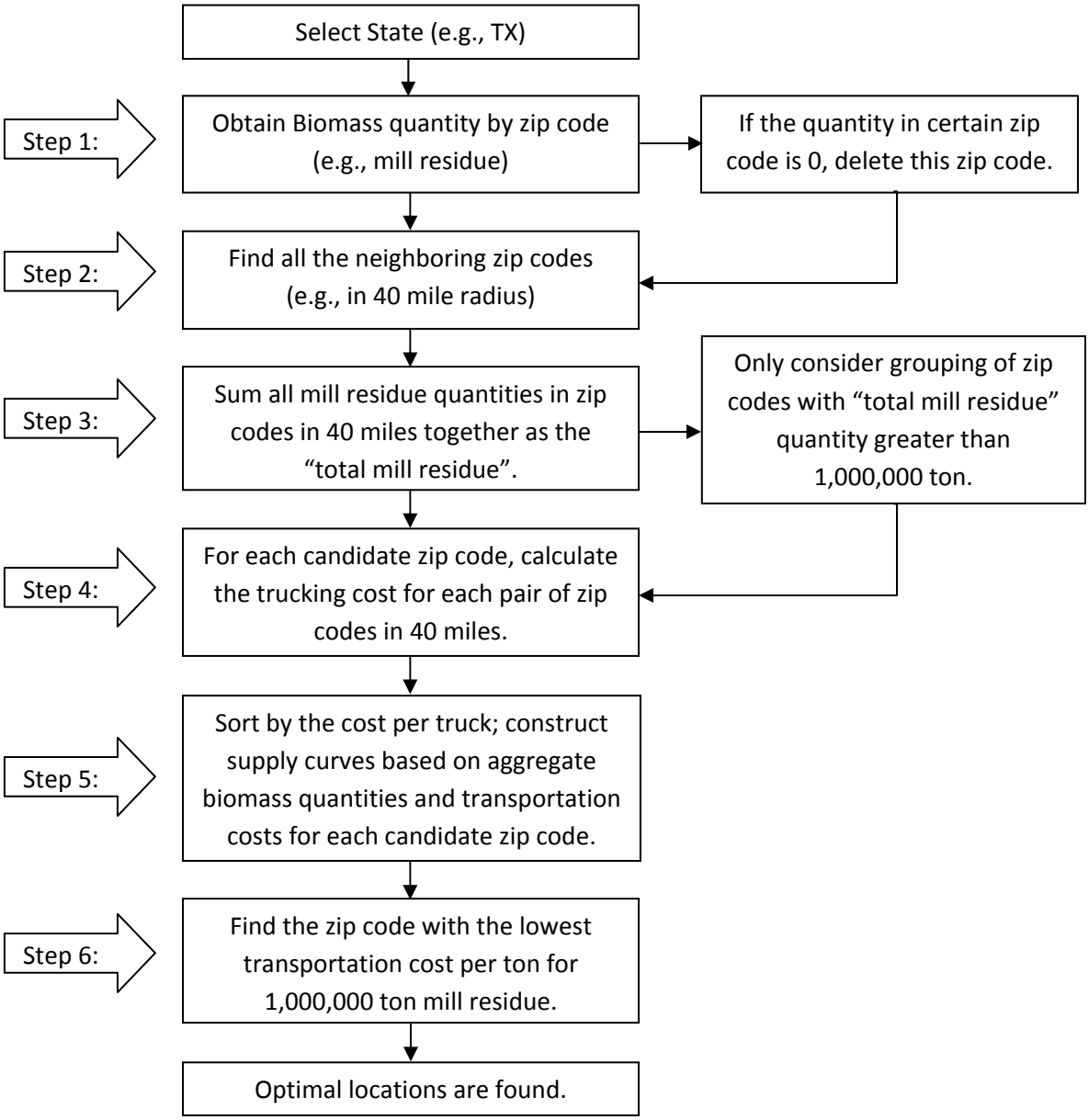

Figure 5.1 Flow chart of biorefinery siting methodology based on trucking cost.

5.2 Biomass Residue Quantities

In this stage, five categories of residues are considered as the feedstocks for the biorefinery facilities, which are logging residues, treatment thinning on timberland, urban wood waste, mill residues and other removals. Each category stands for an exclusive kind of woody residue sources that can be used as the cellulosic feedstocks to produce bioethanol in biorefineries.

- Logging residues: The unused portions of growing-stock and non-growing-stock trees cut or killed by logging and left in the woods.
- Treatment thinning on timberland: The material generated from fuel treatment operations and thinning designed to reduce the risk of loss to wildfire on timberlands.
- Urban wood waste: Urban wood wastes include wood (discarded furniture, pallets, containers, packaging materials and lumber scraps), yard and tree trimmings, and construction and demolition wood.
- Mill Residues: The Forest Service classifies primary mill residues into three categories: bark, coarse residues (chunks and slabs) and fine residues (shavings and sawdust).
- Other removals: Unutilized wood volume from cut or otherwise killed growing stock, from cultural operations such as pre-commercial thinning, or from timberland clearing. Does not include volume removed from inventory through reclassification of timberland to productive reserved forest land.

All the data are by zip code using USDA Forest Service Forest Inventory and Analysis data (2003-2005).

All these woody residue categories can be a significant source of bioenergy feedstock depending on location and concentration, type of material, acquisition, transportation and processing costs. Forestlands are distributed throughout the U.S. and the economics, site-specific characteristics and costs affect the recoverability of logging residues, thinning, and other removals. Primary wood manufacturing mill residues are a relatively convenient and stable source of biomass for cellulosic ethanol because they tend to be clean, uniform, concentrated, have low moisture content, and do not require harvesting costs. These traits make such mill residues desirable feedstocks for energy and biomass needs. The following analysis focused on the mill residues to find the optimal locations of biorefineries in TX and LA, see Figure 5.2.

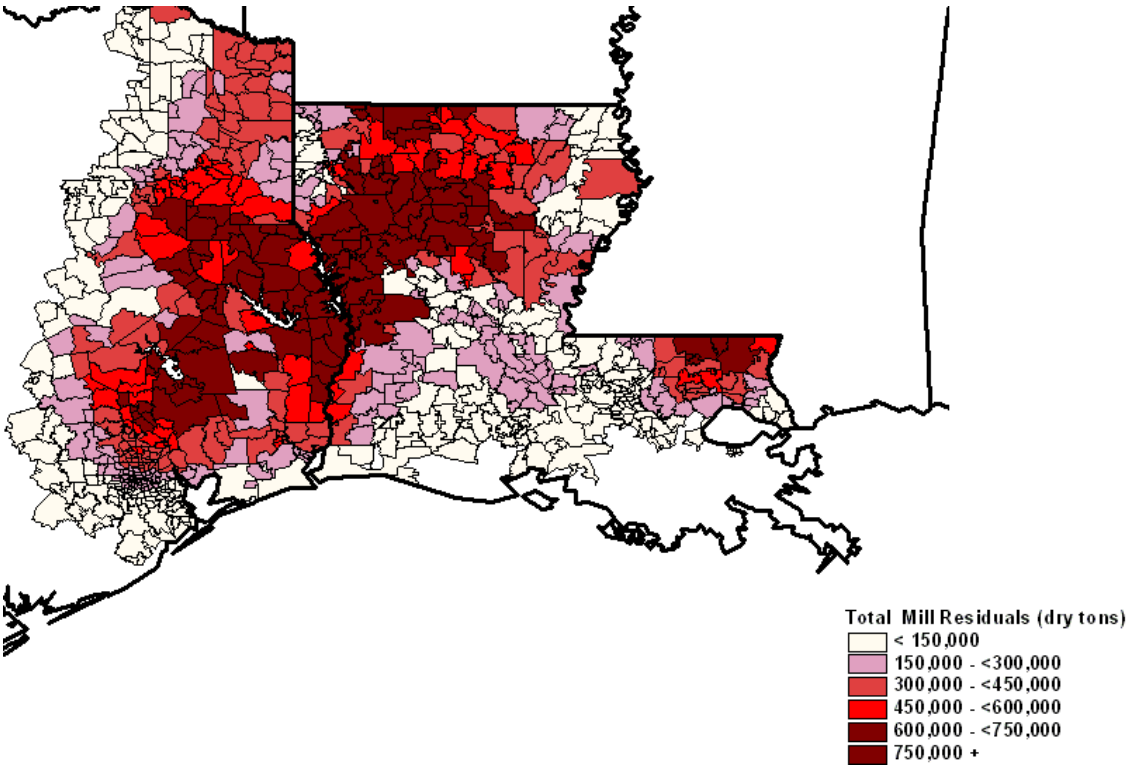

Figure 5.2 The distribution of mill residues quantities in LA and TX.

5.3 Calculating Trucking Transportation Costs

Transportation cost, the cost of moving feedstock or products, is an important component of the overall costs for recovering energy from biomass. It's a very important issue in the decision of locating for bioenergy facilities. Transportation cost typically represents a substantial portion of the total costs of woody biomass, due to the low value of the wood, distances traveled for delivery, and increasing price of diesel fuel. Evaluating the economic feasibility of bioenergy requires comprehensively addressing transportation costs. Transportation cost is highly determined by the geographical location of forests, mills and potential biorefinery facilities.

There are various transportation modes, truck, rail, freight, or the combination of them. In this chapter, trucking transportation costs are analyzed. Trucking costs can be segmented by fixed and variable costs. The components making up variable costs are fuel, labor, tires, and maintenance (Berwick and Farooq 2003).

The fuel price is an external variable that depends on the current market rates, which may vary depend on supply and demand conditions for a geographical location. Meanwhile, fuel economy is a function of engine horsepower, speed, terrain, wind and weight; and the speed is a function of engine horsepower, terrain, wind and weight. In this study, the gross vehicle weight is set at 80,000 pounds, which is the current maximum limitation in TX and LA. The normal tractor and trailer weight are respectively 13,900 pounds and 23,700 pounds; therefore the full pay load is 42,400 pounds. The fuel cost for round trip is calculated as follows:

$$\text{fuel cost} = \left(\frac{\text{fuel price per gallon}}{5.182985} + \frac{\text{fuel price per gallon}}{6.64982}\right) \times \text{driving distance miles} \qquad (5.1)$$

where, 5.182985 is miles per gallon (MPG) when the truck is loaded, and 6.64982 is MPG when the truck is empty.

The labor cost is a readily known variable for paid drivers and is accounted for in the model by time. The average labor cost in LA and TX is set at $15.63/hour.

$$\text{labor cost} = 2 \times \text{labor cost per hour} \times \frac{\text{driving time min}}{60} \tag{5.2}$$

The tire cost consists of tire price and tire wear cost. Tires are weight sensitive and wear more with more weight. In average, the tire cost per mile is set at $0.06155/mile.

$$\text{tire cost} = 2 \times 0.06155 \times \text{driving distance miles} \tag{5.3}$$

The maintenance cost depends on the age of the equipment, weight and operating conditions. The average maintenance cost is estimated at $0.096/mile:

$$\text{maintenance cost} = 2 \times 0.096 \times \text{driving distance miles} \tag{5.4}$$

Components making up fixed costs are equipment costs, license fees, insurance, and overhead expenses. Fixed costs vary in different geographic area and by size of firm. These costs are totaled by category and are estimated on a per mile basis. The total cost per year is estimated at $44276 and the number of driving miles per year is assumed at 100,000 miles. Fix costs for a round trip are:

$$\text{fixed cost} = 2 \times 0.44 \times \text{driving distance miles} \tag{5.5}$$

Combining all variable and fixed costs in formulas (5.1) through (5.5), give the transportation costs per truck for a round trip haul between any pair of zip codes as:

$$\text{cost per truck round trip} =$$
$$\left(\frac{\text{fuel price per gallon}}{5.182985} + \frac{\text{fuel price per gallon}}{6.64982}\right) \times \text{driving distance miles}$$
$$+ 2 \times \text{labor cost per hour} \times \frac{\text{driving time min}}{60}$$
$$+ 2 \times 0.06155 \times \text{driving distance miles}$$
$$+ 2 \times 0.096 \times \text{driving distance miles}$$
$$+ 2 \times 0.44 \times \text{driving distance miles} \qquad (5.6)$$

where, the unit of driving distance is mile, and the unit of driving time is minute;

fuel price per gallon is set at $4.66/gallon;

labor cost per hour is set at $15.63/hour;

and 5.182985 is MPG when the truck is loaded; 6.64982 is MPG when the truck is empty;

and 0.06155 is the estimated tire cost per mile;

and 0.096 is the estimated maintenance cost per mile;

and 0.44 is the estimated fixed cost per miles.

Note that the transportation cost calculated from formula (5.6) is the cost for transportation companies, it's not the real cost that the customers may pay if profit margins are included resulting in a trucking rate.

Total transportation cost is calculated by multiplying cost per truck for a round trip haul and the number of trucks needed to transport all mill residues to the destination. Note in

formula (5.7), the number of trucks needed is not ceiled, so the total cost may be slightly lower than the real cost:

$$\text{total transportaion cost} = \text{cost per truck round trip} \times \frac{\text{mill residue quantities}}{\frac{42{,}400}{2204.59}} \qquad (5.7)$$

where 42,400 pounds is the pay load of per truck; and one ton = 2204.59 pounds.

Dividing the total cost by the total mill residue quantities, gives the transportation cost per ton. This value is a key measure in selecting an optimal zip code.

$$\text{transportation cost per ton} = \frac{\text{total cost}}{\text{mill redsidue quantities}} \qquad (5.8)$$

5.4 Finding Neighboring Zip codes

In order to gather sufficient cellulosic feedstocks to satisfy the demand of a biorefinery, it is necessary to collect mill residue in the neighboring zip codes from the zip code where the potential biorefinery is located. The closer the supply locations, the lower the typical trucking transportation costs. This may not be true if a geographic feature such as a mountain, river, large city, or large water reservoir restrict transportation networks of the neighboring zip codes. In this research, zip codes within a 40 mile radius are analyzed as neighboring zip codes. The sum of the mill residues in a 40 mile radius is generally sufficient in eastern TX and LA to meet the annual demand of a biorefinery (Liu et al. 2008).

For a given zip code, its sphere distances to other zip code are calculated by utilizing the longitudes and latitudes (*http://en.wikipedia.org/wiki/Earth radius*, referenced 06/20/2008).

$$D = \sqrt{(Md\phi)^2 + (N\cos\phi\, d\lambda)^2} \tag{5.9}$$

where ϕ --- the mean latitude,

$d\phi$ --- difference in latitude,

$d\lambda$ --- difference of longitude (in radians),

M --- Earth's radius of curvature in the (north-south) meridian at ϕ,

N --- radius of curvature in the prime normal to M at ϕ.

All zip codes with a sphere distance no more than 40 miles are defined as neighboring zip codes of a centric zip code. Future research will include an 80 mile radius.

For each neighboring zip code, the driving distance and driving time to the given zip code is computed from MapPoint (Microsoft Inc. 2006). From equation (5.6), the transportation cost per truck for a round trip haul between any pair of zip codes is estimated using the driving distance and driving time.

5.5 Top Biorefineries Locations in LA and TX by Zip code

In any zip code for LA and TX, the quantity of mill residues available is less than 1,000,000 tons, the assumed maximum annual demand for a biorefinery that would produce 70,000 gallons of ethanol (Timber Mart South 2008). The sum of the mill residue quantities in the neighboring zip codes within a 40 mile radius is defined as the "total mill residue" of the center zip code where the biorefinery is located.

There are four zip codes, 75930, 75959, 75948, and 77331 in Texas, containing total mill residue quantities greater than 1,000,000 million tons (Table 5.1). These zip codes are highlighted in red color in Figure 5.3. These four zip codes are considered as potential biorefinery locations, and labeled as green dots in map in Figure 5.3.

The supply curves are estimated by sorting the mill residue data by transportation costs per truck for a round trip haul from a zip code with a biorefinery site. The four potential zip codes were examined using this method (Figure 5.4). The zip code 75928 has the lowest trucking cost per ton. For a supply of 1,000,000 annual tons mill residues, the transportation cost is approximately $5.20/ton for this zip code. This zip code also has the highest total of mill residues in Texas, which is 1,331,079 ton.

Table 5.1 The four zip codes with total mill residue quantities beyond 1,000,000 tons in Texas.

State	Zip Code	Total Mill Residue (ton)	Total Transportation Cost ($)
TX	75948	1331078.99	5895383.15
TX	75930	1105029.92	5019429.08
TX	77331	1021945.5	4821877.64
TX	75959	1021793.15	4546664.57

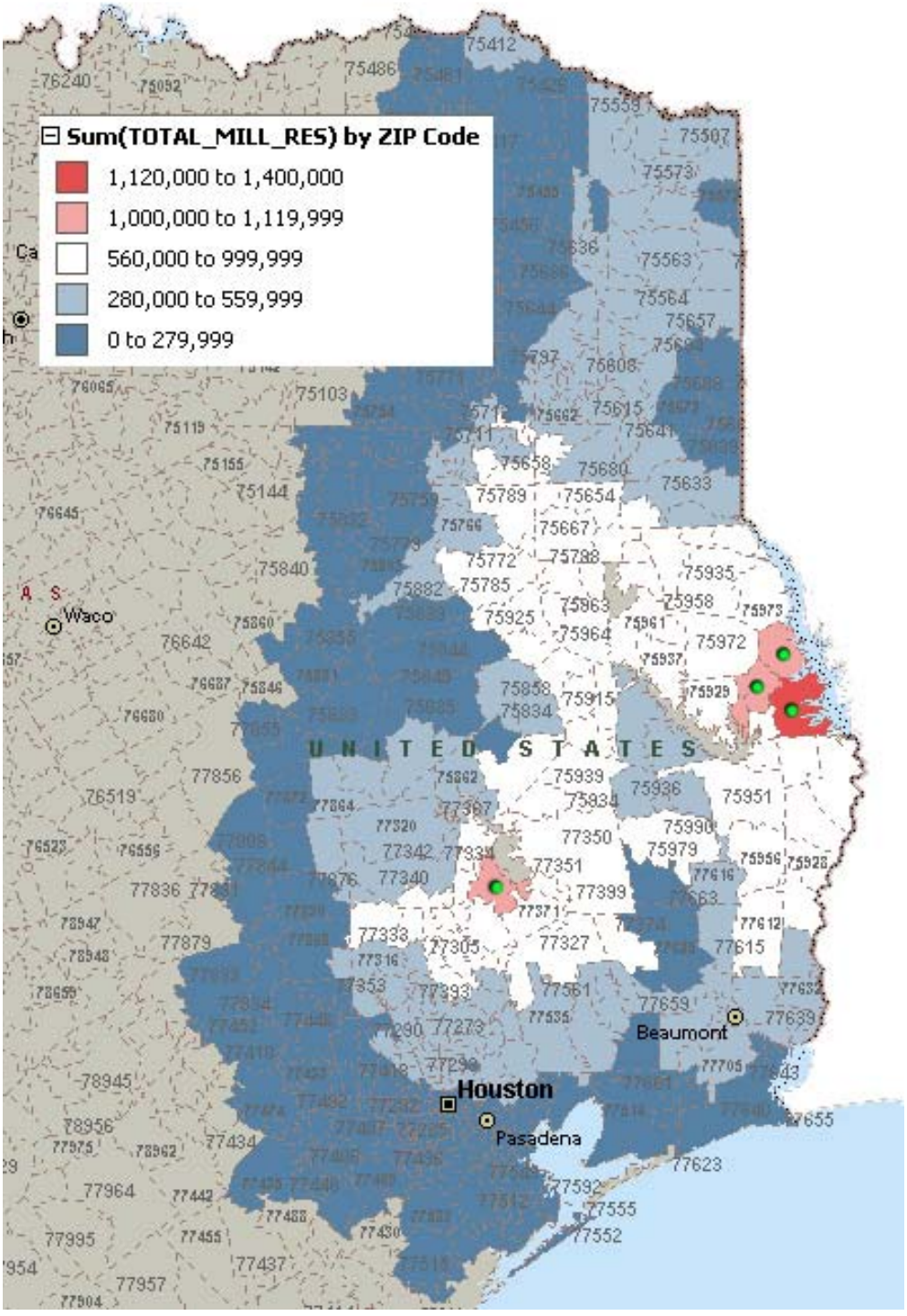

Figure 5.3 The distribution of total mill residues quantities within a 40 mile radius for eastern TX and the four optimal biorefinery locations.

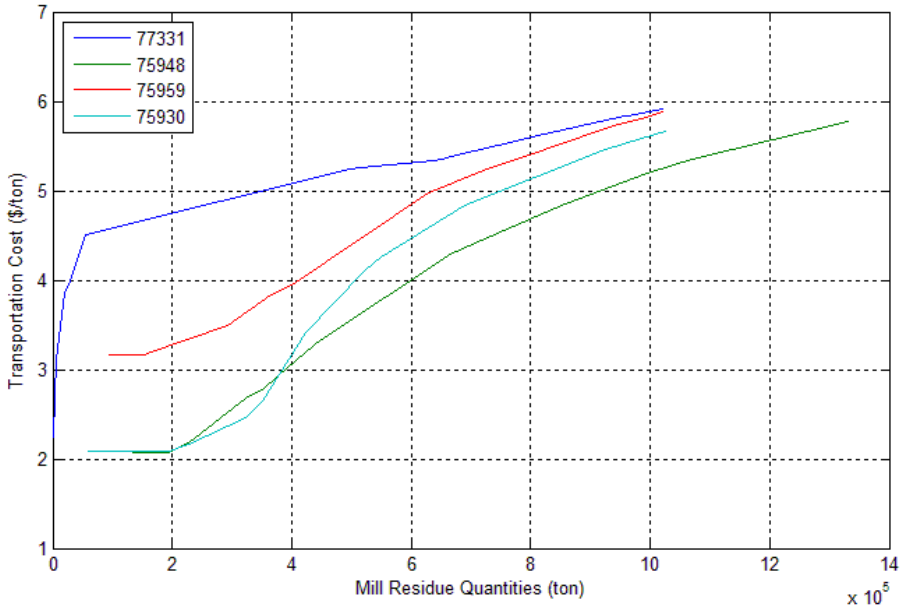

Figure 5.4 Supply curves for four candidate biorefinery locations in TX.

There are 19 zip codes containing total mill residue quantities greater than 1,000,000 tons for a 40 mile procurement radius in LA. They are, in decreasing order of mill residue quantities, 71460, 71462, 71406, 71426, 71449, 71065, 71486, 71458, 71429, 71497, 71457, 71002, 71414, 71066, 71063, 71419, 71038, 71070, and 71450 (Table 5.2), see the corresponding red areas in Figure 5.5. In the first step, five zip codes were picked up as potential sites, 71460, 71462, 714458, 71070, and 71038. These zip codes are located in dispersive geographic regions in LA and in that particular region correspond to higher total mill residue quantities (Figure 5.6). The supply curves for these five potential biorefinery locations are presented in Figure 5.7. The zip codes of 71460 and 71462 correspond to lower transportation costs of approximately $3.50/ton and $3.90/ton, respectively. When compared with the other three zip codes the transportation costs exceed $5.50/ton. From a geographic perspective, the zip codes 71460 and

71462 are both near the Toledo Bend Reservoir, while the other three zip codes are located far away from the Toledo Bend Reservoir (Figure 5.7). Given the transportation cost estimates, the zip codes near Toledo Bend Reservoir are examined in more detail. There are 11 zip codes near the Toledo Bend Reservoir containing total mill residue quantities greater than one million tons. Seven of these zip codes contain the total mill residue quantities greater than that of the zip code 71458, which has been examined in the initial step of the selection process. These seven zip codes will be the potential biorefinery candidates (Figure 5.8). The supply curves for these seven potential biorefinery locations are illustrated in Figure 5.9.

Table 5.2 The 19 zip codes with total mill residue quantities beyond 1,000,000 tons in Louisiana.

State	Zip Code	Total Mill Residue (ton)	Total Transportation Cost ($)
LA	71460	1365551.05	4673991.10
LA	71462	1288203.61	4621543.59
LA	71406	1275495.25	4466529.65
LA	71426	1273272.12	4242813.39
LA	71449	1272172.00	3665809.99
LA	71065	1266826.65	4636708.01
LA	71486	1206287.59	3847089.69
LA	71458	1174919.84	5444621.70
LA	71429	1165590.23	4033257.82
LA	71457	1107036.76	5092675.53
LA	71497	1107036.76	5082557.34
LA	71002	1101337.04	4879258.30
LA	71414	1097768.63	4860514.64
LA	71066	1087754.43	4832870.81
LA	71063	1017585.72	3994870.56
LA	71419	1014244.54	3646166.57
LA	71038	1013500.16	4887323.09
LA	71070	1002649.09	4355162.17
LA	71450	1000125.39	3738292.67

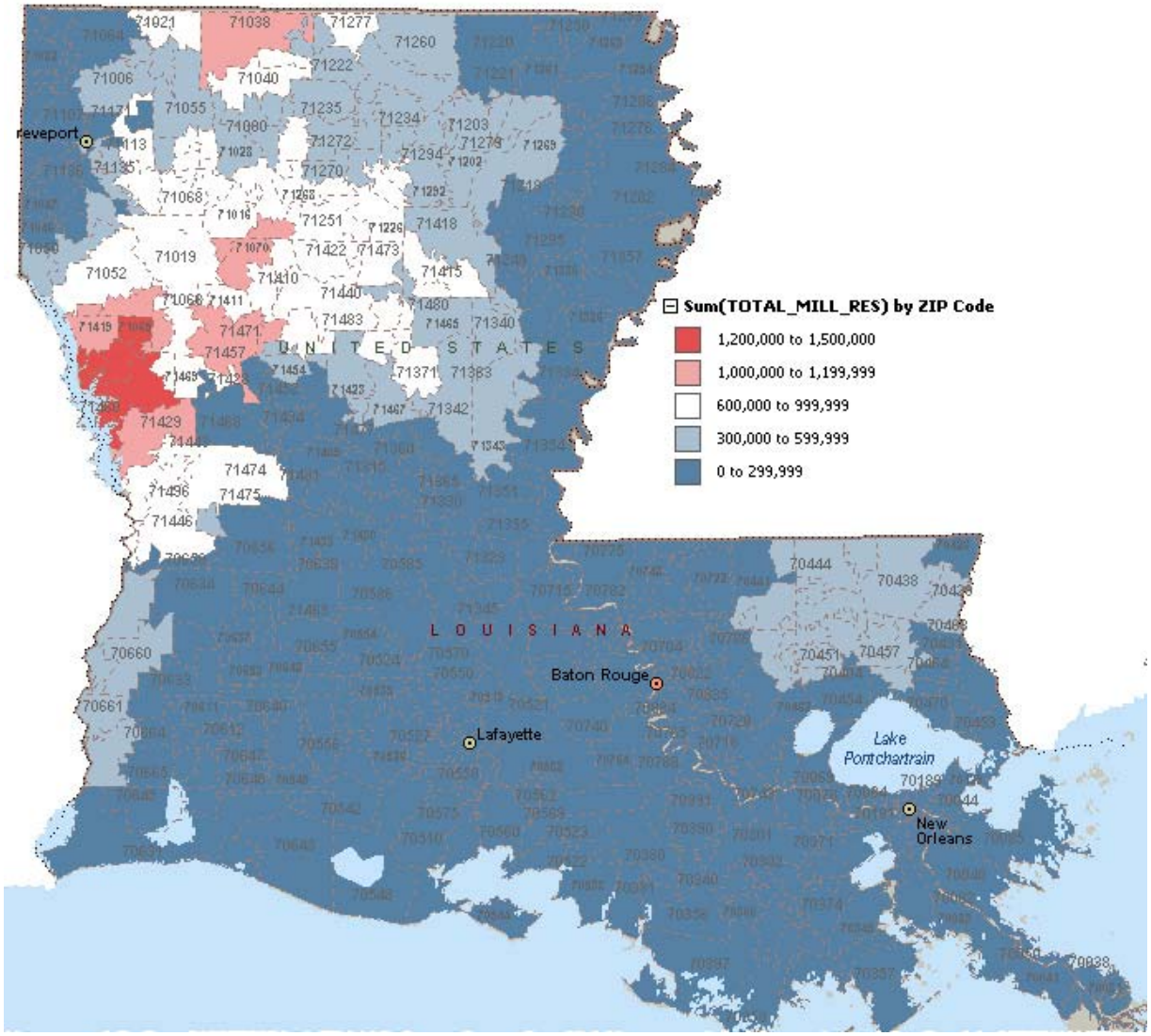

Figure 5.5 The distribution of total mill residues quantities within a 40 mile radius LA.

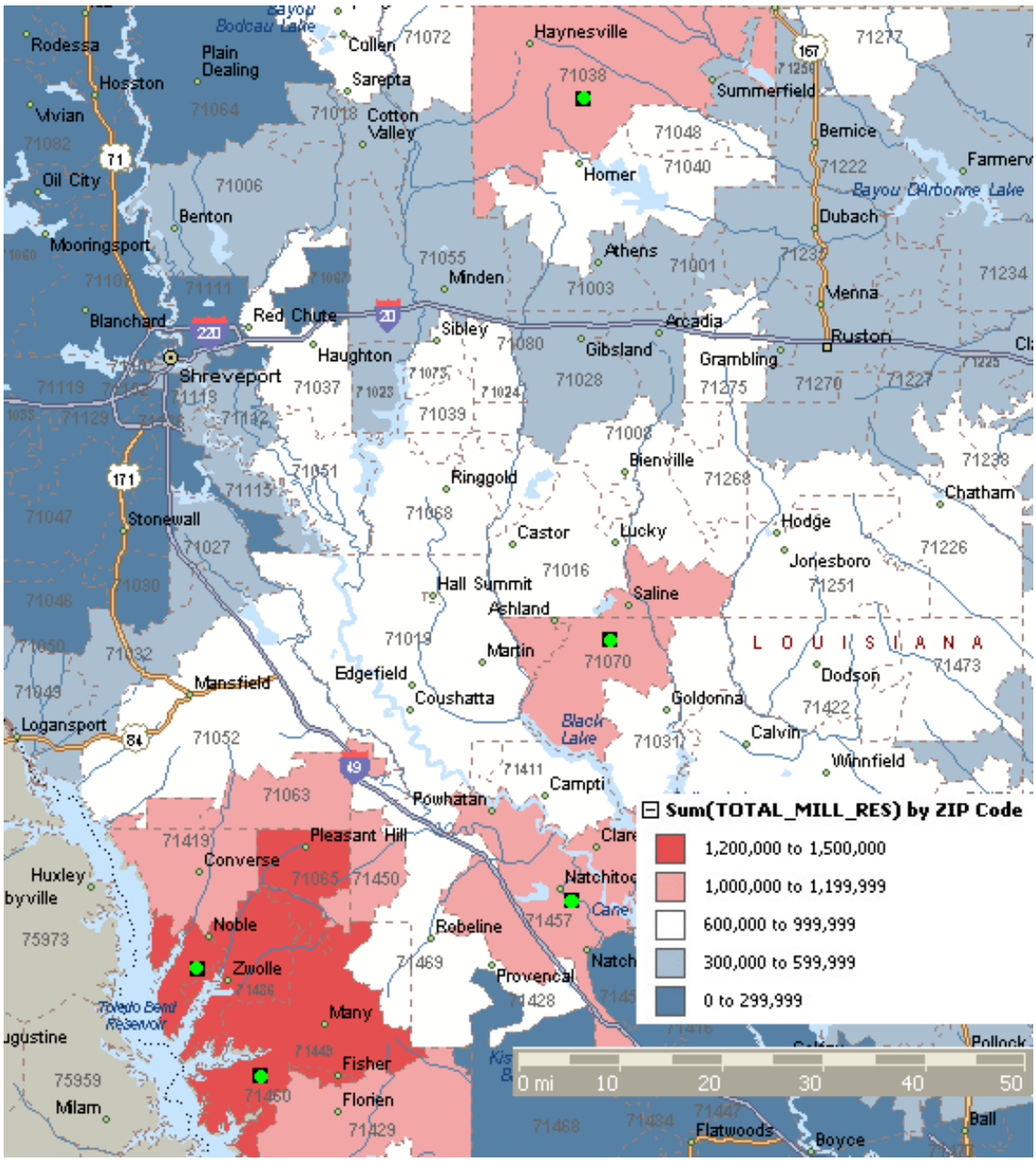

Figure 5.6 Five candidate bio-refinery locations in northwest LA based on total mill residue quantities and geographic dispersion.

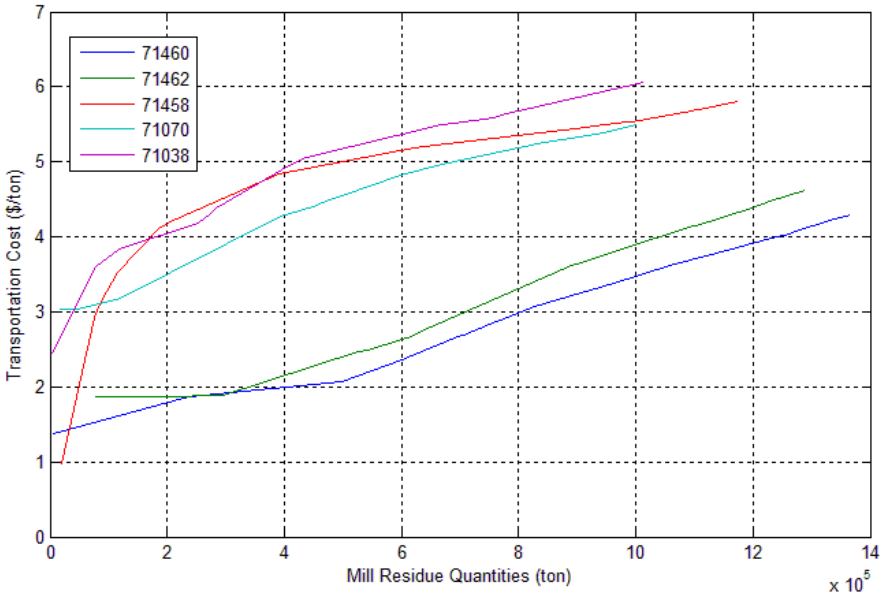

Figure 5.7 Supply curves for five potential bio-refinery locations in LA.

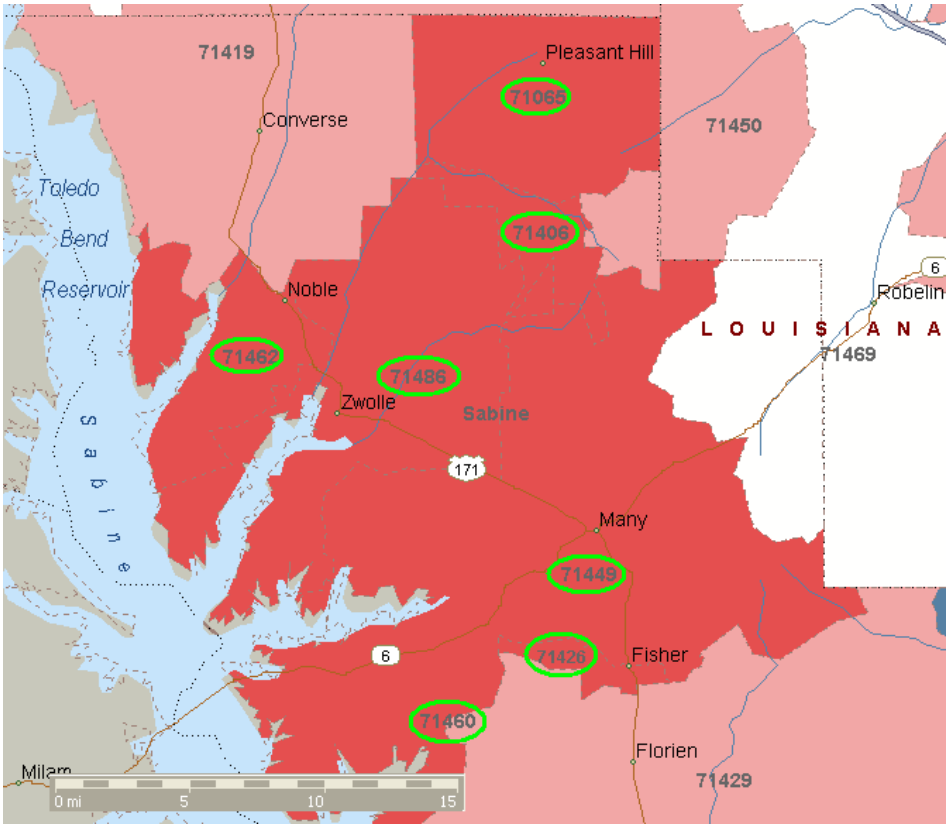

Figure 5.8 Toledo Bend Reservoir area seven potential biorefinery locations in LA.

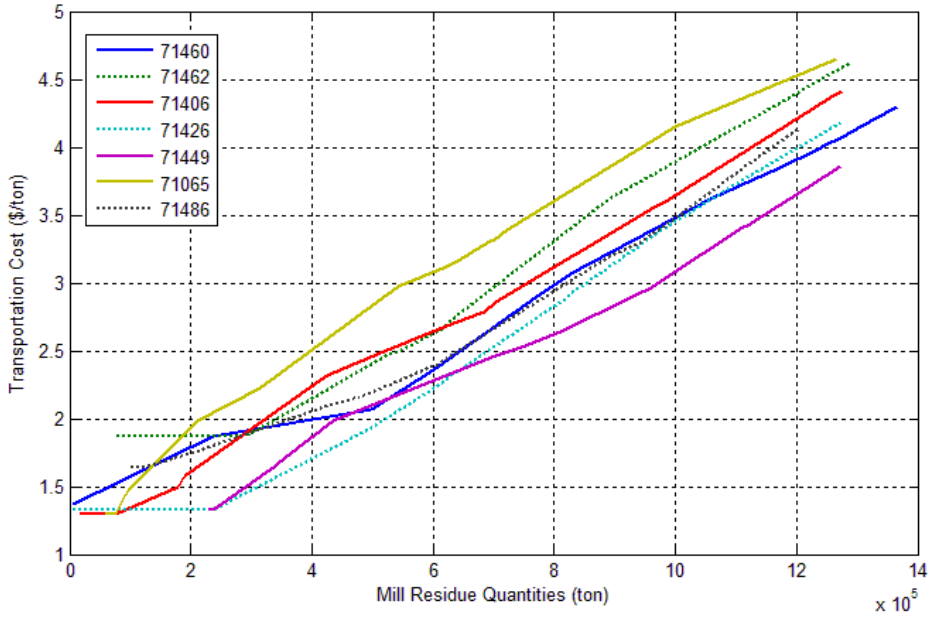

Figure 5.9 Supply curves for seven potential bio-refinery locations in LA near Toledo Bend Reservoir.

The zip code 71449 has the lowest transportation cost, which is approximately $3.10/ton (Table 5.3), followed by 71426, 71460 and 71486, with corresponding approximately transportation costs of $3.50/ton. The zip code 71406 is ranked fifth with transportation costs of approximately $3.60/ton, followed by zip code 71462 with $3.90/ton. The final ranked zip code is 71486 with transportation costs of approximately $4.20/ton. All these trucking costs are lower than that of the best candidate 75928 zip code in TX for a million tons of mill residues.

Based on the analysis of the supply curves, the best biorefinery site is zip code 71449, labeled as “A” in Figure 5.10; followed by zip codes 71426, 71460 and 71486, labeled as “B”; followed by zip code 71406, labeled as “C”. Those are the top five biorefinery locations in LA.

Table 5.3 The detailed information about the best bio-refinery location 71449.

Zip code	neighboring zipcode in 40 miles	driving time (min)	driving distance (mile)	mill residue (ton)	transportation cost per truck round trip ($)	transportation cost ($)	cumulating mill residue (ton)	cumulating cost ($)	cost per ton ($/ton)
71449	71449	0	0	230227	0	0	230,227	$305,502	$1.33
71449	71426	16.78	6.00	6163	25.52	8178.53	236,390	$313,680	$1.33
71449	71486	24.53	11.91	102506	46.07	245570.56	338,897	$559,251	$1.65
71449	71406	29.02	14.60	17137	55.91	49823.65	356,035	$609,075	$1.71
71449	71462	28.53	16.85	79014	61.97	254577.28	435,049	$863,652	$1.99
71449	71429	31.50	16.35	261393	62.10	844075.81	696,443	$1,707,728	$2.45
71449	71469	34.57	17.95	55254	68.18	195887.36	751,697	$1,903,615	$2.53
71449	71065	38.32	20.29	58709	76.68	234085.31	810,406	$2,137,700	$2.64
71449	71419	37.98	24.96	114184	89.55	531681.55	924,591	$2,669,382	$2.89
71449	71439	46.23	26.54	9829	98.26	50220.00	934,420	$2,719,602	$2.91
71449	71457	44.15	27.10	20267	98.74	104052.05	954,687	$2,823,654	$2.96
71449	71450	50.08	28.63	11594	106.12	63975.53	966,282	$2,887,629	$2.99
71449	71063	54.27	29.48	9283	110.66	53414.72	**975,565**	**$2,941,044**	**$3.01**
71449	75948	56.55	32.39	135151	120.00	843251.08	**1,110,717**	**$3,784,295**	**$3.41**
71449	71403	55.80	33.08	7061	121.53	44623.80	1,117,778	$3,828,919	$3.43
71449	75959	63.05	34.88	95571	130.33	647644.76	1,213,349	$4,476,564	$3.69
71449	71446	57.53	37.32	12	134.28	88.52	1,213,362	$4,476,652	$3.69
71449	75930	66.40	37.75	58810	140.10	428403.53	1,272,172	$4,905,056	$3.86

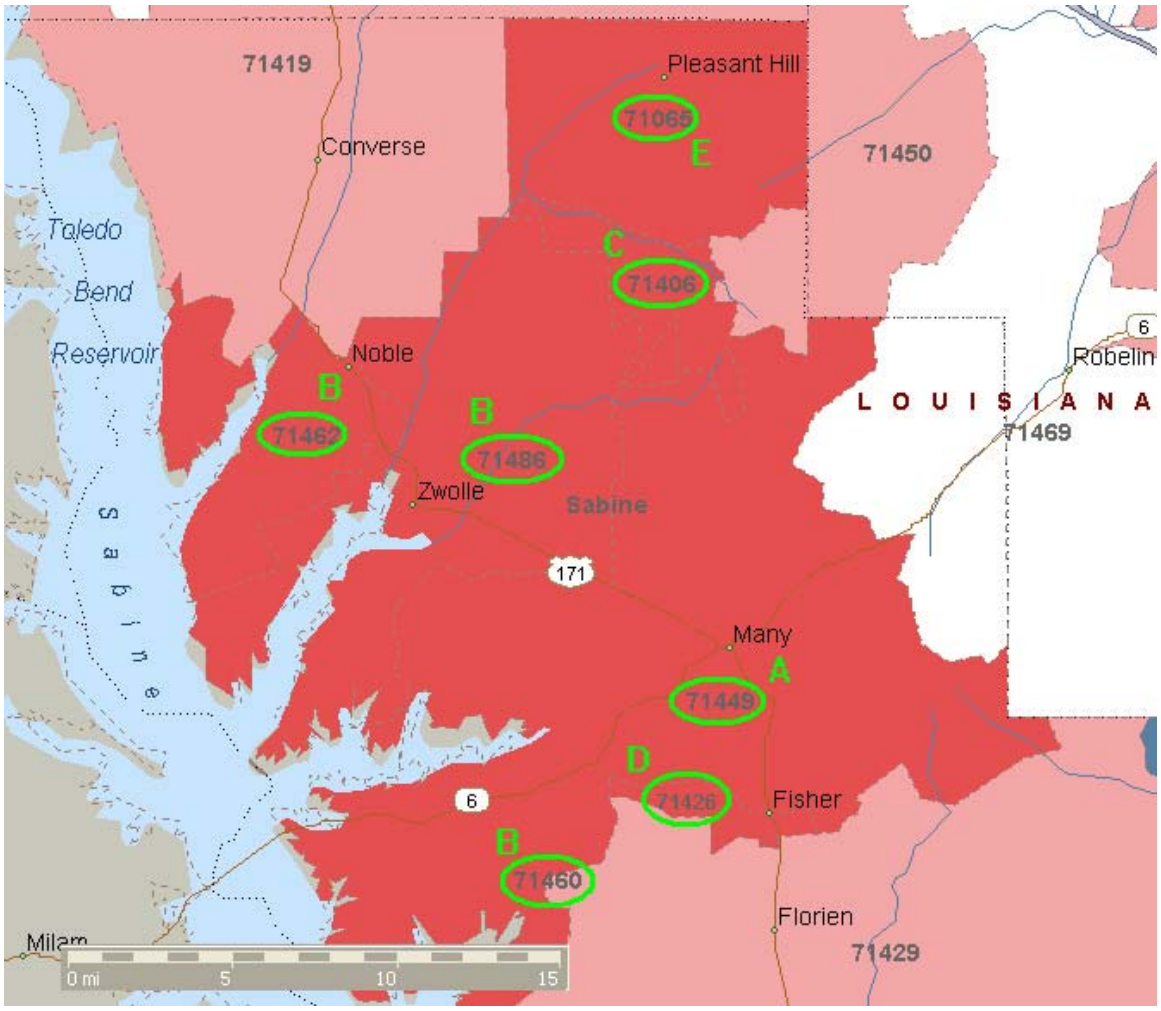

Figure 5.10 The top five biorefinery sites all in LA based on trucking transportation costs.

5.6 Conclusion

In this chapter, the top five biorefinery sites are identified in LA and TX based on the mill residue quantities and trucking costs. Supply curves are constructed to select the better optimal biorefinery locations. The top five locations by zip code are 71449, 71426, 71460, 71486, and 71406. All of these zip codes are near the Toledo Bend Reservoir, a boundary between LA and TX. The best zip code with the lowest trucking costs is 71449. For 1,000,000 tons of mill residue supply per year the transportation cost is approximately $3.10/ton. Note this cost estimate excludes profit margin. The real cost for a biorefinery facility may be substantially higher than the estimated trucking costs presented in this thesis if the biorefinery uses commercial truck hauling instead of trucks owned by the biorefinery plant. The ranking of the best biorefinery locations by zip code may change if profit margin is added, i.e., profit margin or trucking rate quotes vary by the availability of truck transport in a geographic region based on market conditions for commercial trucking.

In the complete study of biorefinery siting in the eastern U.S., forest and agriculture harvesting cost models will be developed with forest and agriculture resource cost database will be constructed. A Microsoft SQL© database will be developed to fuse all datasets in one database and a final modeling algorithm for locating optimal biorefineries in 33 eastern states will be developed. The website of *www.BioSAT.net* is in development which will be a user interface for this modeling system. Railroad networks with intra-model transfer points for approximately 38,000 zip codes will also be used to estimate transportation costs.

Chapter 6

6. Conclusions and Future Research

In the past biofuels and energy from biomass have been explored as an alternative to fossil fuels. The U.S., European, Chinese, and Indian economies are heavily dependent on fossil fuels and recent surges in oil prices are of serious consequence to these economies. The serious problem of greenhouse gases released into the atmosphere from the combustion of fossil fuels and the possible exhaustion of oil reserves have accelerated the interest by scientists, industry leaders, and politicians toward the development of new alternative sources of energy. Biofuels deriving from nonfood feedstocks offer a promising solution for substitution away from petroleum-based energy.

In this thesis, statistical classification methods are used to study the factors that influence landowner attitudes towards harvesting timber. Statistical classification is a procedure in which individual cases are sorted into groups based on one or more quantitative and/or qualitative characteristics of the cases. There are numerous techniques and algorithms for classification problems, such as logistic regression, linear discriminant analysis (LDA), cluster analysis, and classification trees (CT). Most software packages have these functions. Each technique has limitations and advantages. LDA and CT methods are used in this thesis to analyze a survey of 495 private forest landowners that was conducted in seven counties of Tennessee located in Northern Cumberland Plateau region. LDA models and CTs identified approximately the same significant variables and had similar classification rates. Classification

of survey results indicated that 73.3 percent of farmer forest landowners harvested timber, and 69.6 percent of non-farmers who had a length of residency beyond 36.5 years harvested timber. For forest landowners who conducted commercial timber harvests, the importance level of income from the harvest was the overriding factor relative to all other factors. Discriminant analysis results supported the results of CTs. However, the linear discrimination functions and corresponding coefficients did not provide the level of easy-to-interpret two-dimensional detail of CTs, which also detected hidden interactions. These methods provide foresters and land managers with objective and scientific-based tools to assess characteristics of forest landowners likely to harvest trees.

Finding optimal sites for biofuel refineries is an important issue for low cost and competitive biofuel production. Considering the environmental impacts, economic influences, political incentives, and availability of labor, biofuel refinery siting modeling is a complex procedure which has received substantial research support. Bioenergy production is highly geographically dependent on feedstock sources and transportation costs have a strong influence on optimal low cost locations. Many cellulosic biofuel plants derive biofuels from unused agricultural and woody residue. Given the low cost of the unused residues, transportation costs can account for a significant portion of the total biomass fuel costs. New plant sites selected in proximity to unused residues that can be procured at minimum transportation costs will offer strong economic advantages by minimizing raw material costs at the mill gate.

Using supply curves that are based on aggregate biomass quantities and transportation costs by truck provide a feasible method to identify optimal biorefinery locations. In this thesis, this supply curve method was used to find the top five zip code locations in Louisiana and Texas for biorefineries with an annual demand of 1,000,000 tons of mill residues from primary wood manufacturing facilities. The top five optimal zip code locations are 71449, 71426, 71460, 71486, and 71406, all near the Toledo Bend Reservoir, a boundary between Louisiana and Texas. The best zip code location with the lowest trucking cost is 71449. For 1,000,000 annual tons of mill residues, the transportation cost is approximately $3.10/ton. This cost does not include profit margin.

In future research, trucking costs that include profit margin, rail-trucking costs scenarios, consideration for other feedstocks (e.g., urban waste, logging residues, agricultural residues, etc.) will be assessed to estimate optimal biorefinery locations. Woody biomass harvesting costs and agricultural biomass harvesting costs will be added into a future siting model. A real-time, web-based siting model (*www.BioSAT.net*) is in alpha-stage development which will estimate economic supply curves for any given zip code location for 33 Eastern United States.

References

Abt, R. C., Cubbage, F. W., and Pacheco, G. (2000). Southern Forest Resource Assessment Using the Subregional Timber Supply (SRTS) Model. *Forest Products Journal*, 50(4).

Adler, P. R., Del Grosso, S. J., and Parton, W. J. (2007). Life-cycle assessment of net greenhouse-gas flux for bioenergy cropping systems. *Ecological Applications.* 17(3), 675-691.

Agresti, A. (2002). *Categorical data analysis*. John Wiley & Sons, Inc., New York, NY.

Aldenderfer, M. S., and Blashfield, R. K. (1984). *Cluster Analysis*. Sage Publications Inc., Newbury Park, CA.

Allen, R. M., and Bennetto, H. P. (1993). Microbial fuel-cells: electricity production from carbohydrates. *Applied Biochemistry and Biotechnology.* 39-40, 27-40.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*(4): 589-609.

Altman, E. I., Marco, G., and Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking and Finance, 18*(3): 505-529.

Angenent, L. T., Karimk, K., Al-Dahhan, M. H., Wrenn, B. A., Domíguez-Espinosa, R. (2004). Production of bioenergy and biochemicals from industrial and agricultural wastewater. *Trends in Biotechnol.* 22(9), 477-485.

Arellano, A. F., Kasibhatla, P. S., Giglio, L., van der Werf, G. R., Randerson, J. T., and Collatz, G. J. (2006). Time-dependent inversion estimates of global biomass-burning CO emissions

using Measurement of Pollution in the Troposphere (MOPITT) measurements. *Journal of Geophysical Research-Atmospheres, 111*(D9).

Balakrishnama, S., and Ganapathiraju, A. (1998). Linear discriminant analysis - A brief tutorial. *Institute for Signal and Information Processing*. MI.

Balat, M. (2005). Current alternative engine fuels. *Energy Sources,* 27, 569-577.

Bender, M. (1999). Economic feasibility review for community-scale farmer cooperatives for biodiesel. *Bioresource Technology,* 70, 81-87.

Berwick, M., and Farooq, M. (2003). Truck costing model for transportation managers. Upper Great Plains Transportation Institute, North Dakota State University, from http://www.mountain-plains.org/pubs/pdf/MPC03-152.pdf

Bies, L. (2006). The biofuels explosion: Is green energy good for wildlife? *Wildlife Society Bulletin, 34*(4), 1203-1205.

Bothast, R. J. (2005). New technologies in biofuel production. *Agricultural Outlook Forum (AOF)*.

Bouveyron, C. (2007). High-dimensional discriminant analysis. *Communications in Statistics-Theory and Methods, 36*(14): 2607-2623.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 36*(3): 317-346.

Braga-Neto, U., Hashimoto, R., Dougherty, E. R., Nguyen, D. V., and Carroll, R. J. (2004). Is cross-validation better than resubstitution for ranking genes? *Bioinformatics,* 20: 253-258.

Brieman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth Inc., Belmont, CA. 67 p.

Brothier, M., Gramondi, P., Poletiko, C., Michon, U., Labrot, M., and Hacala, A. (2007). Biofuel and hydrogen production from biomass gasification by use of thermal plasma. *High Temperature Material Processes, 11*(2), 231-243.

Brumbley, S. M., Purnell, M. P., Petrasovits, L. A., Nielsen, L. K., and Twine, P. H. (2007). Developing the sugarcane biofactory for high-value biomaterials. *International Sugar Journal, 109*(1297), 5.

Bullen R. A, Arnot, T. C., Lakeman, J. B., and Walsh, F. C. (2006). Biofuel cells and their development. *Biosensors and Bioelectronics, 19*, 607-613.

Buntine, W. (1992). Learning classification trees. *Statistics and Computing, 2*(2): 63-73.

Byrne, J., Shen, B., and Li, X. (1996). The Challenge of Sustainability: Balancing China's Energy, Economic and Environmental Goals. *Energy Policy, 24*(5), 455-462.

Cadenas, A. and Cabezudo, S. (1998). Biofuels as Sustainable Technologies: Perspectives for Less Developed Countries-Food versus Fuel? *Technological Forecasting and Social Change, 58*(1), 83-103.

Canakci, M., and Van Gerpen, J. (2001). Biodiesel production from oils and fats with high free fatty acids. *Trans. ASAE, 44*(6), 1429-1436.

Charles, M. B., Ryan, R., Rtan, N., and Oloruntoba, R. (2008) Public policy and biofuels: The way forward? *Energy Policy*, 26(1), 495.

Chaudhuri, P., Huang, M. C., Loh, W. Y., and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica, 4*(1): 143-167.

Chen, F., and Dixon, R. A. (2007). Lignin modification improves fermentable sugar yields for biofuel production. *Nature Biotechnology, 25*(7), 759-761.

Cherkassky, V. S., and Mulier, F. (1998). *Learning from data: Concepts, theory, and methods*. John Wiley & Sons, Inc., New York, NY.

Chou, P. A., Lookabaugh, T., and Gray, R. M. (1989). Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory, 35*(2): 299-315.

Das, P., Sreelatha, T., and Ganesh, A. (2004). Bio oil from pyrolysis of cashew nut shell-characterisation and related properties. *Biomass & Bioenergy, 27*(3), 265-275.

Davis F., and Higson, S. P. (2007). Biofuel cells--recent advances and applications. *Biosensors and Bioelectronics, 22*.

Demirbas, A. (2000). Biomass resources for energy and chemical industry. *Energy Edu. Sci. Technol, 5*(1), 21-45.

Demirbas, A. (2002). Biodiesel from vegetable oils via transesterification in supercritical methanol. *Energy Conversion and Management, 43*(17), 2349-2356.

Demirbas, A. (2003). Current advances in alternative motor fuels. *Energy Explor Exploit 21*, 475-487.

Demirbas, A. (2004). Bioenergy, global warming, and environmental impacts. *Energy Sources 26*, 225-236.

Demirbas, A. (2006). Biofuel based cogenerative energy conversion systems. *Energy Sources Part a-Recovery Utilization and Environmental Effects, 28*(16), 1509-1518.

Demirbas, A. (2007). Progress and recent trends in biofuels. *Progress in Energy and Combustion Science, 33*(1), 1-18.

Demirbas, M. F., and Balat, M. (2005). Recent advances on the production and utilization trends of bio-fuels: A global perspective. *Energy Conversion and Management,* 47(15-16), 2371-2381.

Dietter, J., Morgner, H., and Difiglio, C. (1997). Using advanced technologies to reduce motor vehicle greenhouse gas emissions. *Energy Policy, 25*(14), 1173-1178.

Draper, N. R., and Smith, H. (1981). *Applied regression analysis*. John Wiley & Sons, Inc., New York, NY.

Du, Z.W., Li, H.R., and Gu, T.Y. (2007). A state of the art review on microbial fuel cells: A promising technology for wastewater treatment and bioenergy. *Biotechnology Advances, 25*(5), 464-482.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*(457): 77-88.

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *Journal of Finance, 32*(3): 875-900.

Energy Information Administration. (2006). International Energy Outlook 2006. U.S. Department of Energy Office of Integrated Analysis and Forecasting Publication DOE/EIA-0484 (2006), Washington, DC. 192 p.

Eriksson, E., and Johansson, T. (2006). Effects of rotation period on biomass production and atmospheric CO2 emissions from broadleaved stands growing on abandoned farmland. *Silva Fennica, 40*(4), 603-613.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. *Advances in knowledge discovery and data mining table of contents*, 1-34.

Fielding, A. (1977). Binary segmentation: The automatic interaction detector and related techniques for exploring data structure. *The Analysis of Survey Data, I*(Exploring Data Structures), 221-257.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2): 179-188.

Foyle, T., Jennings, L., and Mulcahy, P. (2007). Compositional analysis of lignocellulosic materials: Evaluation of methods used for sugar analysis of waste paper and straw. *Bioresource Technology, 98*(16), 3026-3036.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5): 1189-1232.

Frydman, H., Altman, E. I., and Kao, D. L. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *Journal of Finance, 40*(1): 269-291.

Fulton, L., Howes, T., and Hardy, J. (2004). *Biofuels for Transport: An International Perspective*: OECD, International Energy Agency.

Gan, J. B., and Smith, C. T. (2006). A comparative analysis of woody biomass and coal for electricity generation under various CO2 emission reductions and taxes. *Biomass & Bioenergy, 30*(4), 296-303.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika, 61*(1): 101-107.

Geisser, S., and Greenhouse, S. W. (1958). An extension of Box's M results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics, 29*, 885-891.

Gelfand, S. B., Ravishankar, C. S., and Delp, E. J. (1991). An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(2): 163-174.

Gercel, H. F. (2002). The production and evaluation of bio-oils from the pyrolysis of sunflower-oil cake. *Biomass & Bioenergy, 23*(4), 307-314.

Gigler, J. K., Meerdink, G., and Hendrix, E.M.T. (1999). Willow supply strategies to energy plants. *Biomass & Bioenergy, 17*(3), 185-198.

Goldemberg, J. (2000). *World Energy Assessment: Energy and the Challenge of Sustainability*: United Nations Development Programme.

Gommers, A., Thiry, Y., Vandenhove, H., Vandecasteele, C. M., Smolders, E., and Merckx, R. (2000). Radiocesium uptake by one-year-old willows planted as short rotation coppice. *Journal of Environmental Quality, 29*(5), 1384-1390.

Gorski, W. (2006). Biofuels in the ISO, EN and Polish (PN) standards. *Przemysl Chemiczny, 85*(12), 1632-1640.

Graham, R. L., English, B. C., and Noon, C. E. (2000). A Geographic Information System-based modeling system for evaluating the cost of delivered energy crop feedstock. *Biomass & Bioenergy, 18*(4), 309-329.

Graham, R. L., Liu, W., Downing, M., Noon, C. E., Daly, M., and Moore, A. (1997). The effect of location and facility demand on the marginal cost of delivered wood chips from energy crops: A case study of the state of Tennessee. *Biomass & Bioenergy, 13*(3), 117-123.

Graham, R. L., Liu, W., Jager, H. I., English, B. C., Noon, C. E ., and Daly, M. J. (1996). *A regional-scale GIS-based modeling system for evaluating the potential costs and supplies of biomass from biomass crops*. From Proceeding, Bioenergy 96 -- The 7th national bioenergy conference. Nashville, TN.

Granda, C. B., Zhu, L., and Holtzapple, M. T. (2007). Sustainable liquid biofuels and their environmental impact. *Environmental Progress, 26*(3), 233-250.

Gray, K. A. (2007). Cellulosic ethanol - state of the technology. *International Sugar Journal, 109*(1299), 145.

Hall, D. O., and Scrase, J. I. (1998). Will biomass be the environmentally friendly fuel of the future? *Biomass & Bioenergy, 15*(4-5), 357-367.

Hamelinck, C. N., and Faaij, A.P.C. (2006a). Outlook for advanced biofuels. *Energy Policy.* 34(17), 3268-3283.

Hamelinck, C. N., and Faaij, A.P.C. (2006b). Production of advanced biofuels. *International Sugar Journal.* 108(1287), 168-175.

Hill, J., Nelson, E., Tilman, D., Polasky, S., and Tiffany, D. (2006). Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of Sciences of the United States of America.* 103(30), 11206-11210.

Hillring, B. (2006). World trade in forest products and wood fuel. *Biomass & Bioenergy, 30*(10), 815-825.

Hodges, D. G., Young, T. M., and Abt, R. C. (2007) *Regional comparative advantage for woody biofuels production*. Univ. of Tennessee. Department of Forestry, Wildlife and Fisheries. Unpublished proposal manuscript for Southeastern Sun Grant Center.

Hoogwijk, M., Faaij, A., Eickhout, B., de Vries, B., and Turkenburg, W. (2005). Potential of biomass energy out to 2100, for four IPCCSRES land-use scenarios. *Biomass & Bioenergy, 29*(4), 225-257.

Hunt, S. C. (2006). *Biofuels for Transport: Global Potential and Implications for Sustainable Agriculture and Energy in the 21st Century.* Worldwatch Institute. No. 978-1-84407-422-8. Washington, DC.

Husain, S. A., Rose, D. W., and Archibald, S. O. (1998). Identifying agricultural sites for biomass energy production in Minnesota. *Biomass & Bioenergy, 15*(6), 423-435.

Ieropoulos I. A., Greenman, J., Melhuish, C., and Hart, J. (2005). Comparative study of three types of microbial fuel cell. *Enzyme and Microbial Technology 37*(2), 238-245.

Jain, A. K., Tao, Z. N., Yang, X. J., and Gillespie, C. (2006). Estimates of global biomass burning emissions for reactive greenhouse gases (CO, NMHCs, and NOx) and CO2. *Journal of Geophysical Research-Atmospheres, 111*(D6).

Jefferson, M. (2006). Sustainable energy development: performance and prospects. *Renewable Energy, 31*(5), 571-582.

Jensen, K., Menard, J., English, B., Park, W., and Wilson, B. (2002). The wood transportation and resource analysis system (WTRANS): an analysis tool to assist wood residue producers and users. *Forest Products Journal* 52(5): 27-33.

Jorapur, R., and Rajvanshi, A. K. (1997). Sugarcane leaf-bagasse gasifiers for industrial heating applications. *Biomass & Bioenergy, 13*(3), 141-146.

Kaltschmitt, M., and Weber, M. (2006). Markets for solid biofuels within the EU-15. *Biomass & Bioenergy, 30*(11), 897-907.

Kass, G. V. (1975). Significance testing in automatic interaction detection (AID). *Applied Statistics, 24*(2): 178-189.

Kim, H., Guess, F. M., and Young, T. M. (2007). Using data mining tools of decision trees in reliability applications. *IIE Transactions. In Press.*

Kim, H., and Loh, W. Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association,* 96(454).

Kim, H., and Loh, W. Y. (2003). Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics, 12*(3): 512-530.

Klecka, W. R. (1980). *Discriminant analysis*. Sage Publications. Beverly Hills, CA.

Koehler, G. J., and Erenguc, S. S. (1990). Minimizing misclassifications in linear discriminant analysis. *Decision Sciences, 21*(1): 63-85.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 2*(12): 1137-1143.

Kuhn, R., and De Mori, R. (1995). The application of semantic classification trees to natural language understanding. *IEEE Trans. Pattern Anal. Machine Intell.*, 17: 449-460.

Lachenbruch, P. A., and Goldstein, M. (1979). Discriminant analysis. *Biometrics, 35*(1): 69-75.

Lal, R. (2006). Soil and environmental implications of using crop residues as biofuel feedstock. *International Sugar Journal, 108*(1287), 161-167.

Langholtz, M., Carter, R. D., Marsik, M., and Schroeder, R. (2006). Measuring the Economics of Biofuel Availability. from *http://www.esri.com/news/arcuser/1006/biomass1of2.html*

Liu, X., Wang, Y., Young, T. M., Rials, T. G., Hodges, D. G., Hartsell, A., and Guess, F. M. (2008). Optimal biorefinery sites in TX and LA based on transportation costs. *The presentation on the 62nd Forest Product Society Conference*. St. Louis, MI.

Loh, W. Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica, 12*(2): 361-386.

Loh, W. Y., and Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica, 7*(4): 815-840.

Longmire, C., Hodges, D. G., Ostermeier, D. M., and Fly., M. (2007). *Landowner attitudes regarding woodland management on the Northern Cumberland Plateau.* Univ. of Tennessee. Department of Forestry, Wildlife and Fisheries. Unpublished manuscript.

Lynd, L. R. (1996). Overview and evaluation of fuel ethanol from cellulosic biomass: Technology, economics, the environment, and policy. *Annual Review of Energy and the Environment, 21*, 403-465.

Ma, F., and Hanna, M. A. (1999). Biodiesel production: a review. *Bioresource Technology, 70*(1), 1-15.

Matthews, R. W. (2001). Modelling of energy and carbon budgets of wood fuel coppice systems. *Biomass & Bioenergy, 21*(1), 1-19.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, Inc., New York, NY.

McLaughlin, S. B., and Kszos, L. A. (2005). Development of switchgrass (Panicum virgatum) as a bioenergy feedstock in the United States. *Biomass & Bioenergy, 28*(6), 515-535.

Menard, S. W. (2002). *Applied logistic regression analysis*. Sage Publications Inc., Thousand Oaks, CA.

Microsoft Inc. (2006) MapPoint for Windows. Seattle, WA.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K. R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, 41-48.

Milbrandt, A. (2005). A Geographic Perspective on the Current Biomass Resource Availability in the United States., from *http://www.osti.gov/energycitations/product.biblio.jsp?osti_id=861485*

Moller, B., and Nielsen, P.S. (2007). Analysing transport costs of Danish forest wood chip resources by means of continuous cost surfaces. *Biomass & Bioenergy, 31*(5), 291-298.

Morgan, J. N., and Sonquist, J. A. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association, 58*(415-435): 35.

Murray, L. D., Best, L. B., Jacobsen, T. J., and Braster, M. L. (2003). Potential effects on grassland birds of converting marginal cropland to switchgrass biomass production. *Biomass & Bioenergy, 25*(2), 167-175.

Nakagawa, H., Harada, T., Ichinose, T., Takeno, K., Matsumoto, S., Kobayashi, M. (2007). Biomethanol production and CO2 emission reduction from forage grasses, trees, and crop residues. *Jarq-Japan Agricultural Research Quarterly, 41*(2), 173-180.

Nemoto, J., Horikawa, M., Ohnuki, K., Shibata, T., Ueno, H., Hoshino, M. (2007). Biophotofuel cell (BPFC) generating electrical power directly from aqueous solutions of biomass and its related compounds while photodecomposing and cleaning. *Journal of Applied Electrochemistry, 37*(9), 1039-1046.

Nilsson, D. (1999). SHAM - a simulation model for designing straw fuel delivery systems. Part 1: model description. *Biomass & Bioenergy, 16*(1), 25-38.

Noon, C. E., and Daly, M. J. (1996). GIS-based biomass resource assessment with BRAVO. *Biomass and Bioenergy, 10*(2-3), 101-109.

Noon, C. E., Daly, M. J., Graham, R. L., and Zahn, F. B. (1996). *Transportation and site location analysis for regional integrated biomass assessment (RIBA)*. Proceeding for the 7th national bioenergy conference, Nashville, TN.

Ozcimen, D., and Karaosmanoglu, F. (2004). Production and characterization of bio-oil and biochar from rapeseed cake. *Renewable Energy, 29*(5), 779-787.

Parikka, M. (2004). Global biomass fuel resources. *Biomass & Bioenergy, 27*(6), 613-620.

Perlack, R. D., Wright, L. L., Turhollow, A. F., Graham, R. L., Stokes, B. J., and Erbach, D. C. (2005). Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply. Publication DOE/GO-102995-2135/ORNL TM-2005/66. Oak Ridge National Laboratory, Oak Ridge, TN. p 60.

Pintrich, P. R., and De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1): 33-40.

Polagye, B. L., Hodgson, K. T., and Malte, P. C. (2007). An economic analysis of bio-energy options using thinnings from overstocked forests. *Biomass & Bioenergy, 31*(2-3), 105-125.

Poole, D. J., Sharifi, V., Swithenbank, J., Kilgallon, P., Simms, N., Oakey, J. (2007). Continuous analysis of elemental emissions from a biofuel gasifier. *Journal of Analytical Atomic Spectrometry, 22*(5), 532-539.

Poon, T. C. W., Chan, A. T. C., Zee, B., Ho, S. K. W., Mok, T. S. K., Leung, T. W. T. (2001). Application of classification tree and neural network algorithms to the identification of serological liver marker profiles for the diagnosis of hepatocellular carcinoma. *Oncology, 61*(4): 275-283.

Pordesimo, L. O., Hames, B. R., Sokhansanj, S., and Edens, W. C. (2005). Variation in corn stover composition and energy content with crop maturity. *Biomass & Bioenergy, 28*(4), 366-374.

Press, S. J., and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*(364): 699-705.

Ptasinski, K. J., Prins, M. J., and Pierik, A. (2007). Exergetic evaluation of biomass gasification. *Energy, 32*(4), 568-574.

Puhan, S., Vedaraman, N., Rambrahamam, B. V., and Nagarajan, G. (2005). Mahua (madhuca indica) seed oil: A source of renewable energy in India. *Journal of Scientific & Industrial Research, 64*(11), 890-896.

Puppan, D. (2002). Environmental evaluation of biofuels. *Periodica Polytechnica Ser. Soc. Man. Sci. 10*(1), 95-116.

Quinlan, J. R., and Rivest, R. L. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation, 80*(3): 227-248.

Reijnders, L. (2006). Conditions for the sustainability of biomass based fuel use. *Energy Policy, 34*, 863-876.

Rencher, A. C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician, 46*(3): 217-225.

Rencher, A. C. (1993). The contribution of individual variables to Hotelling T2, Wilks' lambda, and R2. *Biometrics, 49*(2): 479-489.

Schapire, R. E. (2002). The boosting approach to machine learning: An overview. *MSRI Workshop on Nonlinear Estimation and Classification*.

Scurlock, J.M.O., Dayton, D. C., and Hames, B. (2000). Bamboo: an overlooked biomass resource? *Biomass & Bioenergy, 19*(4), 229-244.

Searcy, E., Flynn, P., Ghafoori, E., and Kumar, A. (2007). The relative cost of biomass energy transport. *Applied Biochemistry and Biotechnology, 137*, 639-652.

Sheehan, J., Camobreco, V., Duffield, J., Shapouri, H., Graboski, M., and Tyson, K. S. (2000). *An Overview of Biodiesel and Petroleum Diesel Life Cycles*: NREL/TP-580-24772, National Renewable Energy Lab., Golden, CO (US)

Sims, R.E.H., Hastings, A., Schlamadinger, B., Taylor, G., and Smith, P. (2006). Energy crops: current status and future prospects. *Global Change Biology, 12*(11), 2054-2076.

Siotani, M., Hayakawa, T., and Fujikoshi, Y. (1985). *Modern multivariate statistical analysis: A graduate course and handbook*. American Sciences Press, Columbus, OH.

Soccol, C. R., Vandenberghe, L.P.S., Costa, B., Woiciechowski, A. L., de Carvalho, J. C., Medeiros, A.B.P. (2005). Brazilian biofuel program: An overview. *Journal of Scientific & Industrial Research, 64*(11), 897-904.

Solomon, B. D., Barnes, J. R., and Halvorsen, K. E. (2007). Grain and cellulosic ethanol: History, economics, and energy policy. *Biomass & Bioenergy, 31*(6), 416-425.

Sperling, D. (1984). An Analytical Framework for Siting and Sizing Biomass Fuel Plants. *Energy, Vol. 9*(No. 11-12), 1033-1040.

SPSS Inc. (2007) SPSS 16.0 for Windows. Release 16.0.1 Chicago, IL.

Stephanopoulos, G. (2007). Challenges in engineering microbes for biofuels production. *Science, 315*(5813), 801-804.

Strobl, C., Boulesteix, A. L., and Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics and Data Analysis, 52*(1): 483-501.

Taleghani, G., and Kia, A. S. (2005). Technical–economical analysis of the Saveh biogas power plant. *Renew Energy 30*, 441-446.

Timber Mart South (2008). Timber mart-south market news quarterly. The Journal of Southern Timber Market News. 13(1):29.

Torney, F., Moeller, L., Scarpa, A., and Wang, K. (2007). Genetic engineering approaches to improve bioethanol production from maize. *Current Opinion in Biotechnology, 18*(3), 193-199.

Ture, S., Uzun, D., and Ture, I. E. (1997). The potential use of sweet sorghum as a non-polluting source of energy. *Energy, 22*(1), 17-19.

USDA Forest Service Research and Development (2006). Wildland fire and fuels research and development strategic plan: Meeting the needs of the present, anticipating the needs of the future. FS-854

USDA Forest Service, SRS, TPO data, (2003-2005). Environmental Systems Research Institute Inc. (ESRI)

Verbyla, D. L., and Litvaitis, J. A. (1989). Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management, 13*(6): 783-787.

Walsh, M. E. (1998). US bioenergy crop economic analyses: Status and needs. *Biomass & Bioenergy.* 14(4), 341-350.

Weber, J. A. (1993). *The economic feasibility of community-based biodiesel plants.* Master Thesis. University of Missouri, Columbia, 108.

Wierzbicka, A., Lillieblad, L., Pagels, J., Strand, M., Gudmundsson, A., Gharibi, A. (2005). Particle emissions from district heating units operating on three commonly used biofuels. *Atmospheric Environment.* 39, 139-150.

Wilkinson, L. (1992). Tree structured data analysis: AID, CHAID and CART. *Sawtooth/SYSTAT Joint Software Conference*, Sun Valley, ID.

Wu, C. (2007). Cellulose dreams: the search for new means and materials for making ethanol. *Science News.* 172(8), 120-121.

Wu, M., Wu, Y., and Wang, M. (2006). Energy and emission benefits of alternative transportation liquid fuels derived from switchgrass: A fuel life cycle assessment. *Biotechnology Progress.* 22(4), 1012-1024.

Wyman, C. E. (1996). *Handbook on Bioethanol: Production and Utilization*. Taylor & Francis. Washington DC.

Young, T. M. (2007). *Parametric and non-parametric regression tree models of the strength properties of engineered wood panels using real-time industrial data.* Ph.D. Dissertation, Univ. of Tennessee, Knoxville, TN.

Young, T. M., Ostermeier, D. M., Daniel Thomas, J., and Brooks, R.T. (1991). The economic availability of woody biomass for the Southeastern United States. *Bioresource Technology*. 37(1), 7-15.

Zhang, Y., Dub, M. A., McLean, D. D., and Kates, M. (2003). Biodiesel production from waste cooking oil: 1. Process design and technological assessment. *Bioresource Technology.* 89(1), 1-16.

Zhao, W., Chellappa, R., and Phillips, P. J. (1999). *Subspace linear discriminant analysis for face recognition.* Center for Automation Research, Univ. of Maryland, College Park, Technical Report CAR-TR-914, 137-144.

Vita

Yingjin Wang, originally from Wuhan, China, earned her Bachelor and Master degree in Engineering from Huazhong University of Science and Technology respectively in June 2001 and June 2004. She is currently pursuing a second Master degree in Statistics from University of Tennessee and plans to graduate in August, 2008. She has served as a Graduate Teaching Assistant in Department of Statistics, Operations, and Management Science and Graduate Research Assistant in Department of Forestry, Wildlife and Fisheries, Forest Product Center.