12-2009

A Statistical Analysis of Key Factors Influencing the Location of Biomass-using Facilities

Xu Liu
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Xu Liu entitled "A Statistical Analysis of Key Factors Influencing the Location of Biomass-using Facilities." 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 L. 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 Xu Liu entitled “A statistical analysis of key factors influencing the location of biomass-using facilities.” 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 L. 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.)

A statistical analysis of key factors influencing the location of biomass-using facilities

A Thesis
Presented for the
Master of Science Degree
The University of Tennessee, Knoxville

Xu Liu
December 2009

Dedication

This thesis is dedicated to my parents, my parents-in-law, and my husband for their tremendous help, and to my beloved son Ryan L. Yao.

Acknowledgements

Of all those who helped me to write this thesis, I am most indebted to Dr. Timothy M. Young, Dr. Frank M. Guess, and Dr. Russell L. Zaretzki who provided extensive support and understanding on my writing and on my personal life.

Great thanks go to Mr. Jim Perdue (U.S. Forest Service), for his professional identification of the biomass definition and genius development of the website www.BioSAT.net. The data that I used in this thesis are an expansion of the website's database. Thanks to Ms. Sachiko Hurst (UT Office of Bioenergy Programs) whose excellent programming ability is a great help with data extraction from SQL server. She is such a smart and nice lady capable of handling complicated and large databases.

I thank U.S. Forest Service employee and graduate student Mr. Andrew Hartsell, Research Associate Ms. Kerri Norris, Research Associate Ms. Christy Pritchard, and Research Specialist Ms. Andrea Noehmer for their supports in every phase of my data collection and data management with their professional insight and unique technique for data organization. I appreciate Post Doctoral Research Associate Dr. Nicolas André for his help in creating the Visual Basic program for my database generation and for his detailed support in every aspect of computation. I thank Graduate Research Assistant Ms. Xia Huang for her strong ability in GIS identification of ZCTA for each water port's location. I appreciate Mr. Cody Steele, Statistical Communications Specialist at Minitab, Inc., for his professional suggestions to my thesis revision. I thank Graduate Research Assistants Mr. Yan Zeng and Mr. Dillon Carty for their tremendous help in thesis revision and for their continuous encouragement. I thank Ms. Amanda

Silk, the nicest Administrative Assistant, and Mr. Bob Longmire (UT Office of Bioneregy Programs), the most fun person, in the Forest Products Center for their help and kindness.

Abstract

Bioenergy and biofuels are emerging industries in the U.S. economy that will require statistical and economical analyses of woody biomass resources, supply chains, and other key factors that influence the siting of industrial facilities. This thesis develops models using logistic regression to improve the understanding of the key factors that influence the locations of existing wood-using bioenergy and biofuels plants, and other wood-using plants. The scope of the study is 13 Southeastern states.[1] Logistic regression models are developed at the state and regional levels. The resolution of the study is the ZIP Code tabulation area (ZCTA). There are 9,416 ZCTAs in the 13–state study region.

Because a small number of woody biomass-using bioenergy and biofuels plants exist relative to the large number of traditional woody biomass-using facilities (e.g., wood composites, sawmills, and secondary mills), two sample groups are developed. The first group combines all wood-using mills with wood-using bioenergy and biofuels plants, and compares ZCTAs with these types of mills with ZCTAs that do not contain any such facilities. This follows a more modern planning view of total woody biomass management. The second group combines only one type of mill, pulp and paper mills, with wood-using bioenergy and biofuels plants, and compares ZCTAs of these mill types with ZCTAs that do not contain such facilities.

For both groups in the entire study region, logging residues harvesting costs (negative influence) and the availability of thinnings within an 80-mile haul distance (positive influence) are statistically significant factors (p-value < 0.0001) in the logistic models. Population is

[1] Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia.

statistically significant and has a negative influence on site location for six of the thirteen states in the region (p-values ranged from < 0.0001 to 0.0197) for the first group. Twenty-five optimal locations in the Southeastern states (ZCTAs) are predicted from the logistic regression models. A de-clustering algorithm is developed as part of this study to avoid locating potential bioenergy and biofuels sites in close proximity to competing mills within same ZCTA.

Table of Contents

Chapter **Page**

Chapter 1 Introduction **1**

Chapter 2 Literature Review **5**

2.1 Bioenergy 5

2.1.1 Introduction of Bioenergy 5

2.1.2 Importance of Bioenergy 6

2.2 Woody Biomass 7

2.2.1 Concept of Woody Biomass 7

2.2.2 Importance of Using Woody Biomass 8

2.2.3 Feasibility of Woody Biomass 10

2.3 Logistic Model 11

2.3.1 General Introduction of Logistic Model 11

2.3.2 The Origin of Logistic Model 12

2.3.3 Logistic Model Application History 13

2.4 Biomass-using Facilities Siting Models 16

Chapter 3 Methods **18**

3.1 Variables Explanation 18

3.1.1 Response Variables Design 18

3.1.1.1 Coded “1” –existing mill locations 18

3.1.1.2 Coded “0” –“non-probable” locations 19

3.1.2 Explanatory Variables 22

3.2. Data Management and Data Quality 25

3.2.1 Data Management Tools and Database Structures 25

3.2.2 Data Resources and Data Collection Levels 26

3.2.3 Data Accuracy and Consistency 27

3.2.4 Data Management and Missing Value Surrogate Methods 27

3.2.5 Algorithms for Data Generation 29

3.2.5.1 Neighboring ZIP Code Algorithm 29

3.2.5.2 Trucking Cost Generation Algorithm 31
3.3 Logistic Models 32
3.3.1 Logistic Models' Introduction 32
3.3.2 Model Selection Methods and Criteria 33
3.3.3 Model Assessment Tools 34
3.3.3.1 Lift Charts 34
3.3.3.2 Classification Tables 34
3.3.4 Four Optional Models 35

Chapter 4 Results and Discussion 38

4.1 Logistic Regression Results for Group I Biomass-using Facilities 38
4.1.1 Models Assessment for Group I Biomass-using Facilities 38
4.1.2 Predictive Ability Measured by Classification Tables 39
4.1.3 Regional Level Analysis Result for Group I Biomass-using Facilities 41
4.1.4 State Level Analyses for Group I Biomass-using Facilities 45
4.2 Logistic Regression Results for Group II Biomass-using Facilities 47
4.2.1 Models Assessment for Group II Biomass-using Facilities 47
4.2.2 Predictive Ability Measured by Classification Tables 49
4.2.3 Regional Level Analysis Result for Group II Biomass-using Facilities 50
4.2.4 State Level Analyses for the Group II Biomass-using Facilities 53
4.3 De-clustering Algorithm Application to Prediction Results 53

Chapter 5 Summary 60

Chapter 6 Future Research 62

References 64

Appendices 74

A-1 MATLAB codes for generating the neighboring ZIP Code list 75
A-2 SAS codes for data collection and data management of the responses and explanatory variables 78
A-3 SAS codes for de-clustering algorithms 97

Vita 106

List of Tables

Table **Page**

Table 2-1 Numbers of articles in statistical journals containing the word "logit". 15

Table 3-1 Three variables for specifying “non-probable” locations of all wood-using mills with bioenergy and biofuels plants..... 22

Table 3-2 Explanatory variables for two groups of biomass-using facilities. 23

Table 3-3 A ZIP Code pair and its driving distance and driving time. 25

Table 3-4 ZIP Code level variables combination values and surrogating number for missing values. 28

Table 4-1 Model assessment results by BIC criterion for Group I biomass-using facilities. 38

Table 4-2 Classification Table of predictive ability measurement for Group I biomass-using facilities..... 40

Table 4-3 Type 3 analysis of effects for Group I biomass-using facilities. 42

Table 4-4 Analysis of maximum likelihood estimates of parameters for Group I biomass-using facilities..... 43

Table 4-5 State level analysis results for Group I biomass-using facilities. 46

Table 4-6 Model assessment results by BIC criterion for Group II biomass-using facilities. 47

Table 4-7 Classification Table of predictive ability measurement for Group II biomass-using facilities..... 49

Table 4-8 Analysis of maximum likelihood estimates of parameters for Group II biomass-using facilities..... 51

Table 4-9 State level analysis results for Group II biomass-using facilities..... 54

List of Figures

Figure **Page**

Figure 2-1 Standard logistic regression curve (Gershenfeld 1999). 12

Figure 3-1 All wood-using mills with wood-using bioenergy and biofuels plants in 13 Southeastern states. 20

Figure 3-2 Pulp and paper mills with wood-using bioenergy and biofuels plants in 13 Southeastern states. 21

Figure 3-3 Four optional models in SAS® Enterprise Miner. 36

Figure 4-1 Cumulative lift charts that assess four optional models for Group I biomass-using facilities. 39

Figure 4-2 Non-cumulative lift charts that assess four optional models for Group I biomass-using facilities. 40

Figure 4-3 Predictive ability plot based on the classification table for Group I biomass-using facilities. 41

Figure 4-4 Top 25 optimal locations for Group I biomass-using facilities at the 13-state regional level. 44

Figure 4-5 Cumulative lift charts that assess four optional models for Group II biomass-using facilities. 48

Figure 4-6 Non-cumulative lift charts that assess four optional models for Group II biomass-using facilities. 48

Figure 4-7 Predictive ability plot based on the classification table for Group II biomass-using facilities. 50

Figure 4-8 Top 25 optimal locations for Group II biomass-using facilities at the 13-state regional level. 52

Figure 4-9 Top 25 optimal locations after de-clustering for Group I biomass-using facilities. 57

Figure 4-10 Top 25 optimal locations after de-clustering for Group II biomass-using facilities. 58

Chapter 1 Introduction

According to Energy Information Administration (EIA) reports, 84% of U.S. primary energy consumption in 2008 was from fossil fuels. In addition, "in 2008, net imported energy accounted for 26 percent of all energy consumed" in the U.S. (U.S. Department of Energy 2009). Non-renewable fossil fuels exist in complex geo-political regions of the world and should be considered an energy resource with limitations. As oil demands from China and India increase in the future, it is important for the U.S. to have a renewed research emphasis on renewable and long-term sustainable sources of energy (e.g., biomass, solar, etc.). Biomass is considered an environmentally friendly, renewable, and abundant resource from which various useful chemicals and fuels can be produced. In general, the definition of biomass is:

1) "The cell mass produced by a population of living organisms;
2) The organic matter that can be used either as a source of energy or for its chemical components;
3) All the organic matter that derives from the photosynthetic conversion of solar energy" (Government of Canada BioPortal Glossary 2009)

This research uses the second and third definitions of biomass. The goal of the research is to study the factors that influence the location of existing biomass-using facilities in the Southeastern United States. The research has four objectives:

1) Develop an expanded database from the BioSAT (www.BioSAT.net) database to include population, income, employment, water ports, railroad availability, etc.;

2) Develop logistic regression models to identify variables that influence the siting of wood-using bioenergy and biofuels facilities;
3) Predict optimal locations for two groups of biomass-using facilities based on the logistic regression models;
4) Develop a de-clustering algorithm to avoid optimal locations for two groups of biomass-using facilities to have competitive mills within 80 miles distance.

Biomass-using facilities for the last two definitions of biomass include five different types of mills. The five types are primary wood processing mills, secondary wood processing mills, pulp and paper mills, other mills, and wood-using bioenergy and biofuels plants. Given the limitation of the small number of existing wood-using bioenergy and biofuels plants relative to the large number of traditional wood-using facilities (e.g., primary wood processing mills, secondary wood processing mills, and pulp and paper mills), two study groups are developed given the limitations of the logistic regression method:

1) All wood-using mills combined with wood-using bioenergy and biofuels plants;
2) Pulp and paper mills combined with wood-using bioenergy and biofuels plants with primary and secondary wood-using mills as independent variables.

Explanatory variables are categorized into three groups for each ZCTA:

1) Economic variables:
    - Population
    - Family income
    - Employment
    - Population density

- Income per person
- Land area
- Water area

2) Biomass availability variables:
    - Logging residues
    - Other removals
    - Thinnings availability (within 40-mile, 80-mile, 120-mile, and 200-mile haul distances)
    - Urban waste
3) Transportation-related variables:
    - Marginal cost of truck hauling for each ZCTA for annual mill residues demand quantities of 0.5 million dry tons, 1 million dry tons, and 1.5 million dry tons
    - Average cost and total cost truck hauling for total mill residues for each ZCTA's 80-mile haul distance
    - Total quantity of total mill residues for each ZCTA's 80-mile haul distance
    - Logging residues harvesting cost
    - Water port availability
    - Railroad availability

Railroad availability is defined as an ordinal variable, which is ranked as N/A, 1, 2, 3, and 4. Other variables listed above are continuous variables. Logistic models are used to identify

statistically significant factors that influence the location of wood-using facilities for both a 13-state region and for each individual state.

Chapter 2 in the thesis is a literature review that briefly discusses the importance of bioenergy and the feasibility of using woody biomass for energy and biofuels. This chapter also provides a brief introduction to the development of the logistic regression model, its history, and its application in the sciences. Previous research related to the siting of biomass-using facilities is also discussed in this chapter.

Chapter 3 provides detailed explanations of data management methods and of the model selection and the model establishment processes in the thesis. Details and explanations of the database and relevant programming are discussed and referred to in the appendices of the thesis.

Chapter 4 contains the results and the discussion of the results. The predictive models are discussed in this chapter. Significant factors both at the 13-state regional level and at the state-level are explained in detail for the two groups of biomass-using facilities. The 25 optimal locations at the 13-state regional level are discussed with attention given to a de-clustering algorithm application. This algorithm adjusts the predicted possibility to avoid competitive bioenergy and biofuels plants siting closely.

Chapters 5 and 6 are concluding remarks and discussion of future research topics.

Chapter 2 Literature Review

2.1 Bioenergy

2.1.1 Introduction of Bioenergy

Given the economic limitations of fossil fuels in the presence of increasing global demand for energy, bioenergy offers a green-energy solution that is renewable, abundant, and environmentally friendly. According to the United States Department of Agriculture, "bioenergy is renewable energy derived from biological sources, to be used for heat, electricity, or vehicle fuel. Biofuel derived from plant materials is among the most rapidly growing renewable energy technologies" (U.S. Department of Agriculture Economic Research Service 2009). As an alternative to fossil fuels, biofuels are made from biomass resources or from the processing and conversion of derivatives of biomass resources. Biofuels include ethanol, biodiesel, and methanol (Perlack et al. 2005). According to Perlack et al. (2005), biomass includes "any organic matter that is available on a renewable or recurring basis, including agricultural crops and trees, wood and wood residues, plants (including aquatic plants), grasses, animal manure, municipal residues, and other residue materials. There are three main categories of biomass: primary, secondary, and tertiary." This thesis considers facilities for all types of biomass and focuses on the factors that influence site location, e.g., economic variables, biomass availability variables, and transportation-related variables. Biofuel industries are currently receiving more attention and expanding rapidly in Europe, Asia, and the United States (Byrne et al. 1996, Puhan et al. 2005, Soccol et al. 2005). Wright (2006) summarized the worldwide commercial development of bioenergy and the main sources of bioenergy. She found that "biomass electric

generation feedstocks are predominantly forest residues (including black liquor), bagasse, and other agricultural residues" in the U.S., European Union, and Brazil. The U.S. primarily uses starch from maize grain and oil seeds (soy or rapeseed) for biodiesel production.

2.1.2 Importance of Bioenergy

Bioenergy may allow for improved air-quality and enhanced forest management (Parhizkar and Smith 2008). Throughout the last century, worldwide energy consumption has increased 17-fold (United Nations Development Programme 2000). According to the Department of Energy's Annual Energy Review from 2008, the United States began to import energy in the late 1950s. The review also states that "in 2008, net imported energy accounted for 26 percent of all energy consumed" (U.S. Department of Energy 2009). The U.S. economy is dependent on transportation systems that use fossil fuels as a low cost source of transportation. Dependence on oil raises environmental concerns, but dependence on foreign oil adds national security concerns. The combination of environmental and security concerns leads to long-term economic questions. A plethora of literature exists on bioenergy and an extensive review was beyond the scope of this thesis. Biomass Research and Development Board (2008) provided an extensive summary of the economics of biomass feedstocks in the United States. Polagye et al. (2007) used thinnings as an example to analyze the economic impact of bioenergy options from an overstocked forest. A full economic analysis of switchgrass under different scenarios was developed by Kumar and Sokhansanj (2007). Milbrandt (2005) gave a geographic perspective on the availability of biomass resources in the United States. Brechbill and Tyner (2008) performed an economic analysis of corn stover and switchgrass. Summit Ridge Investments, LLC (2007) provided detailed biomass energy information, including hardwood woody biomass

energy opportunity in the Eastern U.S. Galik et al. (2009) analyzed three Southern states aggregate bioenergy potential and the potential supply cost of woody biomass, as well as the interaction between logging residues and roundwood supply. Abt et al. (2000) used the Subregional Timber Supply (SRTS) model to assess southern forest resources. Perez-Verdin et al. (2009) discussed the woody biomass availability for bioethanol conversion in Mississippi.

Retsina and Pylkkanen (2007) talked about emerging technologies that can be used to repurpose the traditional pulp mill into a biorefinery. Demirbas (2005) considered searching biomass residues for cellulosic material that can be used to make bioethanol. Western Governors' Association (2006) focused on using biomass for the production of electricity. In this report, the Western Governors' Association Biomass Task Force did not address the significant contributions that biomass can make in supplying fuels for the transportation sector. The Task Force determined that the Governors' Ethanol Coalition was a preferred venue for the development of policy recommendations related to biomass and transportation fuels. The U.S. Department of Energy (2006) discussed technologies in the forestry industry for energy and cost savings for all mills.

2.2 Woody Biomass

2.2.1 Concept of Woody Biomass

Generally, there are two main categories of biomass. One is forest-derived biomass and the other is agriculture-derived biomass. This thesis focuses on forest-derived biomass, which is defined as "woody biomass" for the remainder of the thesis. Four types of woody biomass are defined in this thesis:

1) Logging residues and other removals from the forest inventory;

2) Forest residues from fuel treatment thinning;

3) Forest products industry processing residues;

4) Urban wood residues.

Logging residues and other removals are categorized for hardwood and softwood, thinnings (within 40-mile, 80-mile, 120-mile, and 200-mile haul distances), total mill residues, and unused mill residues. Table 3-2 provides a detailed explanation of each woody biomass type.

2.2.2 Importance of Using Woody Biomass

Bartuska (2006), a deputy chief in the USDA Forest Service, emphasized the importance of using biomass for energy and discussed other uses and their impact on the environment, economy, and forest management. In the energy aspect, she claimed that only about 4% of our energy was from renewable sources and that energy primarily produced power and heat. About 10% of biomass energy was used for transportation fuels, primarily corn ethanol. The U.S. had a long history of using agricultural and forestry biomass, primarily forest and wood waste, to generate electricity, heat, and steam power. The U.S. could realistically displace 30% of current petroleum consumption with biomass, using an amount approximately equivalent to one billion dry tons of biomass (Bartuska 2006).

In addition to the environmental benefit of reducing greenhouse emissions, Bartuska (2006) summarized four benefits to the economy of using biomass for energy. First, removing excessive levels of forest biomass could reduce the risk of a fire. Second, managing biomass could improve forest productivity and improve forest health. Third, the use of biomass for energy created jobs in rural America while maintaining a forest-based infrastructure. Fourth,

dependence on imported fossil fuels is reduced, thus providing economic growth opportunities for the development of a more robust green economy. Economically, Bartuska (2006) brought forward a notion of "green economics." She pointed out that renewable energy and biobased products from biomass had "green" value because they produce goods and services while adding value to and protecting the environment.

According to the Biomass Research and Development Board (2008), a series of policies supported the development of biofuels. Those policies included the Biomass Research and Development Act of 2000, the Energy Policy Act of 2005 (which mandated increasing domestic use of renewable fuels to 7.5 billion gallons in 2012), the Energy Independence and Security Act (EISA) of 2007 (which established a 36-billion-gallon mandate for biofuels by 2022), and the 2002 and 2008 Farm Bills.

Sedjo (1997) and Forest2Market (2009) provided a general introduction to the current economic impact of woody biomass. Despite the advantages of woody biomass stated above, Caputo (2009) pointed out some disadvantages. He concluded that the potential for an increased demand for woody biomass to drive unsustainable levels of harvesting was dangerous. The negative consequences were damages of biodiversity, soil conservation, and water conservation. The cost of woody biomass associated with harvesting, transporting, storing, and utilizing the material often exceeded its value on the energy market. Some of this was due to the lower ticket price of fossil fuels, which did not include the negative social costs associated with climate change. The lower ticket price of fossil fuels also did not consider the potential for more cost-effective tools, equipment, and logistical processes currently under development for biomass.Caputo (2009) emphasized the fact that federal policies were required to ensure the

sustainability of woody biomass harvesting and to improve the economic feasibility of bioenergy projects.

2.2.3 Feasibility of Woody Biomass

Kaylen et al. (2002) built a mathematical model to analyze the economic feasibility of producing ethanol from lignocellulosic feedstocks. They found that recent technological advancements appeared to make ethanol competitive with gasoline, but only if higher valued chemicals were produced as co-products with the ethanol. The low cost and the chemical composition of crop residues made them attractive as a feedstock. Patton-Mallory (2008) focused on the idea that U.S. Forest Service programs could improve coordination with the use of woody biomass and forest management activities on both federal and private lands. This coordination would occur through improved partnerships, developing and applying new science and technology, expanding markets for bioenergy and biobased products, and facilitating a reliable and predictable supply of biomass. Scurlock (2001) performed detailed research on bioenergy feedstock characteristics by comparing the typical properties of bioenergy feedstocks and biofuels with coal and oil in physical characteristics, chemical characteristics, and composition. He concluded that biomass materials were easier to gasify than coal and that heating values and moisture content of biomass materials were more uniform. However, the bulk density of most biomass feedstocks was generally lower, even after densification.

2.3 Logistic Model

2.3.1 General Introduction of Logistic Model

The logistic regression model, as a member of General Linear Models (GLM), is for categorical response variables. In general, the logistic model transforms the categorical response variables into logarithmic forms, which makes the forms of the coefficients of the explanatory variables consistent with other linear models. The general form of the logistic regression model is:

$$logit[\theta(x)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

The S-shaped curve in Figure 2-1 describes the shape of the logistic function. The X-axis is for explanatory variables and the Y-axis represents the probability of a response category given the values of the explanatory variables. There are three types of logistic regression, which depend on the type of categorical response variable: binary (or binomial) logistic regression, multinominal logistic regression, and ordinal logistic regression. The binary logistic model is used in this thesis and is applicable when the response variable is dichotomous and the explanatory variables are of any type. When categorical response variables have more than two classifications, multinomial logistic regression is used. Ordinal logistic regression is preferred to multinomial logistic regression when the categories of the response variable can be ranked from "low" to "high."

Logistic models have some advantages such as having no stringent assumptions about the explanatory variables. Logistic models do not require a linear relationship between the explanatory variables and the response variables. Moreover, logistic models do not have the stringent assumptions of normally distributed variables or homoscedasticity of the residuals.

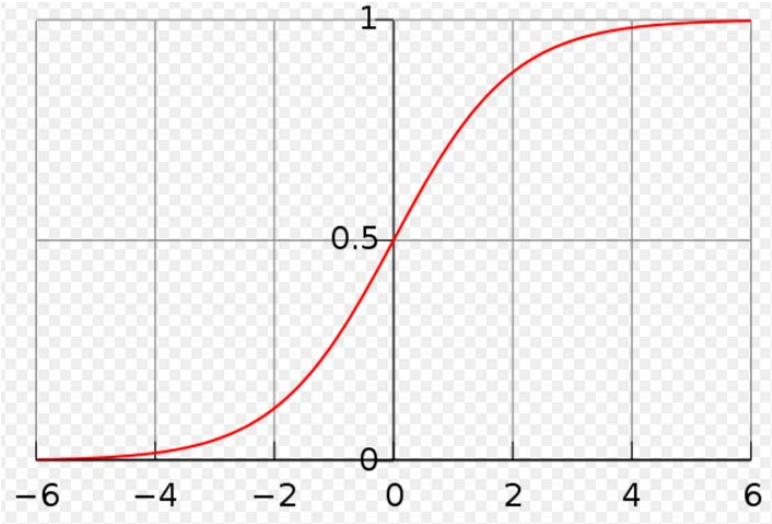

Figure 2-1 Standard logistic regression curve (Gershenfeld 1999).

Classification tables are used to examine the predictive success of a logistic regression model. Lift charts can be used to show the same information as the "Receiver Operating Characteristic" (ROC) curves to assess model fitness. Goodness-of-fit tests, such as the likelihood ratio test, are another way to test the appropriateness of the model. Wald statistics are used to test the significance of individual explanatory variables.

2.3.2 The Origin of Logistic Model

Belgian mathematician Pierre Francois Verhulst first developed the logistic model in 1838 (Cramer 2003). Verhulst suggested that population growth rates have limitations, for example, the rate may depend on population density. The equation is:

$$r = r_0(1 - \frac{N(t)}{K}) \quad (2)$$

In this equation, the function $N(t)$ represents the number of individuals at time t ; the constant r_0 represents the population growth rate in the absence of intra-specific competition; and K is the

carrying capacity, or the maximum number of individuals the environment supports. At low densities($N \ll 0$), the population growth rate r, is maximal and equals r_0. Population growth rates decline to 0 when $N(t) = K$. If $N(t) > K$, the population growth rate becomes negative. The solution of this model is:

$$N_t = \frac{N_0 K}{N_0 + (K - N_0)e^{-r_0 t}} \quad (3)$$

After Verhulst, physiologist T., Brailsford Robertson in 1908 applied the sigmoidal curve to individual growth in animals, plants, and man in two articles (Kingsland 1985). Robertson called his curve the "autocatalytic" curve or the self-accelerating curve when referring to only the accelerating part of the curve.

Pearl and Reed (1920) criticized Robertson's theory and reached the curve in Figure 2-1. Throughout the next twenty years, Pearl and his collaborators applied the logistic growth curve to almost all living populations and used it widely and indiscriminately during their career development. Yule's presidential address (Yule 1925) to the Royal Statistical Society of 1925 was an important publication for logistic model development and he was the person who named the model as *logistic*. By 1924, "logistic" had become a common word in the correspondence between Pearl and Yule (Cramer 2003). Reed and Berkson (1929) applied logistic models to analyzing autocatalytic reactions in chemistry. Reed and Berkson's work marked another early study of the applications of the logistic model.

2.3.3 Logistic Model Application History

Wilson and Worcester (1943) were probably the first to publish an application of the logistic model in bioassay, just before Berkson (1944). However, it was Berkson who persisted and fought for several decades for the application of the logistic model in bioassay. It was not

until the invention of computers and calculators that the ideological conflicted over bioassay and the disadvantage of the logistic model abated. Finney (1971), who had ignored the logistic model in the second edition of his textbook of 1952, made amends in the third edition of 1971 and recognized the power of the logistic model.

Around 1960, the logit terminology and the logistic model were more widely adopted and had their earliest developments in statistics and epidemiology. In statistics, Cox first recognized the power of the logit transformation while dealing with discrete binary outcomes. He wrote a series of papers (Cox 1958, 1966) and an influential textbook (Cox 1970). The rise of the logistic model in statistical literature is illustrated in Table 2-1, which shows how the number of articles increased substantially over time (Cramer 2003). Logistic models reached significant milestones when Berkson (1980) advocated minimum chi-square estimation and when Hosmer and Lemeshow (1989) published a comprehensive textbook about medical applications. In 1973, Nobel Prize winner Daniel L. McFadden linked the logit transformation to the theory of discrete choice in mathematical psychology (McFadden 2001). This success provided a theoretical foundation for the logistic model and thus advanced the use of the probit function in bioassay.

Currently, the logistic model is widely used in every field containing population data or categorical response variables. Those fields include wildlife, fishing, ecology, epidemiology, plant biology, and public health. For example, Ohlmacher and Davis (2003) used a multiple logistic model to predict landslide hazards in the state of Kansas. The explanatory variables in the model included digitized geology, slopes, and landslides. This model successfully indicated that the slope was the most important variable for estimating the probability of a landslide. Soil type and aspect ratio were also considered, but did not increase the predictive power of the

Table 2-1 Numbers of articles in statistical journals containing the word "logit".

Year	Logit
1935-39	-
1949-44	1
1945-49	6
1950-54	15
1955-59	23
1960-64	27
1965-69	41
1970-74	61
1975-79	72
1980-84	147
1985-89	215
1990-94	311

logistic model.

2.4 Biomass-using Facilities Siting Models

This thesis generates complex models for siting biomass-using facilities. The models involved economic influences, transportation-related factors, and the availability of biomass feedstocks. Based on the literature research, there is no paper either in the statistical area or in the forestry area that used logistic regression to examine the significant factors that influence the siting of biomass-using facilities. Papers cited in this chapter are related to siting model research, but used other methods.

Sperling (1984) established a generalized, non-statistical, analytical framework which was combined with a disaggregate microscale approach to identify and further specify the critical factors for assessing the quality of biomass locations in specific regions. The microscale approach was sensitive to variations in soil, topography, land ownership, water supply, water quality, electricity, transportation infrastructure, climate, and other variables. Five variables exerted the strongest influence on the siting and sizing of biomass fuel plants. The five were feedstock supply, fuel distribution, fuel demand, co-product demand, and feedstock processing. The analysis used a systematic framework to identify and integrate all the factors and could provide insight into formulating and analyzing public policies and actions.

Young et al. (1991) used a Geographic Information System (GIS) system, together with a spatial analysis, to assess the economic availability of woody biomass for potential sites for biorefineries in the Southeastern United States. They concluded that Northeast Florida, Southern Georgia, Southern Alabama, and the Coastal Plain of South Carolina were the lowest cost

regions for producing bioenergy from woody biomass. The South Delta of Louisiana, Kentucky, West Virginia, and the mountain regions of Tennessee and Virginia were the highest cost regions.

C.T. Donovan Associates, Inc. and Lee (1996) used a sensitivity analysis to determine the relative importance of various site characteristics to the overall financial performance of a biomass ethanol plant. This study provided information on likely market, market size, environmental regulations and incentives. The purpose of this research was to provide guidance for ethanol plant siting in the Northeastern U.S.

Knut et al. (2000) examined the environmental effects of paper mills in Norway. They studied the particular case of paper production at eight paper mills in relatively pristine environments in Norway. The study calculated the resource use, emissions, and environmental effects of the mills. They found that the actual siting decision also depended on consistent and durable economic and political value judgments, though it was helpful to reduce environmental damage by locating mills in pristine environments.

Moons et al. (2008) looked at the optimal location of new forests in a suburban region under area constraints. The methodology took into account use benefits, non-use benefits, opportunity costs of converting agricultural land, and planting and management costs of the new forest. The recreational benefits of new forest sites were estimated by using function transfer techniques. They found that the net social benefit of an afforestation project varied with the forest sites. The recreational value of a site varied considerably with the available substitutes.

Chapter 3 Methods

3.1 Variables Explanation

3.1.1 Response Variables Design

3.1.1.1 Coded "1" –existing mill locations

Two study groups are generated for the logistic regression models developed in the thesis. Locations that have an existing wood-using facility are coded as "1" in the data set for the logistic models. The two biomass-using facilities groups were:

1) All wood-using mills with wood-using bioenergy and biofuels plants;

2) Pulp and paper mills with wood-using bioenergy and biofuels plants.

Group I biomass-using facilities include primary wood processing mills, secondary wood processing mills, pulp and paper mills, and other mills. As defined by Perlack et al. (2005)[2], primary wood processing mills convert roundwood into other products. These wood processing mills include sawmills, medium density fiberboard (MDF), oriented strand board (OSB), particleboard, plywood, veneer post, pole, piling, dealer, yard, energy and wood chips. Secondary mills in Group I are mills that utilize the products of primary mills. Examples of secondary wood processing mill products include millwork, containers and pallets, buildings, furniture, flooring, paper and paper products. Secondary wood processing mills in this thesis not only include the above products, but also include planed wood products, remanufactured wood products, pallets, boxes, cabinets, trusses, mouldings, kiln dried products, treated wood products,

[2] Definitions of primary mill, second mill, pulp and paper mill are from Perlack et al. (2005).

plants, decking and siding. Other mills include forestry companies, logging mills, and companies that provide equipment and supplies, such as logging machine rental companies.

Pulp and paper mills are included in Group I biomass-using facilities, too. Wood-using bioenergy and biofuels plants (also called "biorefineries") are defined in this analysis as facilities that use all possible wood residues in an integrated biomass conversion process to produce biofuels, biopower, or biochemicals (National Renewable Energy Laboratory 2009). Twenty-nine bioenergy and biofuels plants are located within the 13-state region and are used in the analysis.

The locations of the Group I biomass-using facilities are plotted and displayed in Figure 3-1. Group II biomass-using facilities are plotted and displayed in Figure 3-2. These two plots show that the geographic dispersion of the two groups of woody biomass-using facilities are different. Each state has mills that are in Group I. Oklahoma, Texas, and Florida have the smallest quantity of mills compared to the large volume of mills in the other states. Figure 3-2 illustrates a high concentration of mills in the state of Georgia relative to the other states in this group. Oklahoma does not have any mills in Group II.

3.1.1.2 Coded "0" –"non-probable" locations

Some ZCTAs are not suitable locations to build woody biomass-using facilities because of their geographic and/or economic characteristics. For example, if a ZCTA has no land, no living trees, or is in a big city, this ZCTA is regarded as a "non-probable" location for the woody biomass-using facilities, and we code the response in this ZCTA as "0".

Specifically, three variables in Table 3-1 are used to define "non-probable" locations. A ZCTA is regarded as a "non-probable" location if Sqmiland = 0 (i.e., it has no land and may be

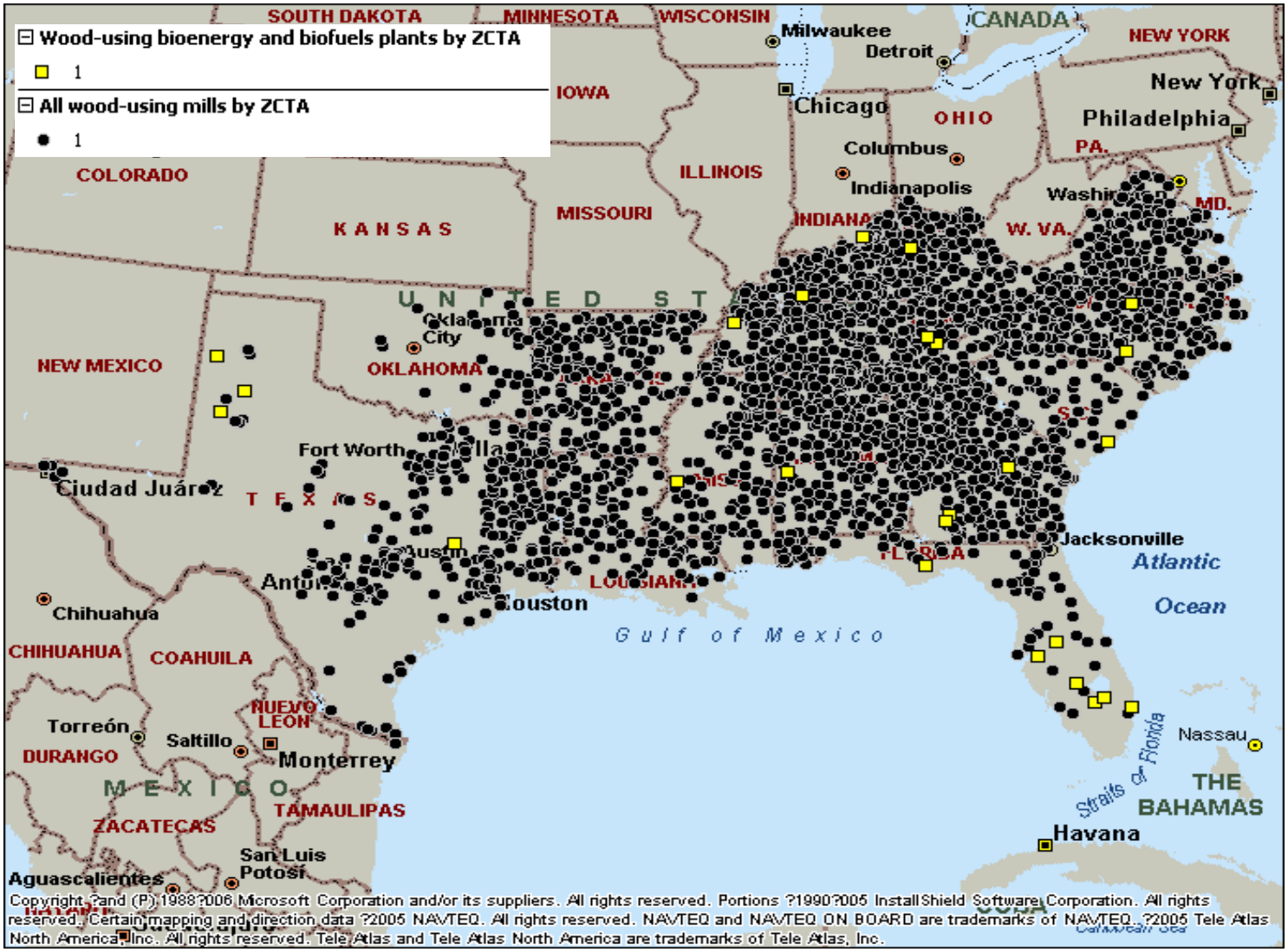

Figure 3-1 All wood-using mills with wood-using bioenergy and biofuels plants in 13 Southeastern states.

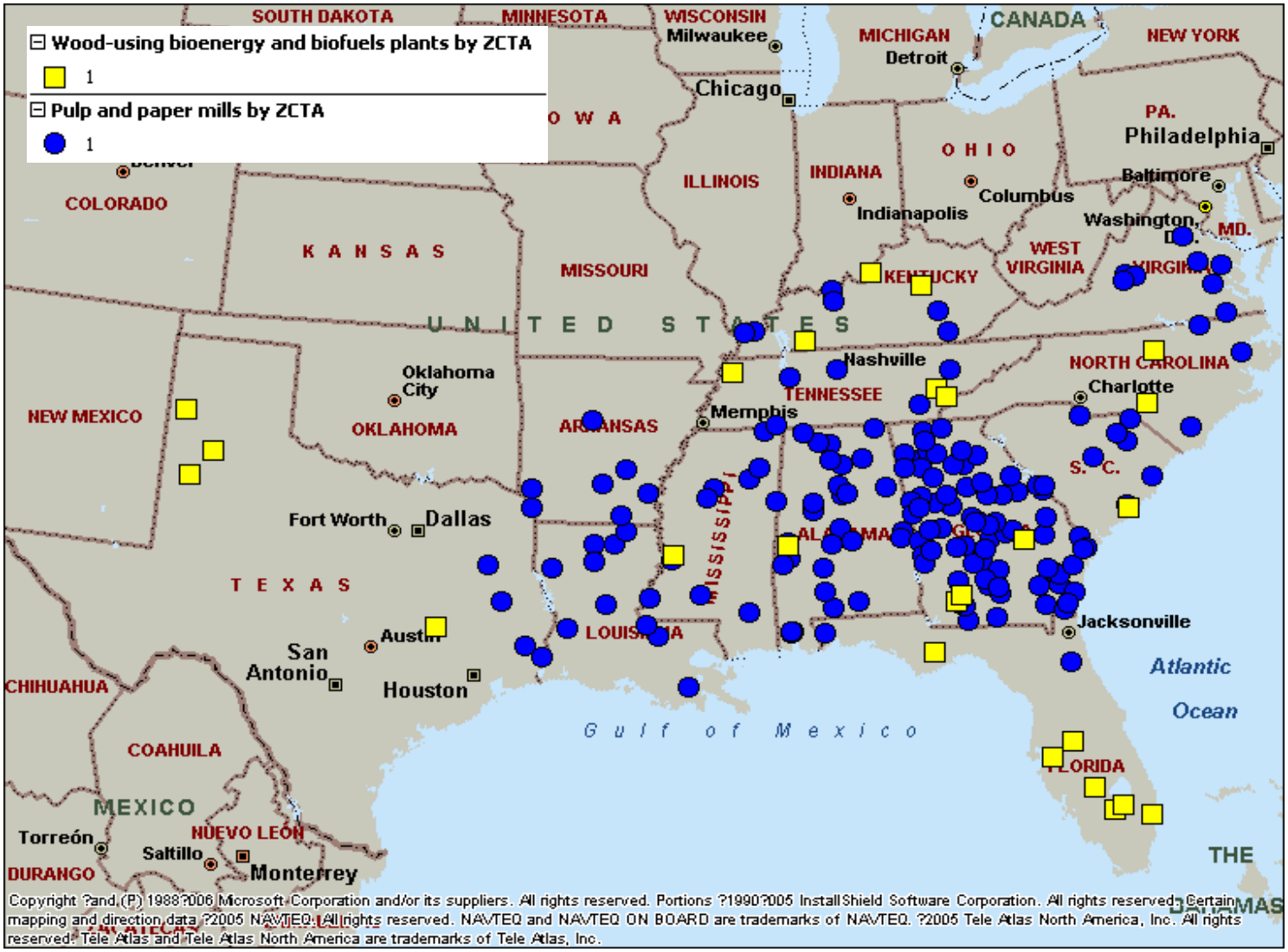

Figure 3-2 Pulp and paper mills with wood-using bioenergy and biofuels plants in 13 Southeastern states.

Table 3-1 Three variables for specifying "non-probable" locations of all wood-using mills with bioenergy and biofuels plants.

Variable Name	Variable Type	Collection level	Explanation
Sqmiland	Continuous	ZCTA	Land area ($mile^2$)
DRY_BIO_TOT	Continuous	ZCTA	Total standing volume in dry tons of all inventory species (trees >= 1.0 inches d.b.h[3].) on forestland (Perlack et al. 2005).
Metropolitan	Binary	City	Metropolitan or not ("1" for metropolitan area and "0" for not)

water, parks, or buildings), DRY_BIO_TOT = 0 (i.e., it has no living trees), or Metropolitan = 1 (i.e., it is in a metropolitan area).

3.1.2 Explanatory Variables

The 31 explanatory variables used for the analysis are listed in Table 3-2 with detailed explanations, units, collection levels, and variable types. For example, "Sqmiwater" is a continuous variable standing for the total water area, in square miles, in a ZCTA. Models of the Group I biomass-using facilitates use the first 28 explanatory variables in Table 3-2. Models of the Group II biomass-using facilities include the last three additional explanatory variables in Table 3-2 to verify their impacts on site locations. The names of the three additional variables are "Primary_mill_total," "Secondary_mill_total," and "Other_Mill_total." These three continuous variables stand for the number of primary wood processing mills, the number of secondary wood processing mills, and the number of other mills in each ZCTA. We also expect the models for Group II biomass-using facilities could exam the relationship between these three mills and the mills in Group II biomass-using facilities.

[3] d.b.h: "The diameter measured at approximately breast high from the ground" (Perlack et al. 2005).

Table 3-2 Explanatory variables for two groups of biomass-using facilities.

Variable name	Variable Type	Collection level	Unit	Explanation
Employment	Continuous	ZCTA	People	Employed person in all industries
Population	Continuous	ZCTA	People	Population in each ZCTA
Population_Density	Continuous	ZCTA	People /mile^2	Population density
Sqmiwater	Continuous	ZCTA	Mile^2	Water area
Median_Family_Income	Continuous	ZCTA	Dollar	Median of family incomes in 1999
Income_index	Continuous	ZCTA	Dollar/ people	Median family income per employed person
LOG_RES_HW	Continuous	ZCTA	Dry tons	Logging residues of hardwood, logging residues of softwood and the total of both. Logging residues are the unused portions of growing-stock and non-growing-stock trees cut or killed by logging and left in the woods (Perlack et al. 2005)
LOG_RES_SW				
LOG_RES_TOT				
OTHR_REM_HW	Continuous	ZCTA	Dry tons	Other removal of hardwood, other removal of softwood and the total of both. Other removal is the unutilized wood volume from cut or otherwise killed growing stock, from cultural operations such as precommercial thinning, or from timberland clearing. Does not include volume removed from inventory through reclassification of timberland to productive reserved forest land (Perlack et al. 2005)
OTHR_REM_SW				
OTHR_REM_TOT				
THIN_40	Continuous	ZCTA	Dry tons	The quantity of the selective removal of trees, primarily undertaken to improve the growth rate or health of the remaining trees, within 40, 80, 120, and 200 miles haul distances (Perlack et al. 2005)
THIN _80				
THIN_120				
THIN_200				
TOTAL_MILL_RES	Continuous	ZCTA	Dry tons	Residues generated from primary mills, secondary mills and pulp and paper mills, which include bark, coarse residues (chunks and slabs), fine residues (shavings and sawdust), sawdust, sander dust, wood chips and shavings, board and cut-offs, miscellaneous scrap wood and black liquor (“solution of lignin-residue and the pulping chemicals used to extract lignin during the manufacture of paper” (Perlack et al. 2005)

Table 3-2 (Continued)

Variable name	Variable Type	Collection level	Unit	Explanation
UNUSED_MILL_RES	Continuous	ZCTA	Dry tons	Mill residues have not be used for wood-using biorefinery facilities or other wood processing mills
MCost_p5M MCost_1M MCost_1p5M	Continuous	ZIP Code	Dollar /mile	Marginal trucking cost of total mill residues within an 80-mile haul distance under 0.5, 1, and 1.5 million dry tons annual demand quantities
TCost_80	Continuous	ZIP Code	Dollar /mile	Total trucking cost of total mill residues within an 80-mile haul distance
ACost_80	Continuous	ZIP Code	Dollar /mile	Average trucking cost of total mill residues within an 80-mile haul distance
TQty_80	Continuous	ZIP Code	Dry tons	Cumulative quantity of mill residues in each ZCTA within an 80-mile haul distance
URBAN_WASTE	Continuous	ZCTA	Dry tons	Municipal solid waste (MSW) and construction and demolition debris (Perlack et al. 2005)
Log_Res_Harvest_Cost	Continuous	County	Dollar /dry ton	Harvesting cost of logging residues
RailroadAvailability	Ordinal	ZIP Code		Railroad accessible index ranked by four railroad companies as N/A, 1, 2, 3, and 4. "N/A" means this ZIP Code has no railroad; "1" means one out of four railroad companies ranks this ZIP Code as having railroad access and so on. Larger numbers mean that the ZIP Code has more railroad access.
Numberports	Continuous	ZCTA	Port	Number of water ports in each ZCTA
Primary_mill_total	Continuous	ZCTA	Mill	Number of primary wood processing mills in each ZCTA
Secondary_mill_total	Continuous	ZCTA	Mill	Number of secondary wood processing mills in each ZCTA
Other_Mill_total	Continuous	ZCTA	Mill	Number of other mills in each ZCTA

3.2. Data Management and Data Quality

3.2.1 Data Management Tools and Database Structures

This research involves large volumes of data that were stored in different formats and that came from different sources. Scrutiny is necessary to ensure data accuracy and data quality. Various software are used for data generation, verification and combination. SAS® 9.1 and SAS® PROC SQL are used for reformatting data from different resources, for merging data sets containing different explanatory variables, and for querying data for verifying the data quality. JMP 7.0.1 and Mirosoft® Excel 2007 are used as supplementary tools for data organization and verification. MATLAB and MapPoint® are used to calculate the driving times and driving distances of ZIP Code pairs (i.e., ZIP1 and ZIP2) as shown in Table 3-3.

There are 82 million records in the data that Table 3-3 describes, which are stored in a SQL server database for the BioSAT model (Young et al. 2008, see www.BioSAT.net). Some data used for this thesis are extracted from the BioSAT SQL server database and include:

Table 3-3 A ZIP Code pair and its driving distance and driving time.

Name	**Explanation**
ZIP1	ZIP Code list of all 13 Southeastern states
ZIP2	80-mile haul distance ZIP Code list within each ZIP1
Driving time	Driving time (minutes) for each pair of ZIP Codes
Driving Distance	Driving distance (miles) for each pair of ZIP Codes

- Marginal cost of delivered total mill residues under 0.5 million, 1 million, and 1.5 million dry tons demand quantity;
- Average cost and total cost of total mill residues within an 80-mile haul distance;
- Total quantity of total mill residues within an 80-mile haul distance.

3.2.2 Data Resources and Data Collection Levels

Because the data come from various sources, the data had different levels of resolution. There are four levels of resolution in the data:

- 5-digit ZIP Code;
- U.S. Census Bureau 5-digit ZIP Code tabulation area (ZCTA);
- City;
- County.

The data sets for the analysis are developed using SAS® PROC SORT and PROC MERGE. Data sets with the same level of resolution are merged directly. Data sets with different levels of resolution are merged based on the corresponding relationship between the hierarchical structures of ZCTAs, ZIP Codes, cities, and counties. Missing values are surrogated after data at the same level were merged.

Data sources are the U.S Census Bureau (2000), U.S. Forest Services (Perlack et al. 2005), various internet sources[4], railroad companies[5], physical, telephone conferences, and

[4] IEA Bioenergy Task 39 (2009), Renewable Fuels Association (2009), University of Wisconsin-Milwaukee Employment and Training Institute (2000), U.S. Army Corps of Engineers Navigation Data Center (2008), U.S. Census Bureau (2000), U.S. etc.

[5] Burlington Northern Santa Fe Railway (2009), CSX Corporation, Inc. (2005), CSX Corporation, Inc. (2009), Norfolk Southern System (2009), and Union Pacific (2009)

emails for data inquiries[6]. Some data are calculated based on a real time trucking cost model or algorithms from Young et al. (2008) and Berwick and Farooq (2003). For example, all marginal costs were generated based on the real-time trucking cost model. More detailed explanations of the two algorithms are in Section 3.2.5.

3.2.3 Data Accuracy and Consistency

To maintain acceptable data quality, ZCTA-level data in the thesis follows the guidelines of the Census Bureau (2000). Note that no ZCTA has missing values for any of the economic variables (e.g., Population, Median_Family_Income, Population_Density, Employment, Sqmiwater and Income_index). Data for woody-biomass using facilities (e.g., primary mills, secondary mills, pulp and paper mills, and other mills) contain approximately 1,800 types of businesses. Several biomass-using facilities have multiple business classifications. For example, a company can produce products that are categorized as coming from different types of biomass-using facilities. This same company may have several office branches in different ZIP Codes and states. Research Specialist Andrea Noehmer categorized the 1,800 business types into 17 businesses groups. The data management procedure considers the data consistence as well as data accuracy.

3.2.4 Data Management and Missing Value Surrogate Methods

Practical and meaningful combination rules are used when combining the subsets of grouped data. When merging ZIP Code level data into ZCTA level, several different methods are used according to the different features of the variables. For example, there are three

[6] Personal communication: Pemberton Truck Lines (Knoxville, TN), 09/ 2008; Skyline Transportation, Inc. (Knoxville, TN), 09/ 2008; Mason Dixon (Knoxville, TN), 09/ 2008; Mason Dixon (Scottsboro, AL), 09/ 2008; Patterson Chip Company (Lily, KY), 11/ 2008; GFI Transport (Mount Joy, PA), 11/ 2008; Tennessee Department of Agriculture (Nashville, TN), 11/ 2008; Carlen Transport Inc (Hampden, ME), 11/ 2008; Gene A. Matt Trucking (Omak, WA), 02/ 2009; GCS Logging (Cambridge, NY), 02/ 2009; Gene A. Matt Trucking (Omak, WA), 02/ 2009.

variables for marginal cost (MC), MC under 0.5 million dry tons annual quantity demanded ("MCost_p5M"), MC under 1 million dry tons annual quantity demanded ("MCost_1M"), and MC under 1.5 million dry tons annual quantity demanded ("MCost_1p5M"). The minimum values for marginal cost among multiple ZIP Codes within each ZCTA are selected, which assumes that potential sites will try to minimize costs. For the average cost and total cost of total mill residues within an 80-mile haul distance ("Acost_80" and "TCost_80"), and also for the total quantity of total mill residues within an 80-mile haul distance ("TQty_80"), the average value among multiple ZIP Codes within each ZCTA is chosen.

In this thesis, a very large number is given to missing values of the cost variables. For MCost_p5M, MCost_1M, and MCost_1p5M the value is 9999. For TCost_80 and ACost_80 the value is 99999999. A "0" is given to missing values of TQty_80. The reason for surrogating missing values as above is to avoid locating potential sites into ZCTAs with a missing value. Table 3-4 provides details of data management methods and surrogated missing value numbers for six ZIP Code level variables.

Table 3-4 ZIP Code level variables combination values and surrogating number for missing values.

Variable Name	**Combination values**	**Surrogating numbers for missing values**
MCost_p5M	Minimum value	9999
MCost_1M	Minimum value	9999
MCost_1p5M	Minimum value	9999
TCost_80	Average value	99999999
ACost_80	Average value	99999999
TQty_80	Average value	0

When data are combined at the ZCTA level, county level, and city level, missing values of all other variables are replaced with "0". The variables Population, Median_Family_Income, Sqmiwater, Sqmiwater, Population_Density, and Income_index are not allowed to have missing values since the economic information is unique for each ZCTA and cannot be substituted by extrapolation. The record length of the data set used for analysis is 8,833 records out of 9,416 ZCTAs. There are 28 explanatory variables for Group I biomass-using facilities and 31 explanatory variables for Group II biomass-using facilities.

3.2.5 Algorithms for Data Generation

Variables like MCost_p5M, MCost_p5M, MCost_1p5M, TCost_80, ACost_80 and TQty_80 are based on the BioSAT model (Young et al. 2008). The SQL server database for BioSAT contains 82 million cost records for 33 Eastern states. This database contains driving time and driving distance for each pair of ZIP Codes in the 13 Southeastern states for up to a 200-mile one-way haul distance. The database was created by first using the neighboring ZIP Code algorithm to generate a list of ZIP Codes with MATLAB modules and then feed into a Visual Basic program.

3.2.5.1 Neighboring ZIP Code Algorithm

The purpose of the algorithm is to find the neighboring ZIP Codes within a target driving distance (e.g. 80 miles) of each ZIP Code. Because calculating the driving distance is much more time consuming than calculating the sphere distance between two ZIP Codes, we first calculate the sphere distances between a given ZIP Code and every other ZIP Code and then filter out those ZIP Codes with a sphere distance over 200 miles to the given ZIP Code. For a given ZIP Code, its sphere distance to another ZIP Code is calculated by utilizing their longitudes and latitudes (Moritz 2000 and Wang 2008):

$$D = \sqrt{(Md\phi)^2 + (N\cos\phi\, d\lambda)^2} \tag{4}$$

ϕ --- 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 .

The driving time and driving distance data are calculated for the ZIP Code pairs with a sphere distance of no more than the target distance (e.g. 80 miles). The detailed procedure is summarized as follows:

1) For a given ZIP Code, compute sphere distance to any other ZIP Code,
2) Get potential neighboring ZIP Codes, which have a sphere distance of no more than a target driving distance (e.g. 80 miles),
3) Calculate the real driving distances to the potential neighboring ZIP Codes. The real-time driving distance is computed by MapPoint® 2006,
4) Among the potential ZIP Codes, those with real-time driving distances no more than the target distance (e.g. 80 miles) are defined as the nearest neighboring ZIP Codes of the given ZIP Code,
5) Repeat the above steps to find the nearest neighboring ZIP Codes for all other ZIP Codes needed.

The MATLAB module for this algorithm is in Table A-1 (Appendix).

3.2.5.2 Trucking Cost Generation Algorithm

The values of MCost_p5M, MCost_1M, MCost_1p5M, TCost_80, ACost_80 and TQty_80 are calculated by using the following algorithm, where ZIP1 is a target ZIP Code and ZIP2 is the neighboring ZIP Code of Zip 1:

Step 1: Sort data by target ZIP Code (Zip 1) and then by driving distance in ascending order.

Step 2: Within each Zip 1,

1. TCost_80 = 0, TQty_80 = 0,

   MCost_p5M = MCost_1M = MCost_1p5M = 9999.

2. Loop from the 1st ZIP2 to the last ZIP2 associated with the current Zip 1

   2.1 prevTCost_80 = TCost_80,

   prevTQty_80 = TQty_80,

   TCost_80 = TCost_80 + cost of the current Zip 2,

   TQty_80 = TQty_80 + quantity of the current Zip 2.

   2.2 If TQty_80 > 1.5 million and MCost_1p5M = 9999

   MCost_1p5M = (TCost_80 - prevTCost_80) / (TQty_80 - prevTQty_80),

   else if TQty_80 > 1.0 million and MCost_1M = 9999

   MCost_1M = (TCost_80 - prevTCost_80) / (TQty_80 - prevTQty_80),

   else if TQty_80 > 0.5 million and MCost_1p5M = 9999

   MCost_1p5M = (TCost_80 - prevTCost_80) / (TQty_80 - prevTQty_80).

3. ACost_80 = TCost_80 / TQty_80.

3.3 Logistic Models

3.3.1 Logistic Models' Introduction

A logistic model is a member of the generalized linear model (GLM) family. Compared to other models that have continuous response variables, the response variable of a logistic model is categorical, i.e., discrete, dichotomous, or ordinary. On the other hand, the explanatory variables have no limitations on data types. Generally, the response variable is dichotomous, such as win/loss or success/failure. The response variable can take the value "1" with probability θ of success, or the value "0" with probability $(1-\theta)$ of failure. The advantage of a logistic model is that the explanatory variables' types can be discrete, continuous, dichotomous, or mixed. Moreover, logistic regression has no limitations on the distributions of the explanatory variables. It is not necessary for the explanatory variables to be normally distributed, to be linearly related, or to have equal variance within each group. However, since the response variable is either "0" or "1" with probability of θ or 1-θ respectively, the function of an explanatory variable on the response variable is not linear. Instead, logistic regression uses a logarithm transformation to the odds $\frac{\theta}{1-\theta}$ to transform the range of response result to a real number. Then the probability θ of success or of "1" is written as :

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{5}$$

where

α = The constant of the equation

β = The coefficient of the predictor variables

An alternative form of the logistic regression equation is:

$$logit[\theta(x)] = \log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \,. \tag{6}$$

3.3.2 Model Selection Methods and Criteria

The goal of logistic regression is to predict an outcome correctly using the most parsimonious model. A parsimonious model includes only explanatory variables that are powerful in predicting the response variable. Three common methods for finding models that contain only variables that are powerful in predicting the response variable are forward selection, backward selection, and stepwise selection. In a common version of forward selection, variables enter the model one by one, where the variable added at each step is the variable that leads to the largest R-square improvement. In backward selection, all of the variables are in an initial model and then the variables are removed from the model one by one to see the improvement in a certain criteria, such as Akaike's Information Criterion (AIC) (McQuarrie and Tsai 1998). In stepwise selection, variables can both enter and exit the model. None of these three methods necessarily identifies the "best model." Because the selection methods work by fitting an automated model to the current data set, they might not examine the combination of variables that produces the best mathematical criteria and they raise the danger of overfitting the model. However, some criteria help protect against the danger of overfitting that emerges from the stepwise procedure. The criterion used in this paper is the Bayesian Information Criterion (BIC), which is defined as

$$-2L_m + m\ln(n), \tag{7}$$

where n is the sample size, L_m is the maximized *log-likelihood* of the model, and m is the number of parameters in the model.

The BIC takes into account both the statistical goodness-of-fit and the number of parameters in order to avoid overfitting the final model (McQuarrie and Tsai 1998).

3.3.3 Model Assessment Tools

3.3.3.1 Lift Charts

According to Larose (2005), lift charts and gain charts (cumulative lift charts) are graphical evaluative methods for assessing the predictive power of models. Lift charts seek to compare response rates with and without the predictive model. For example, we build a model for classifying how many numbers within the whole 300 real-number data set is positive. The model classifies 100 numbers as positive, 80 of which are correct. In addition, in the raw data set 200 of 300 numbers are positive. Then a lift value for the sample size of 100 is (80/100) / (200/300) = 1.2. Lift is a function of sample size, which is why we have to specify that the lift of 1.2 for the model is measured for *n = 100* records. Lift charts are plotted by putting lift values (1.2) on the y-axis and the percentage of samples drawn from the raw data set on the x-axis (In the previous case, the percentage is 200/300=0.666). When comparing different models, the larger the lift value, the better the model is. Cumulative lift charts compare models at the whole sample size level. Non-cumulative lift charts compare models at a decile level.

3.3.3.2 Classification Tables

Classification tables are useful for summarizing the predictive power of a binary logistic regression model. The classification table cross classifies the binary outcome y with a prediction $\hat{y}$ =0 or $\hat{y}$ =1 under a cutoff π_0. The prediction of y is $\hat{y}$ = 1 if $\hat{\pi}_i > \pi_0$ and $\hat{y}$ = 0 if $\hat{\pi}_i \leq \pi_0$. The two useful summaries of predictive power are sensitivity = $P(\hat{y} = 1 \mid y = 1)$ and specificity = $P(\hat{y} = 0 \mid y = 0)$.

The overall proportion of correct classifications is:

$$P(correct\ classification) = P(\hat{y} = 1 \text{ and } y = 1) + P(\hat{y} = 0 \text{ and } y = 0)$$

$$= P(\hat{y} = 1 \mid y = 1)P(y = 1) + P(\hat{y} = 0 \mid y = 0)P(y = 0). \quad (8)$$

The overall proportion is a weighted average of sensitivity and specificity. Sensitivity provides a rate of actual positives ($y = 1$) that are correctly identified ($\hat{y} = 1$). Specificity measures the proportion of negatives ($y = 0$) which are correctly identified ($\hat{y} = 0$).

3.3.4 Four Optional Models

To find the best models for the two biomass-using facilities groups, this thesis uses four ways to build logistic models in SAS® Enterprise Miner. The plot of these four ways is Figure 3-3. All four ways apply the Data Partion node and evaluate stepwise logistic models using the BIC criterion. Differences among them are in the use of the Variable Selection node and the Transform Variables node. For data partation, the data set is partitioned into two parts: 60% of the data was randomly selected as the training set, and 40% of the data was randomly selected as a validation set. The training data set was used to develop the logistic models; the validation data set was used to evaluate the performance of the logistic models. The four ways of building the logistic model are detailed as follows:

1) Data partition and stepwise variable selection for the logistic model using the BIC criterion;
2) Variable Selection[7] node first, followed by data partition and stepwise variable selection for the logistic model using the BIC criterion;
3) All variables except Median_Family_Income are transformed into logarithm form, followed by data partition and stepwise variable selection for the logistic model using the BIC criterion.

[7] The Variable Selection node eliminates variables that do not have an R-square improvement of 0.0005 with the response. Two-way interactions are included with the other explanatory variables.

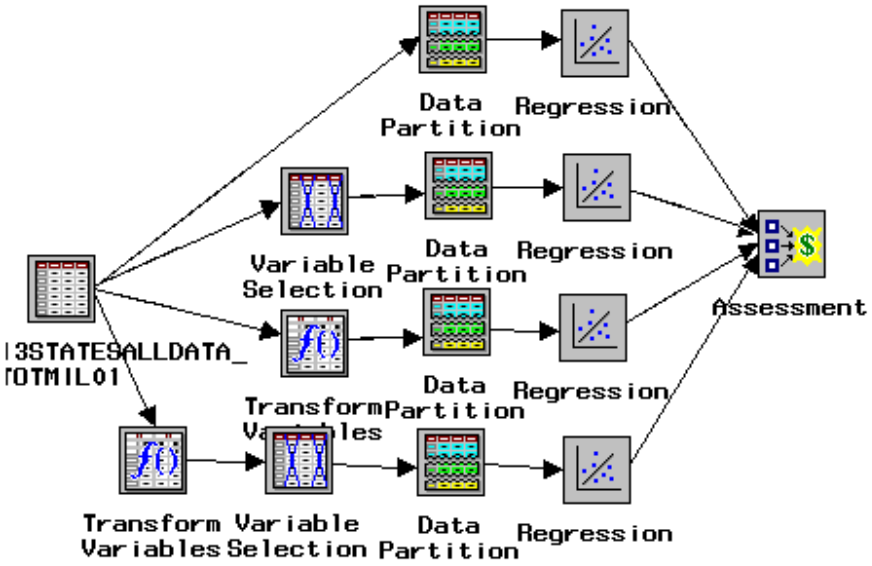

Figure 3-3 Four optional models in SAS® Enterprise Miner.

4) All variables except Median_Family_Income are transformed into logarithm form, followed by the Variable Selection node, followed by data partition and stepwise variable selection for the logistic model using the BIC criterion.

Thus, we name the four optional models in Figure 3-3 from the top to the bottom as "Stepwise only", "Transform all with stepwise", "Variable selection with stepwise", and "All transform variable selection stepwise", respectively.

These four optional models are applied to each biomass-using facilities group within SAS® Enterprise Miner. Then, the four optional models are compared with the BIC criterion and with lift charts and their predictive abilities are evaluated with classification tables. A best model is chosen for each of the two biomass-using facilities groups. Utilizing the chosen best models, the "score functions" are used to predict the scores of being a future Group I or II facility siting location for all ZCTAs without response variable values (i.e., 0 and 1). Based on

the best model of each group of biomass-using facilities, we want to extract the following information of interest:

1) The significant factors in the 13-state regional level based on the selected best model for each biomass-using facilities group,
2) The significant factors in each state level and the cross-state significant factors, based on the selected best model for each biomass-using facilities group,
3) The top twenty-five potential siting locations for each biomass-using facilities group at the 13-state regional level.

Chapter 4 Results and Discussion

This chapter presents the logistic regression results for the two predefined groups of biomass-using facilities. The predicted top 25 potential siting locations for each group of biomass-using facilities are mapped at the 13-state regional level. Lastly, a de-clustering algorithm is used for each group of biomass-using facilities to make the potential locations more feasible.

4.1 Logistic Regression Results for Group I Biomass-using Facilities

The significant factors for the 13-state regional level are discussed. At the state level, significant factors are listed for all 13 states for comparison. The cross-state significant factors are highlighted. The top 25 potential locations are mapped at the 13-state regional level using the software MapPoint® 2006.

4.1.1 Models Assessment for Group I Biomass-using Facilities

The four optional models outlined in Figure 3-3 are applied to the Group I biomass-using facilities. The BIC values are summarized in Table 4-1 for all four optional models. Cumulative

Table 4-1 Model assessment results by BIC criterion for Group I biomass-using facilities.

Model name	BIC Value	Misclassification Rate
"Stepwise only"	2031.6822138	0.1319134318
"Transform all with stepwise"	2033.7368535	0.394709722
"Variable selection with stepwise"	2008.6824715	0.1322569564
"All transform variable selection stepwise"	2034.7414806	0.135692202

and non-cumulative lift charts are presented in Figures 4-1 and 4-2 for all four optional models, respectively. The two lift charts illustrate that the four optional models for Group I biomass-using facilities have no significant difference in terms of the lift values. The BIC values show that the best model is the second (i.e., "Transform all with stepwise") model. Table 4-1 shows that this model has the lowest BIC score, 2008.6824715, and the second-lowest misclassification rate, 0.1322569564. The difference between the misclassification rate for this model and the lowest misclassification rate is less than 0.00035. Recall that the model-building steps for this model are: variable selection, followed by data partition, followed by stepwise variable selectionof the logistic model using the BIC criterion.

4.1.2 Predictive Ability Measured by Classification Tables

Classification tables in Table 4-2 and summarized in Figure 4-3 measure the predictive power of the selected best model. The results for the aforementioned model are:

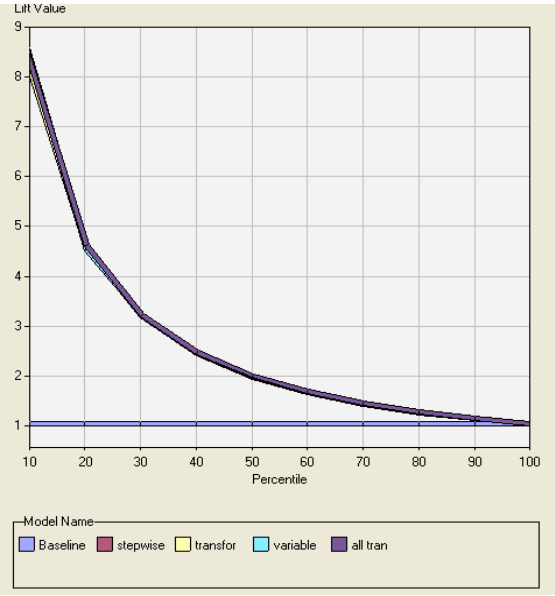

Figure 4-1 Cumulative lift charts that assess four optional models for Group I biomass-using facilities.

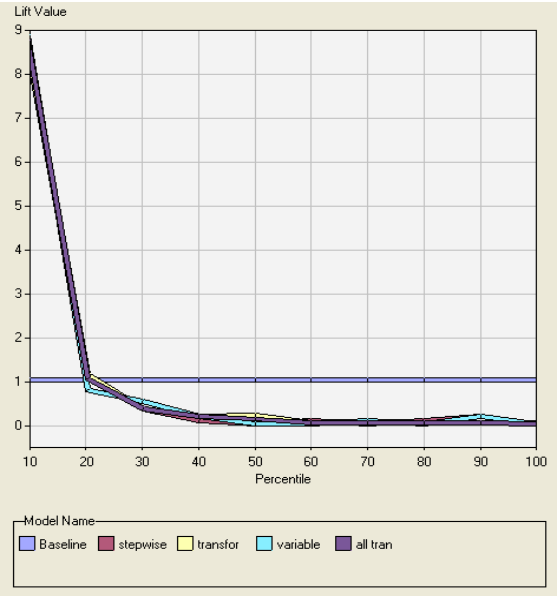

Figure 4-2 Non-cumulative lift charts that assess four optional models for Group I biomass-using facilities.

Table 4-2 Classification Table of predictive ability measurement for Group I biomass-using facilities.

Predictive value Actual value	0	1	Total
0	1218 (87.56%)	173	1391 (47.78%)
1	233	1287 (84.67%)	1520 (52.22%)
Total	145	1460	2911

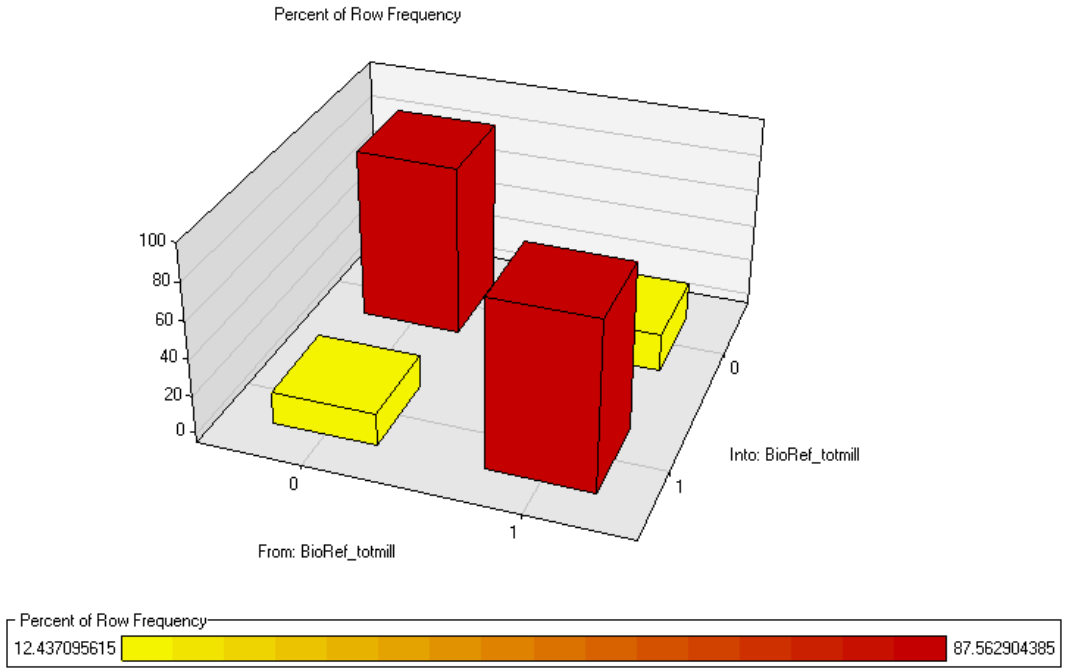

Figure 4-3 Predictive ability plot based on the classification table for Group I biomass-using facilities.

Sensitivity = $P(\hat{y} = 1 \mid y = 1) = 84.67\%$,

Specificity = $P(\hat{y} = 0 \mid y = 0) = 87.56\%$.

The overall proportion of correct classifications is

$$P(correct\ classification) = P(\hat{y} = 1 \mid y = 1)P(y = 1) + P(\hat{y} = 0 \mid y = 0)P(y = 0)$$

$$= 0.8467 * 0.5222 + 0.8756 * 0.4778$$

$$= 86.05\%.$$

4.1.3 Regional Level Analysis Result for Group I Biomass-using Facilities

Using the best model found in Subsection 4.1.1 reduces the twenty eight explanatory variables to the five variables in the model that are statistically significant (p-value < 0.05). The Likelihood ratio goodness-of-fit test (Agresti 2007) shows that the model fits the data well. Significant variables are Median_Family_Income, Thin_80, Unused_MILL_RES, and

Log_Res_Harvest_Cost. RailroadAvailability is significant based on the Type 3 test[8] (p_value <0.0001). Results of the Type 3 test and the maximum likelihood estimates are displayed in Tables 4-3 and 4-4.

Based on the results of the logistic regression, median family income (Median_Family_Income) and harvesting cost for logging residues (Log_Res_Harvest_Cost) have a negative coefficient, so the probability of siting is higher when these variables are lower. The feedstocks such as thinnings within an 80-mile haul distance (THIN_80) and unused mill residues (UNUSED_Mill_RES), and railroad availability (RailroadAvailability 2) all have positive coefficients. The significant variables for Group I biomass-using facilities are consistent with rational economic expectations.

Based on this regression, the best 25 potential locations at the 13-state regional level are estimated and plotted in Figure 4-4, i.e., these 25 ZCTAs having the highest probability. There are ten possible locations in Mississippi, eight in Tennessee, three in Virginia, three in Louisiana,

Table 4-3 Type 3 analysis of effects for Group I biomass-using facilities.

Effect	**DF**	**Wald Chi_Square**	**Pr>ChiSq**
Log_Res_Harvest_Cost	1	144.9197	<0.0001
RailroadAvailability	2	34.0811	<0.0001
Median_Family_Income	1	15.3387	<0.0001
THIN_80	1	184.6736	<0.0001
UNUSED_Mill_RES	1	11.3224	0.0008

[8] The Type 3 test is a more powerful test of parameters for group variables because tests of the parameter estimates can only examine the groups individually (e.g., RailroadAvailability 1 vs N/A and RailroadAvailability 2 vs N/A).

Table 4-4 Analysis of maximum likelihood estimates of parameters for Group I biomass-using facilities.

Parameter	**DF**	**Estimate**	**Wald Chi_Square**	**Pr>ChiSq**
Intercept	1	0.0577	0.10	0.07549
Log_Res_Harvest_Cost	1	-0.00017	144.92	<0.0001
RailroadAvailability 2	1	0.4064	25.59	<0.0001
Median_Family_Income	1	-0.00001	15.34	<0.0001
THIN_80	1	0.00155	184.67	<0.0001
UNUSED_Mill_RES	1	0.0299	11.32	0.0008

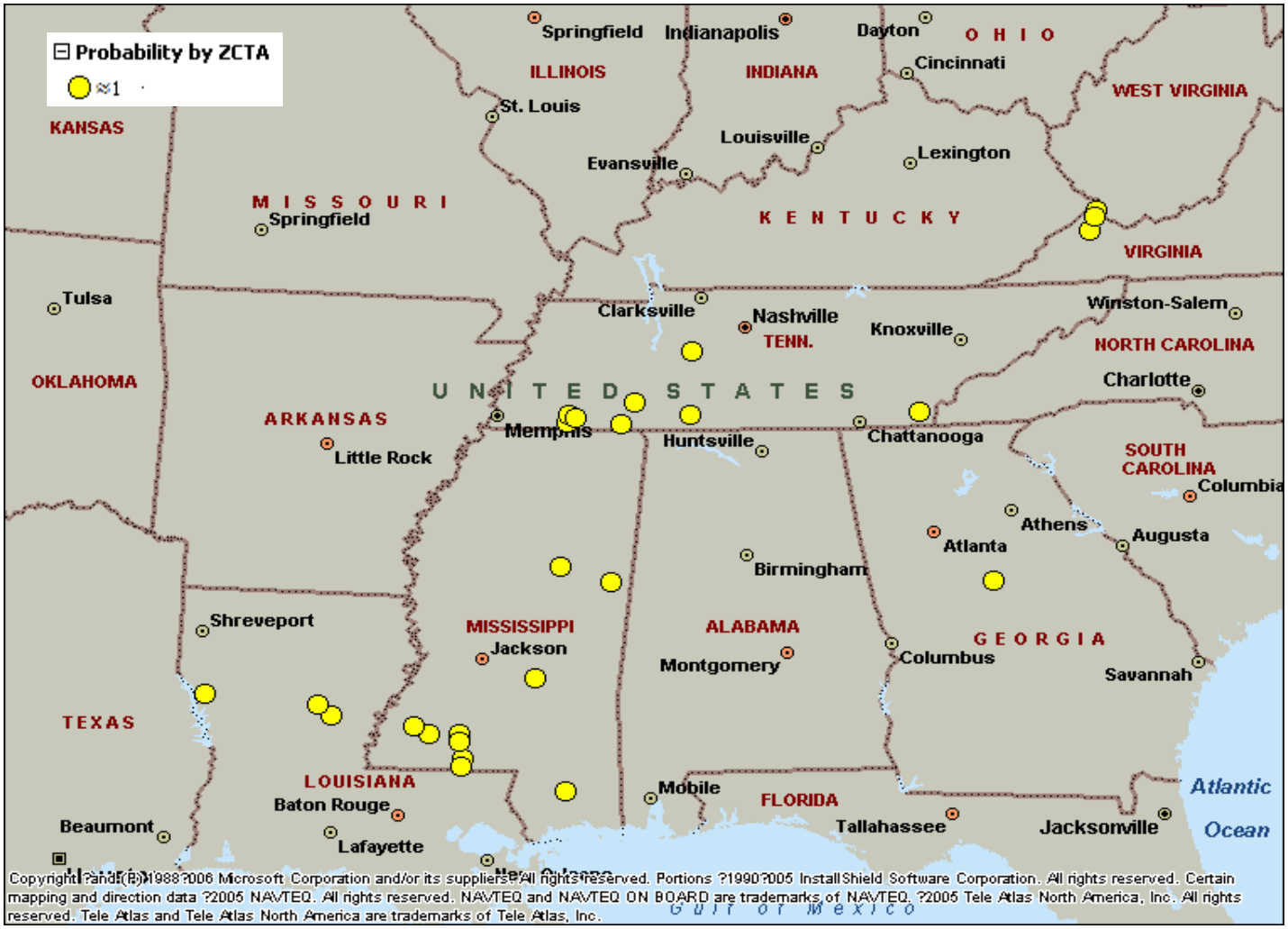

Figure 4-4 Top 25 optimal locations for Group I biomass-using facilities at the 13-state regional level.

and one in Georgia.

4.1.4 State Level Analyses for Group I Biomass-using Facilities

Separate state level analyses are performed to examine the statistically significant variables at the state level. The results in Table 4-5 indicate that population is the most important variable across multiple states. Population is significant in 6 out of the 13 states and has a negative impact on siting decisions for Group I biomass-using facilities. Thinnings within a 40-mile haul distance (THIN_40) is the second-most important variable for Group I biomass-using facilities. Thinnings is significant in 4 out of 13 states and has a positive impact on siting decisions for Group I biomass-using facilities. Population, median family income (Median_Family_Income), and total cost of mill residues within an 80-mile haul distance (TCost_80) all have negative impacts on siting decisions for Group I biomass-using facilities in at least one state. Railroad availability (RailroadAvailability), thinnings within 80-mile and 200-mile haul distances (THIN_40 and THIN_200), water area (sqmiwater), total mill residues within an 80-mile haul distance (TQty_80), and other removal of hardwood and softwood (OTHR_REM_HW and OTHR_REM_SW) all have positive impacts on siting decisions for Group I in at least one state.

The results suggest that significant factors that affect potential locations for Group I are state dependent. For example, South Carolina's potential locations for Group I are ZCTAs with less population and a large volume of thinning within an 80-mile haul distance. North Carolina's best locations for Group I are ZCTAs with less population and a large volume of thinnings within a 160-mile haul distance. Arkansas' Group I preferred sites which have low family income and large volume of other removals of hardwood.

Table 4-5 State level analysis results for Group I biomass-using facilities.

State Name		TN	TX	FL	AL	LA	AR	KY	VA	SC	GA	NC	OK	MS
Population	-		X[9]	X		X		X		X		X		
THIN_40	+		X	X	X	X								
RailroadAvailability 1	+	X		X				X						
Median_Family_Income	-				X		X	X						
THIN_80	+								X	X				
RailroadAvailability 2	+	X						X						
THIN_200	+	X												
Sqmiwater	+	X												
TCost_80	-								X					
OTHR_REM_HW	+						X							
TQty_80	+										X			
OTHR_REM_SW	+										X			
THIN_160	+											X		

[9] The "X" in red means that the sign is the opposite of the possible sign listed in the column after the variable names.

In summary, state-level results suggest that the better locations for Group I are ZCTAs with low population, low family income, water availability, large volume of feedstocks, access to railroads, and low harvesting costs.

4.2 Logistic Regression Results for Group II Biomass-using Facilities

The significant factors for the 13-state regional level are discussed. At the state level, significant factors are listed for all 13 states for comparison. The cross-state significant factors are highlighted. The top 25 potential locations are mapped at the 13-state regional level.

4.2.1 Models Assessment for Group II Biomass-using Facilities

The same four optional models as in Figure 3-3 are used for the Group II biomass-using facilities. The BIC values are summarized in Table 4-6 for all four optional models. Cumulative and non-cumulative lift charts are presented in Figures 4-5 and 4-6 for all four optional models, respectively. The two lift charts illustrate that the four models for Group II biomass-using facilities have no significant difference in terms of the lift values. The BIC values show that the

Table 4-6 Model assessment results by BIC criterion for Group II biomass-using facilities.

Model name	**BIC value**	**Misclassification Rate**
"Stepwise only"	365.90341332	0.0265087422
"Transform all with stepwise"	400.19496559	0.0321489002
"Variable selection with stepwise"	374.5732677	0.0276367738
"All transform variable selection stepwise"	411.82338301	0.0344049633

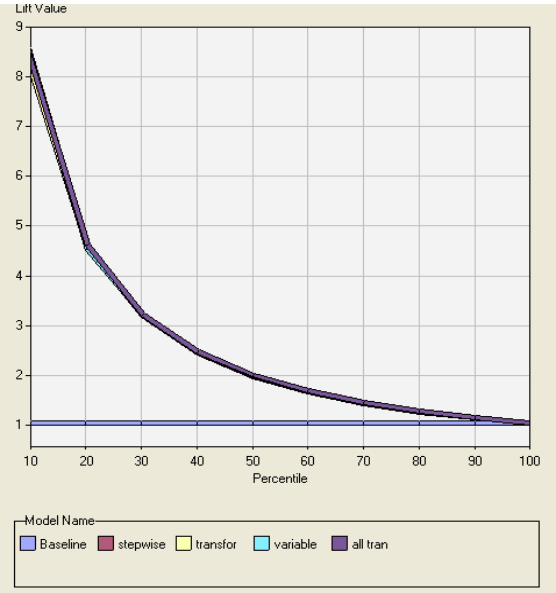

Figure 4-5 Cumulative lift charts that assess four optional models for Group II biomass-using facilities.

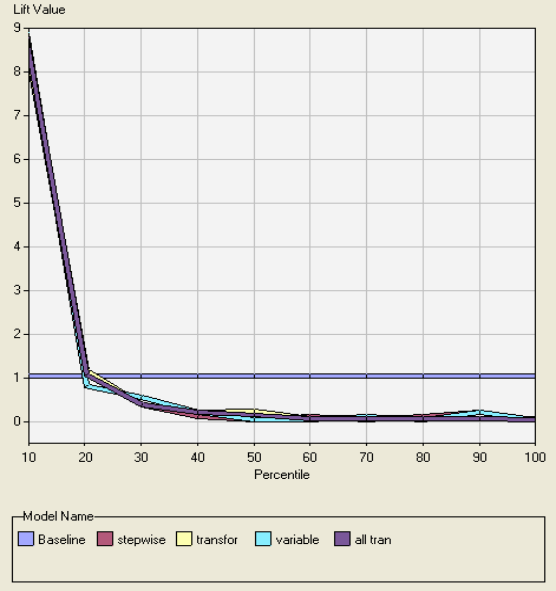

Figure 4-6 Non-cumulative lift charts that assess four optional models for Group II biomass-using facilities.

best model came from the first (i.e., "Stepwise only") model in Figure 3-3, which has the lowest BIC score, 365.90341332, and the lowest misclassification rate, 0.0265087422, as shown in Table 4-6. Recall the model-building steps for this model are data partition with stepwise logistic model with the BIC criterion. The analyses for the 13-state regional level and for the state level are performed with this model.

4.2.2 Predictive Ability Measured by Classification Tables

Classification tables in Table 4-7 and Figure 4-7 measure the predictive power of the best model. The results show that this model is preferred:

Sensitivity = $P(\hat{y} = 1 \mid y = 1) = 65.22\%$,

Specificity = $P(\hat{y} = 0 \mid y = 0) = 99.58\%$.

The overall proportion of correct classifications is

$$P(correct\ classification) = P(\hat{y} = 1 \mid y = 1)P(y = 1) + P(\hat{y} = 0 \mid y = 0)P(y = 0)$$

$$= 0.6522 * 0.0649 + 0.9958 * 0.9351$$

$$= 97.35\%$$

Table 4-7 Classification Table of predictive ability measurement for Group II biomass-using facilities.

Predictive value / **Actual value**	**0**	**1**	**Total**
0	1651 (99.58%)	7	1658 (93.51%)
1	40	75 (65.22%)	115 (6.49%)
Total	1691	82	1773

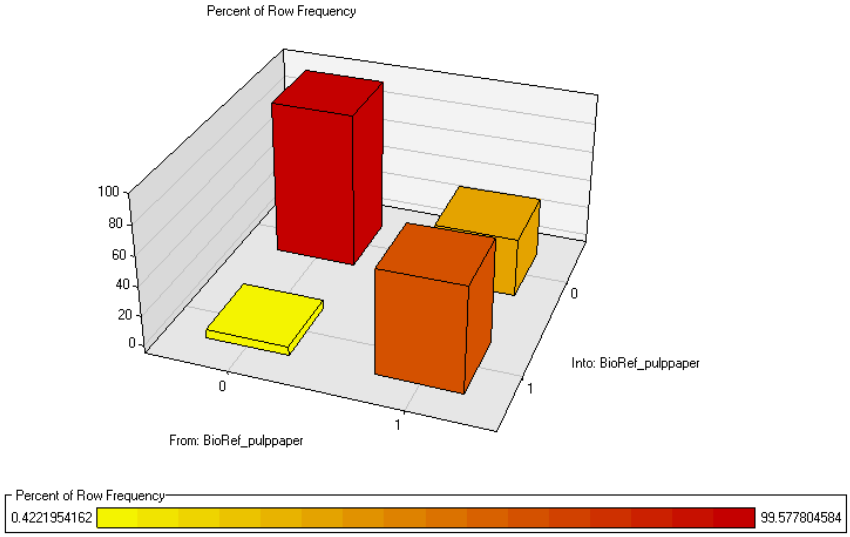

Figure 4-7 Predictive ability plot based on the classification table for Group II biomass-using facilities.

4.2.3 Regional Level Analysis Result for Group II Biomass-using Facilities

Using the best model found in the previous subsection reduces the thirty-one explanatory variables to the five variables that were statistically significant (p-values < 0.05). The Likelihood ratio goodness-of-fit test (Agresti 2007) shows that the model fits the data well. Significant variables include Population, thinnings within an 80-mile haul distance (Thin_80), harvesting cost of logging residues (Log_Res_Harvest_Cost), number of primary wood processing mills (Primary_mill_total), and number of secondary wood processing mills (Secondary_mill_total). Since there are no grouping variables in this model, Type 3 test is not performed. Maximum likelihood estimates are listed in Table 4-8.

Based on this regression, the best 25 locations for the Group II biomass-using facilities are presented in Figure 4-8. There are seven possible locations in Georgia, six in North Carolina,

Table 4-8 Analysis of maximum likelihood estimates of parameters for Group II biomass-using facilities.

Parameter	**DF**	**Estimate**	**Wald Chi_Square**	**Pr>ChiSq**
Intercept	1	-1.7817	42.12	0.07549
Population	1	-0.00011	40.26	<0.0001
THIN_80	1	0.00125	80.93	<0.0001
Log_Res_Harvest_Cost	1	-0.00035	35.53	<0.0001
Primary_mill_total	1	0.8492	11.95	0.0005
Secondary_mill_total	1	0.4710	22.06	<0.0001

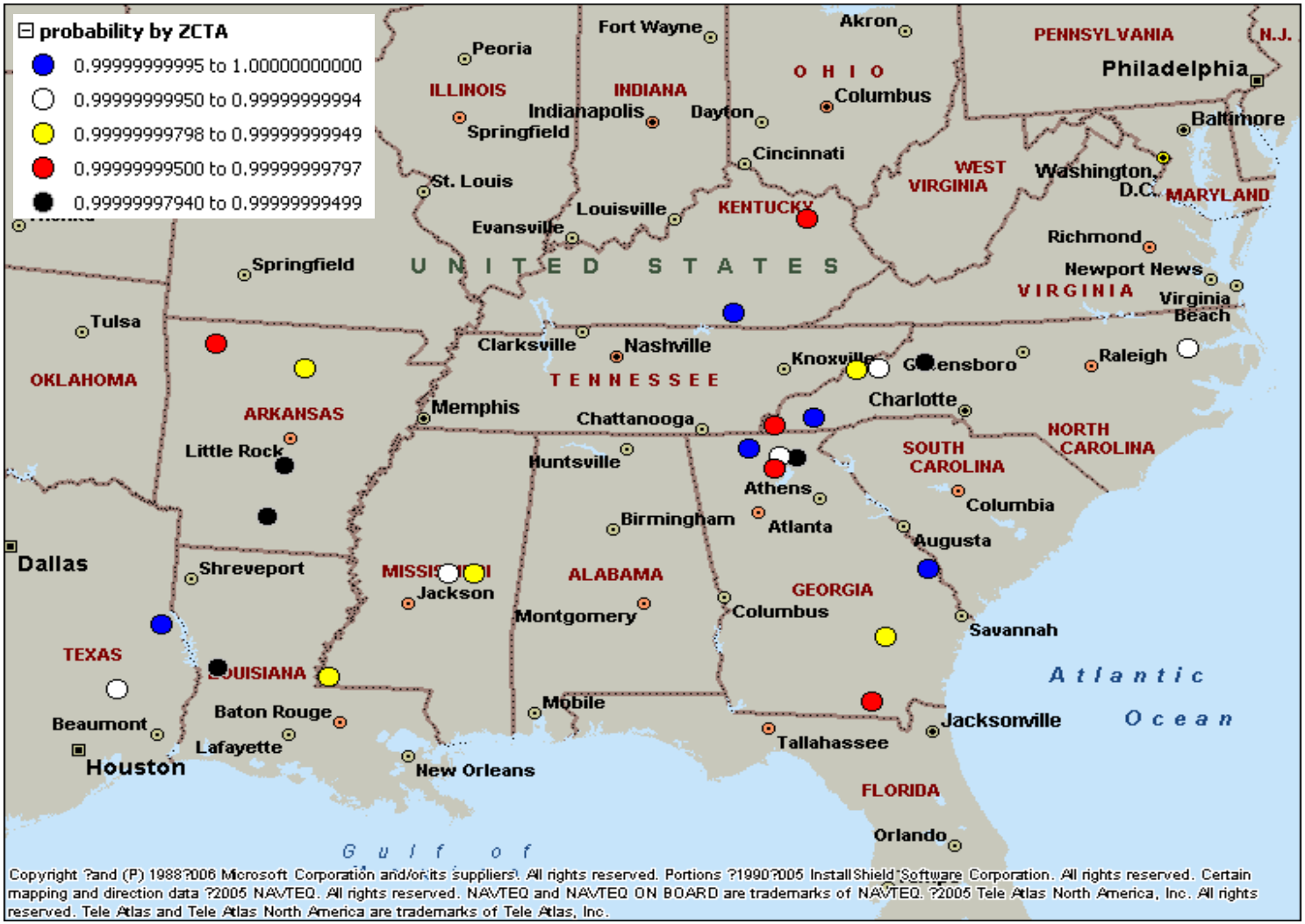

Figure 4-8 Top 25 optimal locations for Group II biomass-using facilities at the 13-state regional level.

four in Arkansas, three in Mississippi, and two in Kentucky and Texas, respectively.

4.2.4 State Level Analyses for the Group II Biomass-using Facilities

Separate state level analyses are performed to examine the statistically significant variables at the state level. The analysis in each state is summarized in Table 4-9. Primary wood processing mills and other removals of softwood are the only variables that are statistically significant (p-values < 0.05) in more than one state. Population is a positive significant variable instead of expected negative impact in Texas. South Carolina, Arkansas, and Mississippi have no significant variables. Oklahoma has no regression results because there are no data for the response variable in this state. All of the other significant variables that are significant at the 13-state regional level show positive or negative influences in at least one state. Variables of thinnings within different haul distances are significant at the state level in different states.

4.3 De-clustering Algorithm Application to Prediction Results

The logistic regression models are helpful in identifying the best 25 locations for both Group I and Group II biomass-using facilities. However, the practicality of the selected ZCTA locations may be questionable due to their proximity to existing wood-using mills that compete for the same resource. A “de-clustering algorithm” is developed as part of this thesis to avoid identifying ZCTAs for bioenergy and biofuels plants that have other biomass-using mills in the same ZCTA.

For each group, based on the existing mills, the maximum number of nearby primary, secondary, and pulp and paper mills in three different radius ranges are computed (e.g., 0-20 miles, 20-40 miles, and 40-80 miles). These maximum numbers of nearby mills are defined as

Table 4-9 State level analysis results for Group II biomass-using facilities.

State Name		LA	NC	TX	FL	KY	VA	GA	AL	TN	SC	AR	MS	OK
Primary_Mill_total	+	X	X											
OTHR_REM_SW	+				X	X								
Population	-			X										
RailroadAvailability 1	+				X									
Sqmiwater	+						X							
THIN_40	+							X						
THIN_80	+						X							
THIN_160	+								X					
THIN_120	+					X								
THIN_200	+			X										
Other_Mill_total	+								X					
LOG_RES_HW	+	X												
UNUSED_MILL_RES	+									X				

"tolerance numbers." The tolerance numbers are notated as $t_{20,prim}$, $t_{40,prim}$, $t_{80,prim}$, $t_{20,sec}$, $t_{40,sec}$, $t_{80,sec}$, $t_{20,pulp}$, $t_{40,pulp}$, and $t_{80,pulp}$, standing for tolerance numbers of primary mills in 0-20, 20-40, 40-80 miles, tolerance numbers of secondary mills in 0-20, 20-40, 40-80 miles, tolerance numbers of pulp and paper mills in 0-20, 20-40, 40-80 miles, respectively. For example, in Group II, the tolerance numbers are: $t_{20,prim} = t_{20,sec} = t_{20,pulp} = 0$, $t_{40,prim} = 75$, $t_{40,sec} = 516$, and $t_{40,pulp} = 19$. This is interpreted to mean that an existing mill in Group II cannot have any nearby mill within 0-20 miles and can have at most 75 primary mills, 516 secondary mills and 19 pulp and paper mills within 20-40 miles. Correspondingly, for each of the potential ZCTA locations, the numbers of nearby primary, secondary, and pulp and paper mills in the ranges of 0-20, 20-40, 40-80 miles are computed and defined as $n_{20,prim}$, $n_{40,prim}$, $n_{80,prim}$, $n_{20,sec}$, $n_{40,sec}$, $n_{80,sec}$, $n_{20,pulp}$, $n_{40,pulp}$, and $n_{80,pulp}$, respectively. For each group of mills, based on these tolerance numbers, and on the nearby numbers of each ZCTA, a penalty factor is constructed for each ZCTA as follows:

$$f_{all} = \exp\left(-c\left(\frac{n_{20,prim}}{t_{20,prim}} + \frac{n_{20,sec}}{t_{20,sec}} + \frac{n_{20,pulp}}{t_{20,pulp}} + \frac{n_{40,prim}}{t_{40,prim}} + \frac{n_{40,sec}}{t_{40,sec}} + \frac{n_{40,pulp}}{t_{40,pulp}} + \frac{n_{80,prim}}{t_{80,prim}} + \frac{n_{80,sec}}{t_{80,sec}} + \frac{n_{80,pulp}}{t_{80,pulp}}\right)\right), \quad (9)$$

$$f_{pulp} = \exp\left(-c\left(\frac{n_{20,prim}}{t_{20,prim}} \times 1_{n_{20,prim} > t_{20,prim}} + \frac{n_{20,sec}}{t_{20,sec}} \times 1_{n_{20,sec} > t_{20,sec}} + \frac{n_{20,pulp}}{t_{20,pulp}} + \frac{n_{40,prim}}{t_{40,prim}} \times 1_{n_{40,prim} > t_{40,prim}} + \frac{n_{40,sec}}{t_{40,sec}} \times 1_{n_{40,sec} > t_{40,sec}} + \frac{n_{40,pulp}}{t_{40,pulp}} + \frac{n_{80,prim}}{t_{80,prim}} \times 1_{n_{80,prim} > t_{80,prim}} + \frac{n_{80,sec}}{t_{80,sec}} \times 1_{n_{80,sec} > t_{80,sec}} + \frac{n_{80,pulp}}{t_{80,pulp}}\right)\right), \quad (10)$$

where $1_{a>b}$ is a sign function, equal to 1 if $a > b$ and 0 if $a \leq b$, and $c = \ln(2)$, i.e., a 0.5 penalty factor is generated when a nearby number equals the tolerance number. The penalty factor f_{all} is for Group I, and f_{pulp} is for Group II.

The difference between formulas (9) and (10) is that formula (10) for f_{pulp} uses sign functions. Recall that in Chapter 4, Subsection 4.2.3, for the Group II biomass-using facilities, we find that primary and secondary wood processing mills have positive significant effects on the site location of pulp and paper mills. Thus, if a nearby number of primary or secondary mills does not exceed the corresponding tolerance number, it would not contribute to the penalty factor f_{pulp}.

We multiply the penalty factor of a ZCTA and its predicted probability as a potential site location in order to obtain an adjusted predicted probability. This new adjusted predicted probability considers the competition between the future mills and the existing mills to try to keep the future mills away from the existing mills. The mapping of 25 "de-clustered" siting locations for the Group I biomass-using facilities is illustrated in Figure 4-9. Figure 4-10 displays a map of 25 "de-clustered" siting locations for the Group II biomass-using facilities.

For Group I, Tennessee has only one potential location after de-clustering compared with eight locations before de-clustering. Mississippi has four potential de-clustered locations compared with ten locations before de-clustering. Ten out of the 25 de-clustered locations are in Florida.

For Group II, only six locations remain after de-clustering. Before de-clustering, the potential locations are mainly in Georgia and North Carolina. After the de-clustering, Georgia

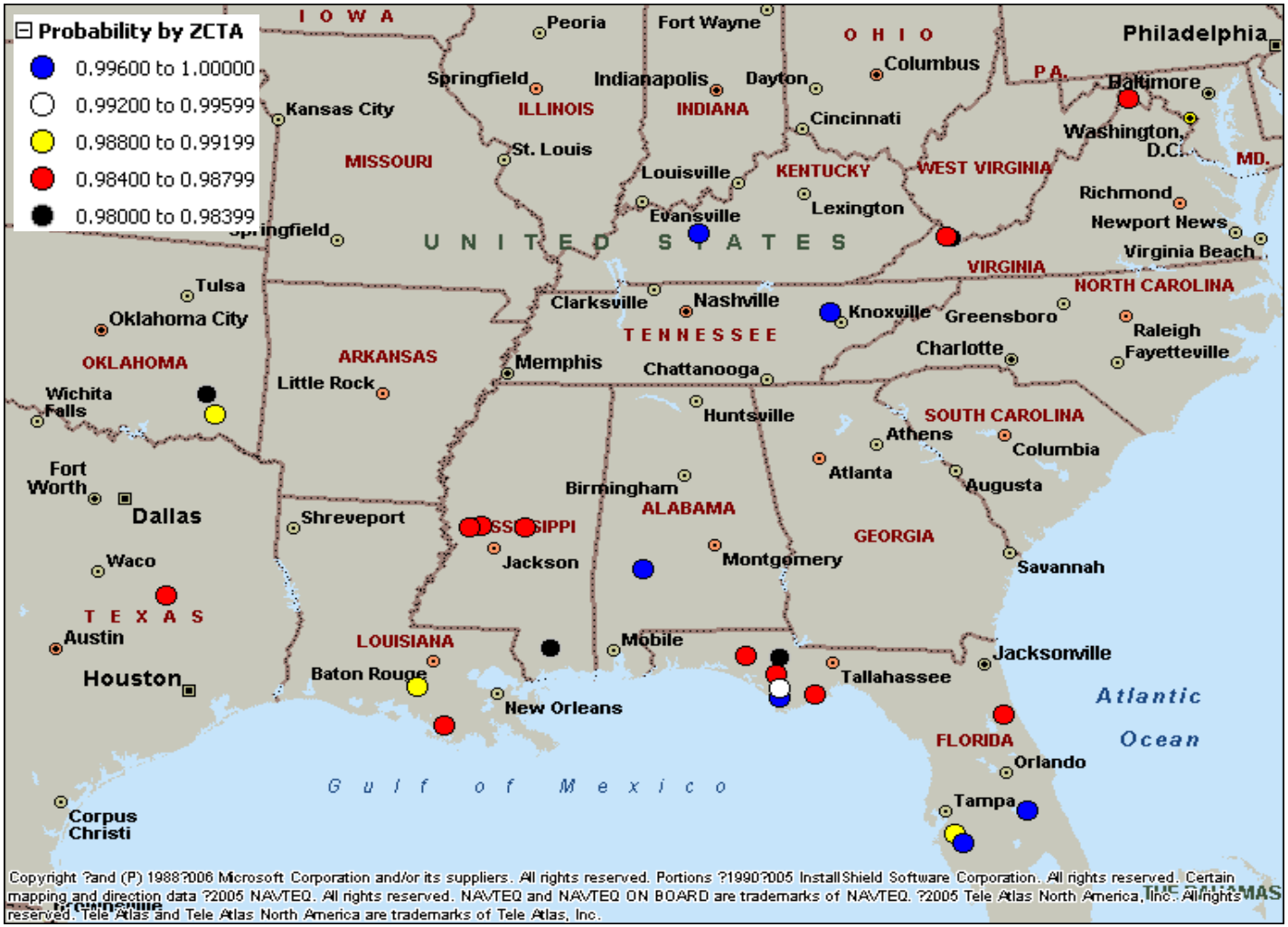

Figure 4-9 Top 25 optimal locations after de-clustering for Group I biomass-using facilities.

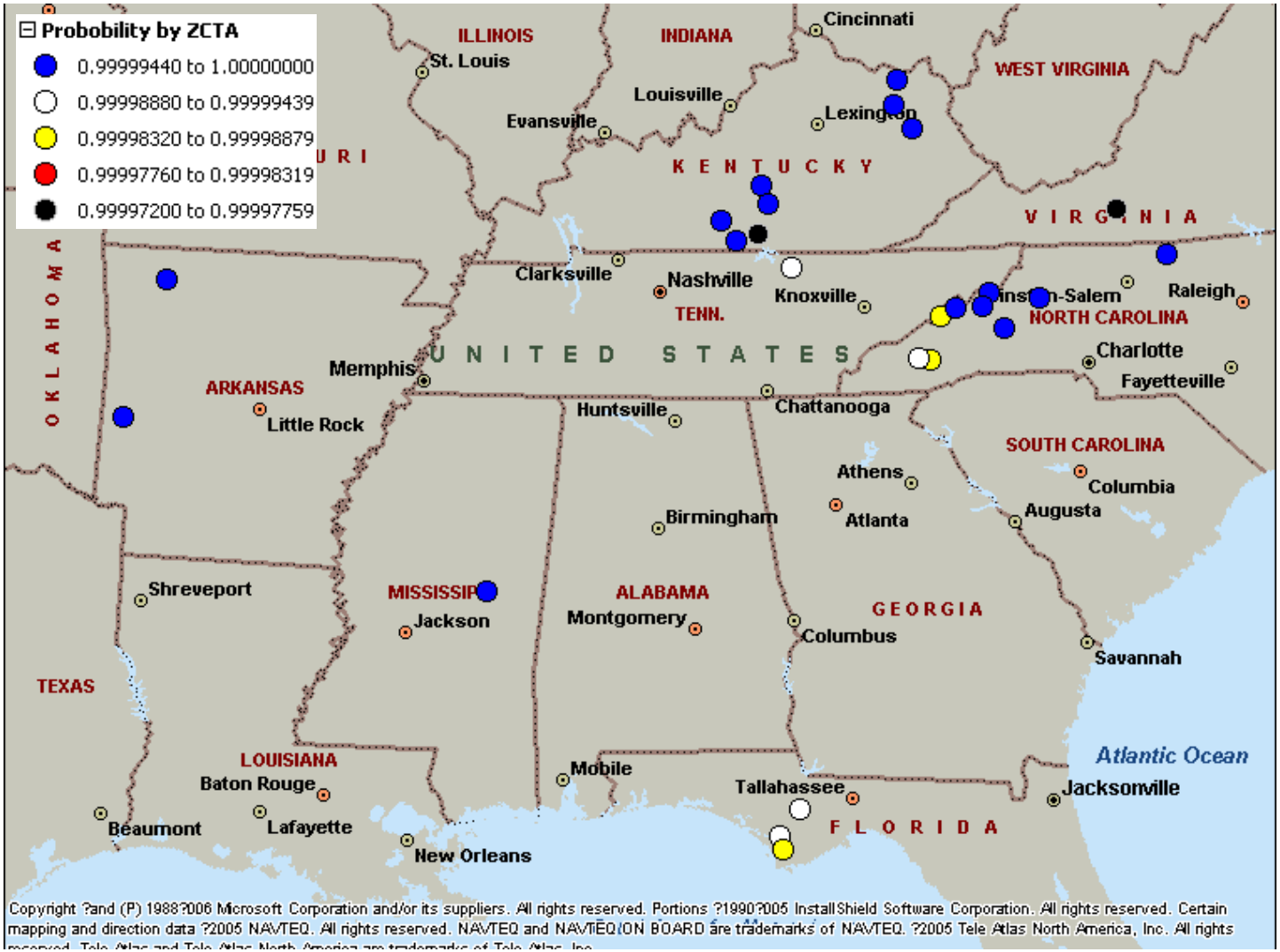

Figure 4-10 Top 25 optimal locations after de-clustering for Group II biomass-using facilities.

does not have a potential location. North Carolina gains potential locations, increasing from six to nine after de-clustering. Kentucky has eight potential locations after de-clustering.

Chapter 5 Summary

Two woody biomass-using facilities groups are studied in this thesis because the number of existing woody biomass-using bioenergy and biofuels plants is relatively small when compared with the large number of traditional woody biomass-using facilities. The analysis of Group I biomass-using facilities, which combined all woody biomass-using mills with wood-using bioenergy and biofuels plants, provides a modern planning view of total woody biomass management. Based on the research in this thesis, harvesting costs of logging residues and family income are statistically significant and have negative impacts on siting Group I biomass-using facilities in the 13-state region. Thinnings within an 80-mile haul distance, unused mill residues, and railroad availability are significant variables with positive impacts on siting Group I biomass-using facilities. In state level analyses, population is statistically significant and has a negative influence on siting locations in six of the 13 states (p-values ranged from <0.0001 to 0.0197) for Group I biomass-using facilities.

The analysis at the 13-state regional level for Group II biomass-using facilities, which combined pulp and paper mills with wood-using bioenergy and biofuels plants, provides statistical analysis results of the relationship between primary wood processing mills, secondary wood processing mills, and pulp and paper mills with bioenergy and biofuels plants. Primary wood processing mills and secondary wood processing mills are significant variables and have positive impacts in siting Group II biomass-using facilities. This observation reveals that the existing primary wood processing mills and secondary wood processing mills may compete with the future Group II biomass-using facilities, but they are still important feedstock providers (of feedstock such as wood chips) and may have a synergistic relationship with Group II biomass-

using facilities. Another positive variable is thinnings within an 80-mile haul distance. Population and the harvesting cost of logging residues are significant and have negative impacts on siting locations of Group II biomass-using facilities. In the state level analyses for Group II biomass-using facilities, no significant variable exists across the 13 states.

For both groups in the entire study region, statistically significant factors (p-value < 0.0001) in the logistic models are the harvesting cost of logging residues, which has a negative influence on siting decision, and the availability of thinnings within an 80-mile haul distance, which has a positive influence.

Twenty-five optimal locations (ZCTAs) are predicted and mapped at the 13-state regional level for each biomass-using facilities group. A de-clustering algorithm is also developed as part of this study to avoid competition between future mills and existing mills by keeping future mills away from existing mills.

In addition to the logistical model results, the database built for this study is an important outcome of the thesis. This database will not only benefit future research on this topic, but also support the public domain website www.BioSAT.net. The database currently has 14 types of biomass with real-time trucking cost models. Combining this research's data with the website database, such as including the economic information about population and employment for each ZCTA, could broaden the informational characteristics available about the ZCTAs and make the website more comprehensive.

Chapter 6 Future Research

The application of logistic models to the analysis of optimal location problems in woody biomass-using facilities is new. A brand new de-clustering algorithm is used to adjust the prediction results. This research refreshes people's view of forestry decision-making, but much more effort is needed to deepen and broaden the analysis. Generally, there are four parts to improve.

First, further research needs to collect more detailed data, such as soil type, rainfall, elevation, barge accessibility, building permits, and environmental regulations. As a high pollution industry, environmental policy may affect the site locations considerably. The more detailed the data set for analysis, the more useful the results.

Second, more analysis methods could be compared. This research focuses on the binary logistic regression model, but the ordinal logistic regression model may be more feasible if there is a way to define the response variable by something more than just zero and one. The other analysis methods that could be applied to this research are decision trees. One decision tree, named "Entropy Reduction," has been applied to this data on a trial basis and shows promising results. However, due to time limitations, a comprehensive and detailed discussion of decision trees applications are left for future research. There are several good decision tree methods possibly suitable for this analysis in addition to Entropy Reduction, such as Cruise and Guide.

Third, the de-clustering algorithm used to set potential locations away from existing competing mills in certain ZCTAs could also be capable of de-clustering potential locations themselves, avoiding the selection of potential locations that cluster to each other.

At last, this study schema is capable of being extended to all Eastern states, even across the whole country, because the data structure will be maintained across any number of states. The difficulty of extending this model lies in the increasing difficulty of collecting data as detailed as the five digit ZIP Code level or ZCTA level that appear in this thesis.

References

Abt, R. C., Cubbage, F. W., & Pacheco, G. (2000). Southern forest resource assessment using the Subregional Timber Supply (SRTS) Model. *Forest Products Journal*, 50(4), 25-33.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Bartuska, A. (2006). *Why biomass is important: The role of the USDA forest service in managing and using biomass for energy and other uses*. Washington, DC: Research & Development, USDA Forest Service. Retrieved from <http://purl.access.gpo.gov/GPO/LPS92802>.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227), 357-365.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *The Annals of Statistics*, 8(3), 457-487.

Berwick, M. & Farooq, M. (2003). *Truck costing model for transportation managers*. MPC report, no. 03-152. Fargo, ND: Upper Great Plains Transportation Institute, North Dakota State University.

Biomass Research and Development Board (2008). *The economics of biomass feedstocks in the United States: A review of the literature*. Occasional Paper No. 1. Retrieved from <http://www.brdisolutions.com>.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22(1), 134–167.

Brechbill, S. & Tyner, W. E. (2008). *The economics of renewable energy: Corn stover and switchgrass*. Retrieved from <http://www.ces.purdue.edu/extmedia/ID/ID-404.pdf>.

Burlington Northern Santa Fe Railway (2009). *Burlington Northern Santa Fe terminal list* [Data file]. Retrieved from <http://www.bnsf.com/bnsf.was6/refFilesStation/StationCentralController>.

Byrne, J., Shen, B., & Li, X. (1996). The challenge of sustainability: Balancing China's energy, economic and environmental goals. *Energy Policy*, 24(5), 455-462.

Caputo, J. (2009). *Sustainable forest biomass: Promoting renewable energy and forest stewardship* [PDF document]. Environmental and Energy Study Institute. Retrieved from <http://www.eesi.org/files/eesi_sustforbio_final_070609.pdf>.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society,* 20(2), 215-242.

Cox, D. R. (1966). Some procedures connected with the logistic qualitative response curve. In *Research Papers in Statistics: Festschrift for J. Neyman*. Ed. F. David. New York: Wiley. 55-71.

Cox, D. R. (1970). *Analysis of binary data*. London: Methuen.

Cramer, J. S. (2003). *The origins and development of the logit model*. Manuscript, University of Amsterdam and Tinbergen Institute. Retrieved from <http://www.cambridge.org/resources/0521815886/1208_default.pdf>.

CSX Corporation, Inc. (2005). *CSX state fact sheets* [Data file]. Retrieved from <http://www.csx.com/?fuseaction=about.state_facts>.

CSX Corporation, Inc. (2009). *CSX system map* [Graphic illustration of CSX railroad system]. Retrieved from <http://www.csx.com/share/media/media/docs/PrintableSystemMap-REF24028.pdf>.

C.T. Donovan Associates, Inc. & Lee, R. L. (1996). *Siting an ethanol plant in the Northeast.* Washington, DC: Northeast Regional Biomass Program.

Demirbas, A. (2005). Bioethanol from cellulosic materials: A renewable motor fuel from biomass. *Energy Sources*, 27, 327-337.

Finney, D. (1971). *Probit Analysis* (3rd ed.). Cambridge, U.K.: Cambridge University Press.

Forest2Market (2009). *Wood biomass energy: A Forest2Market research report*. Retrieved from <http://www.nafoalliance.org/LinkClick.aspx?fileticket=V8T%2F%2BRpgob0%3D&tabid=65&mid=510>.

Galik, C. S., Abt, R.C., & Wu, Y. (2009). Forest biomass supply in the Southeastern United States: Implications for industrial roundwood and bioenergy production. *Journal of Forestry*, 107(2), 69-77.

Gardner T. (2009). *Open and planned US cellulosic ethanol plants* [Data file]. Retrieved from <http://uk.reuters.com/article/oilRpt/idUKN1952406520090219>.

Gershenfeld, N. (1999). *The Nature of Mathematical Modeling*. Cambridge, U.K.: Cambridge University Press.

Government of Canada BioPortal Glossary (2009). *BioBasics* [Biomass definition]. Retrieved from <http://www.biobasics.gc.ca/english/View.asp?mid=411&x=696>.

Graham, R. L., English, B. C., & 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.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

IEA Bioenergy Task 39 (2009). *Status of 2nd generation biofuels demonstration facilities*. [Data file]. Retrieved from <http://biofuels.abc-energy.at/demoplants>.

Kaylen, M., Van Dyne, D.L., Choi, Y., & Blasé, M. (2000). Economic feasibility of producing ethanol from lignocellulosic feedstocks. *Bioresource Technology*, 72(1), 19-32.

Kingsland, S. E. (1985). *Modeling nature episodes in the history of population ecology*. Chicago: The University of Chicago Press.

Knut, S. L., Hallgeir, B., & Kjell, J. (2000). Siting of paper mills: Is a pristine environment an industrial resource? *Environmental Science Technology*, 34 (4), 546–551.

Kumar, A. & Sokhansanj, S. (2007). Switchgrass (Panicum vigratum, L.) delivery to a biorefinery using integrated biomass supply analysis and logistics (IBSAL) model. *Bioresource Technology*, 98(5), 1033-1044.

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Mapemba, L., Epplin, F., Taliaferro, C., & Huhnke, R. (2007). Biorefinery feedstock production on conservation reserve program land. *Review of Agricultural Economics*, 29(2), 227-246.

McFadden, D. (2001). Economic choices. *The American Economic Review*, 91(3), 351-378.

McQuarrie, A. D. R. & Tsai, C.-L. (1998). *Regression and time series model selection*. Singapore [u.a.]: World Scientific Publishing.

Milbrandt, A. (2005). *A geographic perspective on the current biomass resource availability in the United States*. Technical Report NREL/TP-560-39181. Golden, CO: National

Renewable Energy Laboratory. Retrieved from <http://www.nrel.gov/docs/fy06osti/39181.pdf>.

Moons, E., Saveyn, B., Proost, S., & Hermy, M. (2008). Optimal location of new forests in a suburban region. *Journal of Forest Economics*, 14(1), 5-27.

Moritz, H. (2000). Geodetic reference system 1980. *Journal of Geodesy*, 74(1), 128-162.

National Renewable Energy Laboratory (2009). *What is a biorefinery?* [Biorefinery definition]. Retrieved from <http://www.nrel.gov/biomass/biorefinery.html>.

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

Norfolk Southern System (2009). *Norfolk Southern system map* [Graphic illustration of Norfolk Southern railroad system]. Retrieved from <http://www.nscorp.com/nscportal/nscorp/pdf/systemmap2008.pdf>.

Ohlmacher, G. C. & Davis, J. C. (2003). Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas. *Engineering Geology*, 69(3-4), 331–343.

Parhizkar, O. & Smith, R. L. (2008). Application of GIS to estimate the availability of Virginia's biomass residues for bioenergy production. *Forest Products Journal*, 58(3), 71-76.

Patton-Mallory, M. (2008). *Woody biomass utilization strategy*. FS-899. Washington D.C.: U.S. Deptartment of Agriculture, Forest Service. Retrieved from <http://www.fs.fed.us/woodybiomass/strategy/documents/FS_WoodyBiomassStrategy.pdf>.

Pearl, R. & Reed, L. J. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation. *Proceedings from the National Academy of Sciences*, 6(6), 275-288.

Perez-Verdin, G., Grebner, D.L., Sun, C., Munn, I. A., Schultz, E. B., & Matney, T. G. (2009). Woody biomass availability for bioethanol conversion in Mississippi. *Biomass & Bioenergy*, 33(3), 492-503.

Perlack, R. D., Wright, L. L., Turhollow, A. F., Graham, R. L., Stokes, B. J., & Erbach, D. C. (2005). *Biomass as feedstock for a bioenergy and bioproducts industry: The technical feasibility of a billion-ton annual supply*. Oak Ridge, TN: Oak Ridge National Laboratory.

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

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

Ragavan, A. J. (2008). *How to use SAS to fit multiple logistic regression models* [PDF document]. SAS global forum 2008. Retrieved from <http://www2.sas.com/proceedings/forum2008/369-2008.pdf>.

Reed, L. J. & Berkson, J. (1929). The application of the logistic function to experimental data. *Journal of Physical Chemistry* 33(5), 760-779.

Renewable Fuels Association (2009). *Biorefinery locations* [Data file]. Retrieved from <http://www.ethanolrfa.org/industry/locations/>.

Retsina, T. & Pylkkanen, V. (2007). *AVAP™, a novel biorefinery concept*. TAPPI Web Exclusives. Atlanta, GA: American Process. Retrieved from <http://www.americanprocess.com/doc/AVAP_Paper.pdf>.

Scurlock, J. (2001). *Bioenergy Feedstock Characteristics*. Bioenergy Feedstock Information Network. Oak Ridge, TN: Oak Ridge National Laboratory. Retrieved from <http://bioenergy.ornl.gov/papers/misc/biochar_factsheet.html>.

Sedjo, R. A. (1997). The economics of forest-based biomass supply. *Energy Policy*, 25(6), 559-566.

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

Sperling, D. (1984). An analytical framework for siting and sizing biomass fuel plants. *Energy*, 9(11-12), 1033-1040.

Summit Ridge Investments, LLC (2007). *Eastern hardwood forest region woody biomass energy opportunity*. [S.l.]: Summit Ridge Investments, LLC.

The Biomass Research and Development Board (2008). *Increasing feedstock production for biofuels: Economic drivers, environmental implications, and the role of research*. Washington DC: Biomass Research & Development Initiative.

Union Pacific (2009). *UPRR system map* [Graphic illustration of the Union Pacific railroad system]. Retrieved from <http://www.uprr.com/aboutup/maps/sysmap.shtml>.

United Nations Development Programme (2000). *World energy assessment: Energy and the challenge of sustainability*. Ed. J. Goldemberg. Washington D.C.: UNDP/ UN-DESA/World Energy Council.

University of Wisconsin-Milwaukee Employment and Training Institute (2000). *Workforce and household/income data* [Data file]. Available from <http://www4.uwm.edu/eti/PurchasingPower/ETIshapefiles.htm>.

U.S. Army Corps of Engineers Navigation Data Center (2008). *U.S. waterway data* [Data file]. Retrieved from <http://www.ndc.iwr.usace.army.mil/data/datapwd.htm>.

U.S. Census Bureau (2000). *Population, land area, and water area data* [Data file]. Available from <http://www.census.gov/geo/www/gazetteer/places2k.html>.

U.S. Department of Agriculture Economic Research Service (2009). *Bioenergy* [Bioenergy definition]. Retrieved from <http://www.ers.usda.gov/features/bioenergy/>.

U.S. Department of Energy (2006). *Forest products industry technology roadmap* [PDF document]. Retrieved from <http://www.agenda2020.org/PDF/FPI_Roadmap%20Final_Aug2006.pdf>.

U.S. Department of Energy (2009). *Annual energy review 2008: Energy perspectives* [Web document]. Retrieved from <http://www.eia.doe.gov/emeu/aer/ep/ep_frame.html>.

Van den Broek, R., Vleeshouwers, R., Hoogwijk, M., van Wijk, A., & Turkenburg, W. (2001). The energy crop growth model SILVA: Description and application to eucalyptus plantations in Nicaragua. *Biomass & Bioenergy*, 21(5), 335-349.

Wang, Y. (2008). *Comparing linear discriminant analysis with classification trees using forest landowner survey data as a case study with considerations for optimal biorefinery siting*. Master Thesis, University of Tennessee, Knoxville, TN.

Western Governors' Association (2006). *Clean and diversified energy initiative-biomass*. Retrieved from <http://www.westgov.org/wga/initiatives/cdeac/Biomass-summary.pdf>.

Wilson, E. B. & Worcester, J. (1943). The determination of L. D. 50 and its sampling error in bio-assay. *Proceedings of the National Academy of Sciences*, 29(2), 79-85.

Wright, L. (2006). Worldwide commercial development of bioenergy with a focus on energy crop-based projects. *Biomass and Bioenergy*, 30(8-9), 706-714.

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, University of Tennessee, Knoxville, TN.

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

Young, T. M., Perdue, J. H., Hartsell, A., Abt, R. C., Hodges, D. G., & Rials, T. G. (2008). A real-time, web-based optimal Biomass Site Assessment Tool (BioSAT): Module 1 - An economic assessment of mill residues for the southern U.S. *Forest Inventory and Analysis (FIA) Symposium 2008; October 21-23, 2008; Park City, UT.* Retrieved from <http://www.fs.fed.us/rm/pubs/rmrs_p056/rmrs_p056_42_young.pdf>.

Yule, G. U. (1925). The growth of population and the factors which control it. *Journal of the Royal Statistical Society,* 138, 1-59.

Appendices

A-1 MATLAB codes for generating the neighboring ZIP Code list

```
function main(dist_lowerBnd, dist_upperBnd, showSphereDist)

% Ref: http://en.wikipedia.org/wiki/Earth_radius
% [zipset, ziploc] = zipInfo;
% Assume that all the latitude and longitude is expressed in RADIAN, and
% the data was sorted by ascending latitude and then by ascending
% longitude.

% dist_lowerBnd and dist_upperBnd are in miles.

if nargin == 2
    showSphereDist = 0;
end

load zipCode.mat zipInfo stateAbb stateCode
% zipInfo = zipInfo(1:5,:); % for testing only
zipset = zipInfo(:, 1);
ziploc = zipInfo(:, 3 : 4);
% zipwood = zipInfo(:, 5 : 7);
n = length(zipset);
for i = 1 : n
    for j = 1 : length(stateCode)
        if zipInfo(i, 2) == stateCode(j)
            zipstate(i, 1:2) = stateAbb(j, 1:2);
            break;
        end
    end
end
clear zipInfo stateAbb stateCode

dist_upperBnd = dist_upperBnd * 1.6093; %km
dist_lowerBnd = dist_lowerBnd * 1.6093;

ER = 6378.137; %Equatorial radius
PR = 6356.7523; %Polar radius
ER2 = ER*ER;
PR2 = PR*PR;
ERPR2 = (ER*PR)^2;
M0 = PR^2 / ER; %M(phi=0)
N0 = ER; %N(phi=0)
% Let phi = the latitude and lambda = the longitude
phi_threshold = dist_upperBnd / M0;
lambda_threshold = dist_upperBnd / (N0 * cos(pi*50/180));

ziploc = ziploc .* pi / 180;
fid = fopen('ziplist.txt','wt');
firsttime = 1;
for i = 1 : n
    cnt = 0;
    pntA = ziploc(i, :);
```

```
    for j = i + 1 : n
        pntB = ziploc(j, :);
%         if (withinSE13(zipstate(i, 1:2), zipstate(j, 1:2)))
%             continue;
%         end
        if (pntB(1) - pntA(1) <= phi_threshold)
            if abs(pntB(2) - pntA(2)) <= lambda_threshold
                lngDist = CalLngDist(pntA, pntB, ER2, PR2, ERPR2);
                latDist = CalLatDist(pntA, pntB, ER2, PR2);
                dist = sqrt(lngDist^2 + latDist^2);
                if (dist <= dist_upperBnd && dist > dist_lowerBnd)
                    cnt = cnt + 1;
                    if firsttime ~= 1
                        fprintf(fid,'\n');
                    end
                    firsttime = 0;
                    if zipset(i) < 1000
                        % the first two digits are "00" in this ZIP Code
                        fprintf(fid,'00%d,',zipset(i));
                    else if zipset(i) < 10000
                            % the first digit is "0" in this ZIP Code
                            fprintf(fid,'0%d,',zipset(i));
                        else
                            fprintf(fid,'%d,',zipset(i));
                        end
                    end
                    if zipset(j) < 1000
                        % the first two digits are "00" in this ZIP Code
                        fprintf(fid,'00%d',zipset(j));
                    else if zipset(j) < 10000
                            % the first digit is "0" in this ZIP Code
                            fprintf(fid,'0%d',zipset(j));
                        else
                            fprintf(fid,'%d',zipset(j));
                        end
                    end
                    if (showSphereDist ~= 0)
                        fprintf(fid,',%g',dist/1.6093);
                    end

                    fprintf(fid,'\n');
                    if zipset(j) < 1000
                        % the first two digits are "00" in this ZIP Code
                        fprintf(fid,'00%d,',zipset(j));
                    else if zipset(j) < 10000
                            % the first digit is "0" in this ZIP Code
                            fprintf(fid,'0%d,',zipset(j));
                        else
                            fprintf(fid,'%d,',zipset(j));
                        end
                    end
                    if zipset(i) < 1000
                        % the first two digits are "00" in this ZIP Code
                        fprintf(fid,'00%d',zipset(i));
```

```
                    else if zipset(i) < 10000
                            % the first digit is "0" in this ZIP Code
                            fprintf(fid,'0%d',zipset(i));
                        else
                            fprintf(fid,'%d',zipset(i));
                        end
                    end
                    if (showSphereDist ~= 0)
                        fprintf(fid,',%g',dist/1.6093);
                    end
                end
            end
        else
            break;
        end
    end
end
fclose(fid);


function lngDist = CalLngDist(pntA, pntB, ER2, PR2, ERPR2)
phi = (pntA(1) + pntB(1)) / 2;
dPhi = abs(pntB(1) - pntA(1));
M = ERPR2 / (ER2*cos(phi)^2 + PR2*sin(phi)^2)^1.5;
lngDist = M * dPhi;

function latDist = CalLatDist(pntA, pntB, ER2, PR2)
phi = (pntA(1) + pntB(1)) / 2;
dLambda = abs(pntB(2) - pntA(2));
N = ER2 / sqrt(ER2*cos(phi)^2 + PR2*sin(phi)^2);
latDist = N * cos(phi) * dLambda;

function res = withinSE13(state1, state2)
se13 = ['LA'; 'TX'; 'OK'; 'AR'; 'VA'; 'KY'; 'TN'; 'NC'; 'SC'; 'AL'; 'GA';
'MS'; 'FL'];

res = 0;
for i = 1 : 13
    if state1 == se13(i, 1:2)
        for j = 1 : 13
            if state2 == se13(j, 1:2)
                res = 1;
                break;
            end
        end

        break;
    end
end
return;
```

A-2 SAS codes for data collection and data management of the responses and explanatory variables

```
libname bi 'D:\Thesis\data and paper\data\Combination';

/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/
      /*** Data Management ***/
/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/
/* Import data sets from excel files */
/* census data */
proc import out=bi.census
      datafile= "D:\Thesis\data and paper\data\Combination\census data.xls"
      DBMS=EXCEL replace;
      sheet="ZIPHHOLD (2)$";
run;

/* import data of matching ZIP Codes with ZCTAs*/
proc import out= bi.match
      datafile= "D:\Thesis\data and paper\data\Combination\ZIP Code with
matched zcta.xls"
      DBMS=EXCEL replace;
run;

/* import harvest data */
proc import out= bi.harvest
      datafile= "D:\Thesis\data and paper\data\Combination\Harvest cost.xls"
      DBMS=EXCEL replace;
run;

/* import quantity data */
proc import out= bi.qty
      datafile= "D:\Thesis\data and paper\data\Combination\quantity by
zcta.xls"
      DBMS=EXCEL replace;
run;

/* import railroad data */
proc import out= bi.railroad
      datafile= "D:\Thesis\data and paper\data\Combination\railroad index by
ZIP Code .xls"
      DBMS=EXCEL replace;
run;

/*Price data management*/
proc import out=bi.price
      datafile= "D:\Thesis\data and paper\data\Combination\Price of 13 south
states_meanvalueadded.xls"
      DBMS=EXCEL replace;
run;

/* mill data */
proc import out= bi.mill
```

```
      datafile= "D:\Thesis\data and paper\data\Combination\mill location.xls"
      DBMS=EXCEL replace;
run;

/* large city list */
proc import out=bi.largecity
      datafile= "D:\Thesis\data and paper\data\Combination\large city.xls"
      DBMS=EXCEL replace;
run;

/* biofuel facility locations */
Proc import out=bi.biofuel
      datafile= 'D:\Thesis\data and paper\data\Combination\Biofuel
location.xls'
      DBMS=excel  replace;
run;

/* total quantity and average cost */
proc import out= BI.TOTALQTYAVGCOST
      datafile= "D:\Thesis\data and
paper\data\Combination\TotalQtyAvgCost.xls"
      DBMS=EXCEL replace;
      sheet="Sheet1$";
run;

/* marginal price for 0.5 million demand */
proc import out= BI.MC_p5M
      datafile= "D:\Thesis\data and paper\data\Combination\MC_p5M1M1p5M.xls"
      DBMS=EXCEL replace;
      sheet="p5M$";
run;

/* marginal price for 1 million demand */
proc import out= BI.MC_1M
      datafile= "D:\Thesis\data and paper\data\Combination\MC_p5M1M1p5M.xls"
      DBMS=EXCEL replace;
      sheet="1M$";
run;

/* marginal price for 1.5 million demand */
proc import out= BI.MC_1p5M
      datafile= "D:\Thesis\data and paper\data\Combination\MC_p5M1M1p5M.xls"
      DBMS=EXCEL replace;
      sheet="1p5M$";
run;

/* state abbreviation list*/
proc import out= BI.StateAbb
      datafile= "D:\Thesis\data and paper\data\Combination\State Name
Abbreviation.xls"
      DBMS=EXCEL replace;
      sheet="Sheet1$";
run;
```

```
/** Categoried new mill location data management **/
proc import out= BI.newmills_category
      datafile= "D:\Thesis\data and
paper\data\Combination\newmills_category.xls"
      DBMS=EXCEL replace;
run;
data bi.newmills_category;

      set BI.newmills_category;
      format  Abb_State_Name $2.;
      format   City_Name $20.;
      format Zipcode $5.;
      informat Abb_State_Name $2.;
      informat City_Name $20.;
      informat Zipcode $5.;
run;
data bi.south13newmills_category;
      set bi.newmills_category;
      where Abb_State_Name in ('FL','SC', 'TN',
'TX','VA','AL','AR','GA','KY','LA','MS','NC','OK' );
      drop Category  Business   Coded_mill_type;
run;
/* remove the duplicated mill locations */
proc sort data=bi.south13newmills_category nodup
      out=bi.south13newmills_categorysort;
      by Company Zipcode Business_Category;
run;
/* Unique recodes are created, but the same company in the same ZIP Code may
be in different categories */
proc sort data=bi.south13newmills_categorysort nodupkey
      out=bi.south13newmills_clear;
      by Address Zipcode Business_Category;
run;
/* format the data*/
data BI.south13newmills_clear;
set BI.south13newmills_clear;
      format  Abb_State_Name $2.;
      format   City_Name $20.;
      format Zipcode $5.;
      format Business_category $32.;
      informat Abb_State_Name $2.;
      informat City_Name $20.;
      informat Zipcode $5.;
      informat Business_category $32.;
run;

/* mill locations with category codes */
proc import out=bi.mills_categorycode
      datafile= "D:\Thesis\data and
paper\data\Combination\newmills_category_code.xls"
      DBMS=EXCEL replace;
run;
proc sort data=bi.mills_categorycode;
by Business_category;
```

```
run;
data bi.codedmills;
      merge bi.south13newmills_clear
      bi.mills_categorycode;
      by Business_category;
run;

/* find mills with multiple catogories */
proc import out=bi.codemillwithsameloccat
      datafile= "D:\Thesis\data and paper\data\Combination\131 companies in
same location having different mill categories.xls"
      DBMS=EXCEL replace;
run;
proc sort data=bi.codemillwithsameloccat;
by Company Address Zipcode;
run;
/*131 data items have same location with different categories. They were
cleared out and given a single category*/
data bi.millmess;
      merge bi.codedmills
      bi.codemillwithsameloccat(in=insameloccat);
      by Company address Zipcode;
      if insameloccat;
run;
proc sort data=bi.millmess nodupkey;
      by  Company address Zipcode Coded_mill_type;
run;
data bi.millmess131;
      set bi.millmess;
      retain Milltype_coded;
      by Company ;
      if First.Company then do;
            Milltype_coded=Coded_mill_type;
      end;
      cnt + 1;
      if Coded_mill_type<Milltype_coded then Milltype_coded=Coded_mill_type;
run;
proc sort data=bi.millmess131 nodupkey;
      by Company Address Zipcode Milltype_coded;
run;

/* mill locations with an unique record */
proc import out=bi.codedmill6168clear
      datafile= "D:\Thesis\data and paper\data\Combination\6168 clear mill
locations-no replication in same or different locations.XLS"
      DBMS=EXCEL replace;
run;
data bi.codedmill6168clearwithcode;
      merge bi.codedmill6168clear(in=in6168)
      bi.codedmills;
      if in6168;
run;

/* merge the above two sets as the final new mill category data set*/
```

```
data bi.finalmillcoded(drop=Company Address City_name Abb_State_Name
Business_Category );
      set bi.codedmill6168clearwithcode(drop=Coded_mill_type )
      bi.millmess131(drop=Milltype_coded Coded_mill_type cnt);
run;

/* export this data set to an excel file for the further JMP operations*/
proc EXPORT DATA=bi.finalmillcoded
      OUTFILE="D:\Thesis\data and paper\data\from Andrea\finalmillcoded.xls"
      DBMS=EXCEL replace;
run;
/* for this data set, the JMP summary operation is used to create the next
imported excel data set */

/* coded mills at ZIP Code level */
proc import out=bi.codedmillszipcode
      datafile= "D:\Thesis\data and paper\data\Combination\Final version of
coded 2812 mills in zipcode level with transpose into 4 kinds of mills.xls"
      DBMS=EXCEL replace;
run;
proc sort data=bi.codedmillszipcode;
by Zipcode;
run;
/** finish new mill data category mangement **/

/* waterway data */
proc import out= bi.ports
      datafile= "D:\Thesis\data and paper\data\Combination\Waterway_ZCTA.xls"
      DBMS=EXCEL replace;
run;
proc sort data=bi.ports nodupkey;
by Latitude Longtitude ;
run;
proc sql;
create table bi.numberports as
select ZCTA, Sum(Ports) as Numberports
from bi.ports
group by ZCTA;
quit;
/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/
       /*** finished all the data management works ***/
/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/

/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/
/*** Start to merge the data sets ***/
/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/

/*Merge all data sets in the level of Zipcode by Zipcode, and also matching
the ZIP Codes with the corresponding ZCTAs*/
/* sort data sets for merging */
proc sort data=bi.railroad
      out=bi.s_railroad;
      by Zipcode;
run;
```

```
proc sort data=bi.match
      out=bi.s_match;
      by Zipcode;
run;
proc sort data=bi.totalqtyavgcost
      out=bi.s_totalqtyavgcost;
      by Zipcode;
run;
proc sort data=bi.mc_p5M
      out=bi.s_mc_p5M;
      by Zipcode;
run;
proc sort data=bi.mc_1M
      out=bi.s_mc_1M;
      by Zipcode;
run;
proc sort data=bi.mc_1p5M
      out=bi.s_mc_1p5M;
      by Zipcode;
run;
Proc sort data=bi.clearMills;
      by Zipcode;
run;
proc sort data=bi.biofuel nodup
      out=bi.s_biofuel;
      by Zipcode;
run;
/* merge them */
data bi.RailroadMCWithMissingValue; /*the data set name does not imply all
data in it*/
      merge bi.s_railroad (drop = state_name City IN = inRailroad)
      bi.s_match (IN = inMatch drop = State_FIPS County_FIPS  Latitude
Longitude)
      bi.s_totalqtyavgcost (IN = inTtlQtyAvgCst)
      bi.codedmillszipcode
      bi.s_biofuel(in=inbiofuel)
      bi.s_mc_p5M
      bi.s_mc_1M
      bi.s_mc_1p5M;
      by Zipcode;
      if inMatch ;
run;
/* surrogate the missing values */
data bi.RailroadMCByZipcode;
      set bi.RailroadMCWithMissingValue;
      if Railroad_Availability = . then do;
            CSX = 0;
            NS = 0;
            BNSF = 0;
            UP = 0;
            Railroad_Availability = 0;
      end;
      if CSX = . then CSX = 0;
      if NS = . then NS = 0;
```

```
      if BNSF = . then BNSF = 0;
      if UP = . then UP = 0;
      if CSX > 0 then CSX = 1;
      if CumTC_80 = . then CumTC_80 = 99999999;
      if CumQty_80 = . then CumQty_80 = 0;
      if AC_80 = . then AC_80 = 99999999;
      if MC_p5M = . then MC_p5M = 9999;
      if MC_1M = . then MC_1M = 9999;
      if MC_1p5M = . then MC_1p5M = 9999;
      if Milltotal=. then Milltotal=0;
      if Primary_mill=. then Primary_mill=0;
      if Secondary_mill=. then Secondary_mill=0;
      if Pulp_and_paper_mill=. then Pulp_and_paper_mill=0;
      if Other_Mill=. then Other_Mill=0;
      if Bioref=. then Bioref=0;
run;

/* convert the above data from the ZIP Code level to the ZCTA level */
proc sort data = bi.RailroadMCByZipcode;
      by ZCTA;
run;
/* combining the information in the same ZCTA */
data bi.RailroadMC (keep = ZCTA City State_Name County_Name CID
                           RailroadAvailability TCost_80 TQty_80 ACost_80
MCost_p5M MCost_1M
                           MCost_1p5M Mills_total Primary_mill_total
                           Secondary_mill_total Pulp_and_paper_mill_total
Other_Mill_total
                           Biorefineries );
      set bi.RailroadMCByZipcode;
      format RailroadAvailability $4.;
      informat RailroadAvailability $4.;
      retain Railroad_CSX Railroad_NS Railroad_BNSF Railroad_UP MCost_p5M
MCost_1M MCost_1p5M
      mills_total Primary_mill_total Secondary_mill_total
Pulp_and_paper_mill_total
      Other_Mill_total Biorefineries;
      by ZCTA;
      if First.ZCTA then do;
            Railroad_CSX = CSX;
            Railroad_NS = NS;
            Railroad_BNSF = BNSF;
            Railroad_UP = UP;
            MCost_p5M = MC_p5M;
            MCost_1M = MC_1M;
            MCost_1p5M = MC_1p5M;
            cnt = 0;
            TCost_80 = 0;
            TQty_80 = 0;
            ACost_80 = 0;
            Mills_total=0;
            Primary_mill_total=0;
            Secondary_mill_total=0;
            Pulp_and_paper_mill_total=0;
```

```
            Other_Mill_total=0;
            Biorefineries=0;
      end;
      if CSX > Railroad_CSX then Railroad_CSX = CSX;
      if NS > Railroad_NS then Railroad_NS = NS;
      if BNSF > Railroad_BNSF then Railroad_BNSF = BNSF;
      if UP > Railroad_UP then Railroad_UP = UP;
      if MC_p5M < MCost_p5M then MCost_p5M = MC_p5M;
      if MC_1M < MCost_1M then MCost_1M = MC_1M;
      if MC_1p5M < MCost_1p5M then MCost_1p5M = MC_1p5M;
      TCost_80 + CumTC_80;
      TQty_80 + CumQty_80;
      ACost_80 + AC_80;
      /*Ports_total+NumberPorts;*/
      Mills_total+Milltotal;
      Primary_mill_total+Primary_mill;
      Secondary_mill_total+Secondary_mill;
      Pulp_and_paper_mill_total+Pulp_and_paper_mill;
      Other_Mill_total+Other_Mill;
      Biorefineries+Bioref;
      /*T_numbermills+numbermills*/
      /*numbermills were substituted by T_numbermills*/;
      cnt + 1;
      if Last.ZCTA then do;
            TCost_80 = TCost_80 / cnt;
            TQty_80 = TQty_80 / cnt;
            ACost_80 = ACost_80 / cnt;
            Railroad_Availability = Railroad_CSX + Railroad_NS +
Railroad_BNSF + Railroad_UP;
            if Railroad_Availability = 0 then RailroadAvailability = 'N/A';
            else RailroadAvailability = put(Railroad_Availability, $4.);
            output bi.RailroadMC;
      end;
run;

/*Merge RailroadMC, census, quantity and waterway data by ZCTA*/
proc sort data = bi.RailroadMC;
      by ZCTA;
run;
proc sort data=bi.qty out=bi.s_qty;
      by ZCTA;
run;
proc sort data=bi.census out=bi.s_census;
      by ZCTA;
run;
proc sort data=bi.numberports
      out=bi.s_numberports;
      by ZCTA ;
run;
/* merge by ZCTA */
data bi.CensusRailroadMCQtyMillMissing;
      merge bi.RailroadMC (in = inRailroadMC)
      bi.s_census (in = inCensus drop = Zipcode)
      bi.s_qty (in = inQty)
```

```
      bi.s_numberports;
      by ZCTA;
      if inRailroadMC and inCensus;
run;
/* surrogate missing values */
data bi.CensusRailroadMCQtyMill;
      set bi.CensusRailroadMCQtyMillMissing;
      if DRY_BIO_HW = . then DRY_BIO_HW = 0;
      if DRY_BIO_SW = . then DRY_BIO_SW = 0;
      if DRY_BIO_TOT = . then DRY_BIO_TOT = 0;
      if LOG_RES_HW = . then LOG_RES_HW = 0;
      if LOG_RES_SW = . then LOG_RES_SW = 0;
      if LOG_RES_TOT = . then LOG_RES_TOT = 0;
      if OTHR_REM_HW = . then OTHR_REM_HW = 0;
      if OTHR_REM_SW = . then OTHR_REM_SW = 0;
      if OTHR_REM_TOT = . then OTHR_REM_TOT = 0;
      if THIN_40 = . then THIN_40 = 0;
      if THIN_80 = . then THIN_80 = 0;
      if THIN_120 = . then THIN_120 = 0;
      if THIN_160 = . then THIN_160 = 0;
      if THIN_200 = . then THIN_200 = 0;
      if TOTAL_MILL_RES = . then TOTAL_MILL_RES = 0;
      if UNUSED_MILL_RES = . then UNUSED_MILL_RES = 0;
      if URBAN_WASTE = . then URBAN_WASTE = 0;
      if Numberports=. then Numberports=0;
run;

/*Merge CensusRailroadMCQtyMill and price data by State_Name*/
proc sort data = bi.CensusRailroadMCQtyMill;
      by State_Name;
run;
proc sort data=bi.price out=bi.s_price; /*NOTE: only 33 states are in the
price data set*/
      by State_Name;
run;
proc sort data=bi.StateAbb;
      by State_Name;
run;
/* merge */
data bi.CensusRailroadMCQtyMillPrice bi.notIn33States;
      merge bi.CensusRailroadMCQtyMill (in = inCensorRailroadMCQtyMill)
      bi.s_price (in = inPrice )
      bi.StateAbb(keep = State_Name Abb_State_Name);
      by State_Name;
      if inCensorRailroadMCQtyMill and inPrice then output
bi.CensusRailroadMCQtyMillPrice;
      else if inCensorRailroadMCQtyMill then output bi.notIn33States;
run;

/* merge the harvest data in by CID*/
proc sort data = bi.harvest
      out = bi.s_harvest;
      by CID;
run;
```

```
proc sort data = bi.CensusRailroadMCQtyMillPrice;
      by CID;
run;
/* merge */
data  bi.CensusRailroadMCQtyMillPriHvest;
      merge bi.CensusRailroadMCQtyMillPrice
(in=inCensusRailroadMCQtyMillPriBio) bi.s_harvest (in = inHarvest);
      by CID;
      if inCensusRailroadMCQtyMillPriBio;
run;

/* merge the large city data in by state and city names*/
proc sort data= bi.largecity
      out=bi.s_largecity;
      by  Abb_State_Name City;
run;
proc sort data=bi.CensusRailroadMCQtyMillPriHvest;
      by Abb_State_Name City;
run;
data bi.South33statesAllDatamissing;
      merge
bi.CensusRailroadMCQtyMillPriHvest(in=inCensusRailMCQtyMillPriHvest)
bi.s_largecity;
      by Abb_State_Name City;
      if inCensusRailMCQtyMillPriHvest;
run;

/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/
     /*  End of merging data  */
/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/

/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/
/*Definition of the response variable for Group 1:  MILLS-TOTAL*/
/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/

/* surrogate missing values */
data bi.South33StatesAllData_totmill;
      set bi.South33statesAllDatamissing;
      if Metropolitan=. then Metropolitan=0;
      if Log_Res_Harvest_Cost = . then Log_Res_Harvest_Cost =9999;
      if DRY_BIO_TOT=0 then BioRef_totmill=0;
      if Sqmiland=0 then BioRef_totmill=0;
      if Metropolitan=1 then BioRef_totmill = 0;
      if Mills_total  >= 1 or Biorefineries>=1 then BioRef_totmill = 1;
      if BioRef_totmill=0 or BioRef_totmill=1 then datapartation=1;
      if BioRef_totmill=. then datapartation=0;
run;
/* NOTE: The data set BI.SOUTH33STATESALLDATA has 22179 observations and 79
variables. */

/* all data in the Southeastern 13 states */
data bi.South13StatesAllData_totmill(drop=State_Name);
      set bi.South33StatesAllData_totmill(drop=CID City County_Name
Abb_State_Name City
```

```
                                                    SW_Clean_Mill_Res_Price
HW_Clean_Mill_Res_Price SW_and_HW_clean_Comb_price
                                                    SW_UnClean_Mill_res_Price
HW_UnClean_Mill_res_Price SW_and_HW_Comb_UnClean_Price
                                                    Total_Mill_res_Price
Unused_Mill_res_Price SW_Dry_Bio_Price HW_Dry_Bio_Price
                                                    SW_and_HW_Dry_Bio_Comb_Price
SW_Log_Res_Price HW_Log_Res_Price
                                                    Comb_Log_Res_Price
Pulp_SW_Price Pulp_HW_Price Sawtimber_SW_Price
                                                    Sawtimber_HW_Price
Comb_Sawtimber_Price    Pulp_Growth_SW_Price Pulp_Growth_HW_Price
                                                    Comb_Pulp_Growth_Price
Sawtimber_Growth_SW_Price Sawtimber_Growth_HW_Price
                                                    Comb_Sawtimber_Growth_Price
Urban_Waste_Price SW_Other_Removals_Price
                                                    HW_Other_Removals_Price
Comb_Other_Rem_Price Thinnings_Price Comb_Pulp_Price
                                                    Mills_total
Primary_mill_total Secondary_mill_total Pulp_and_paper_mill_total
Other_Mill_total
                                                    Biorefineries Metropolitan
DRY_BIO_HW DRY_BIO_SW DRY_BIO_TOT Sqmiland Population_Density );
      where State_Name in ('Tennessee','Florida','Alabama', 'Louisiana',
'Texas','Oklahoma','Arkansas',
                                                    'Virginia','Kentucky','North
Carolina', 'South Carolina', 'Georgia','Mississippi' );
run;

proc sql;
      /* the existing and non-probable mill locations*/
      create table bi.South13StatesAllData_totmill01 as
      select * from bi.South13StatesAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      /* the potential mill locations */
      create table bi.South13StatesAllData_totmillmis as
      select * from bi.South13StatesAllData_totmill
      where BioRef_totmill =.;
quit;

/* For each of the 13 states, three data sets are generated: */
/* all ZCTA loations, the existing and non-probable mill locations, and the
potential mill locations*/
data bi.TennesseeAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Tennessee';
run;
proc sql;
      create table bi.TennesseeAllData_totmill01 as
      select * from bi.TennesseeAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.TennesseeAllData_totmillmis as
```

```
      select * from bi.TennesseeAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.TexasAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Texas';
run;
proc sql;
      create table bi.TexasAllData_totmill01 as
      select * from bi.TexasAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.TexasAllData_totmillmis as
      select * from bi.TexasAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.FloridaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Florida';
run;
proc sql;
      create table bi.FloridaAllData_totmill01 as
      select * from bi.FloridaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.FloridaAllData_totmillmis as
      select * from bi.FloridaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.AlabamaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Alabama';
run;
proc sql;
      create table bi.AlabamaAllData_totmill01 as
      select * from bi.AlabamaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.AlabamaAllData_totmillmis as
      select * from bi.AlabamaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.LouisianaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Louisiana';
run;
proc sql;
      create table bi.LouisianaAllData_totmill01 as
      select * from bi.LouisianaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;
```

```
      create table bi.LouisianaAllData_totmillmis as
      select * from bi.LouisianaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.OklahomaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Oklahoma';
run;
proc sql;
      create table bi.OklahomaAllData_totmill01 as
      select * from bi.OklahomaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.OklahomaAllData_totmillmis as
      select * from bi.OklahomaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.ArkansasAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Arkansas';
run;
proc sql;
      create table bi.ArkansasAllData_totmill01 as
      select * from bi.ArkansasAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.ArkansasAllData_totmillmis as
      select * from bi.ArkansasAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.VirginiaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Virginia';
run;
proc sql;
      create table bi.VirginiaAllData_totmill01 as
      select * from bi.VirginiaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.VirginiaAllData_totmillmis as
      select * from bi.VirginiaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.KentuckyAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Kentucky';
run;
proc sql;
      create table bi.KentuckyAllData_totmill01 as
```

```
      select * from bi.KentuckyAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.KentuckyAllData_totmillmis as
      select * from bi.KentuckyAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.NorthCarolinaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'North Carolina';
run;
proc sql;
      create table bi.NorthCarolinaAllData_totmill01 as
      select * from bi.NorthCarolinaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.NorthCarolinaAllData_totmillmis as
      select * from bi.NorthCarolinaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.SouthCarolinaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'South Carolina';
run;
proc sql;
      create table bi.SouthCarolinaAllData_totmill01 as
      select * from bi.SouthCarolinaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.SouthCarolinaAllData_totmillmis as
      select * from bi.SouthCarolinaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.GeorgiaAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Georgia';
run;
proc sql;
      create table bi.GeorgiaAllData_totmill01 as
      select * from bi.GeorgiaAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.GeorgiaAllData_totmillmis as
      select * from bi.GeorgiaAllData_totmill
      where BioRef_totmill =.;
quit;

data bi.MississippiAllData_totmill;
      set bi.South13StatesAllData_totmill;
      where State_Name = 'Mississippi';
run;
```

```
proc sql;
      create table bi.MississippiAllData_totmill01 as
      select * from bi.MississippiAllData_totmill
      where BioRef_totmill = 0 or BioRef_totmill = 1;

      create table bi.MississippiAllData_totmillmis as
      select * from bi.MississippiAllData_totmill
      where BioRef_totmill =.;
quit;

/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/
/*Definition of the response variable for Group 2:  PULP AND PAPER MILL*/
/*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*/

/* surrogate missing values */
data bi.South33States_pulppaper;
set bi.South33statesAllDatamissing;
if Metropolitan=. then Metropolitan=0;
if Log_Res_Harvest_Cost = . then Log_Res_Harvest_Cost =9999;
/*Create the score of BioRefScore*/
if DRY_BIO_TOT=0 then BioRef_pulppaper=0;
if Sqmiland=0 then BioRef_pulppaper=0;
if Metropolitan=1 then BioRef_pulppaper = 0;
if Pulp_and_paper_mill_total  >= 1 or Biorefineries>=1 then BioRef_pulppaper
= 1;
if BioRef_pulppaper=0 or BioRef_pulppaper=1 then datapartation=1;
if BioRef_pulppaper=. then datapartation=0;
run;
/* NOTE: The data set BI.SOUTH33STATESALLDATA has 22179 observations and 79
variables. */

/* all data in the Southeastern 13 states */
data bi.South13States_pulppaper;
set bi.South33States_pulppaper(drop=CID City County_Name  Abb_State_Name City
                                                    SW_Clean_Mill_Res_Price
HW_Clean_Mill_Res_Price SW_and_HW_clean_Comb_price
                                                    SW_UnClean_Mill_res_Price
HW_UnClean_Mill_res_Price SW_and_HW_Comb_UnClean_Price
                                                    Total_Mill_res_Price
Unused_Mill_res_Price SW_Dry_Bio_Price HW_Dry_Bio_Price
                                                    SW_and_HW_Dry_Bio_Comb_Price
SW_Log_Res_Price HW_Log_Res_Price
                                                    Comb_Log_Res_Price
Pulp_SW_Price Pulp_HW_Price Sawtimber_SW_Price
                                                    Sawtimber_HW_Price
Comb_Sawtimber_Price    Pulp_Growth_SW_Price Pulp_Growth_HW_Price
                                                    Comb_Pulp_Growth_Price
Sawtimber_Growth_SW_Price Sawtimber_Growth_HW_Price
                                                    Comb_Sawtimber_Growth_Price
Urban_Waste_Price SW_Other_Removals_Price
                                                    HW_Other_Removals_Price
Comb_Other_Rem_Price Thinnings_Price Comb_Pulp_Price
```

```
                                                    Mills_total
Primary_mill_total Secondary_mill_total Pulp_and_paper_mill_total
Other_Mill_total
                                                    Biorefineries Metropolitan
DRY_BIO_HW DRY_BIO_SW DRY_BIO_TOT Sqmiland Population_Density );
Where State_Name in ('Tennessee','Florida','Alabama', 'Louisiana',
'Texas','Oklahoma', 'Arkansas',
                                                    'Virginia','Kentucky','North
Carolina', 'South Carolina',  'Georgia','Mississippi' );
run;

proc sql;
/* the existing and non-probable mill locations */
create table bi.South13States_pulppaper01 as
select * from bi.South13States_pulppaper
where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

/* the potential mill locations */
create table bi.South13States_pulppapermis as
select * from bi.South13States_pulppaper
where BioRef_pulppaper =.;
quit;

/* For each of the 13 states, three data sets are generated: */
/* all ZCTA loations, the existing and non-probable mill locations, and the
potential mill locations */
data bi.Tennessee_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Tennessee';
run;
proc sql;
      create table bi.Tennessee_pulppaper01 as
      select * from bi.Tennessee_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Tennessee_pulppapermis as
      select * from bi.Tennessee_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Florida_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Florida';
run;
proc sql;
      create table bi.Florida_pulppaper01 as
      select * from bi.Florida_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Florida_pulppapermis as
      select * from bi.Florida_pulppaper
      where BioRef_pulppaper =.;
quit;
```

```
data bi.Alabama_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Alabama';
run;
proc sql;
      create table bi.Alabama_pulppaper01 as
      select * from bi.Alabama_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Alabama_pulppapermis as
      select * from bi.Alabama_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Louisiana_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Louisiana';
run;
proc sql;
      create table bi.Louisiana_pulppaper01 as
      select * from bi.Louisiana_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Louisiana_pulppapermis as
      select * from bi.Louisiana_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Oklahoma_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Oklahoma';
run;
proc sql;
      create table bi.Oklahoma_pulppaper01 as
      select * from bi.Oklahoma_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Oklahoma_pulppapermis as
      select * from bi.Oklahoma_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Arkansas_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Arkansas';
run;
proc sql;
      create table bi.Arkansas_pulppaper01 as
      select * from bi.Arkansas_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Arkansas_pulppapermis as
      select * from bi.Arkansas_pulppaper
      where BioRef_pulppaper =.;
```

```
quit;

data bi.Virginia_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Virginia';
run;
proc sql;
      create table bi.Virginia_pulppaper01 as
      select * from bi.Virginia_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Virginia_pulppapermis as
      select * from bi.Virginia_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Kentucky_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Kentucky';
run;
proc sql;
      create table bi.Kentucky_pulppaper01 as
      select * from bi.Kentucky_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Kentucky_pulppapermis as
      select * from bi.Kentucky_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.NorthCarolina_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'North Carolina';
run;
proc sql;
      create table bi.NorthCarolina_pulppaper01 as
      select * from bi.NorthCarolina_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.NorthCarolina_pulppapermis as
      select * from bi.NorthCarolina_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.SouthCarolina_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'South Carolina';
run;
proc sql;
      create table bi.SouthCarolina_pulppaper01 as
      select * from bi.SouthCarolina_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.SouthCarolina_pulppapermis as
```

```
      select * from bi.SouthCarolina_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Georgia_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Georgia';
run;
proc sql;
      create table bi.Georgia_pulppaper01 as
      select * from bi.Georgia_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Georgia_pulppapermis as
      select * from bi.Georgia_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Mississippi_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Mississippi';
run;
proc sql;
      create table bi.Mississippi_pulppaper01 as
      select * from bi.Mississippi_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Mississippi_pulppapermis as
      select * from bi.Mississippi_pulppaper
      where BioRef_pulppaper =.;
quit;

data bi.Texas_pulppaper;
      set bi.South13States_pulppaper;
      where State_Name = 'Texas';
run;
proc sql;
      create table bi.Texas_pulppaper01 as
      select * from bi.Texas_pulppaper
      where BioRef_pulppaper = 0 or BioRef_pulppaper = 1;

      create table bi.Texas_pulppapermis as
      select * from bi.Texas_pulppaper
      where BioRef_pulppaper =.;
quit;
```

A-3 SAS codes for de-clustering algorithms

```
libname ts 'D:\Nancy\Thesis\Decluster';

/* create a data set of ZIP Code pairs with driving distances of less than 80
miles*/
data ts.ls80distance(drop= Drivingtime);
      set ts.ls80;
run;
proc sql;
      create table ts.ls80d as
      select * from ts.ls80distance
      where Drivingdistance<80;
quit;

/* rename the zipcode variable as ZIP2 in the two data sets below for merging
*/
proc datasets library = ts;
      modify codedmillszipcode ;
      rename Zipcode=ZIP2;
run;
quit;
proc datasets library = ts;
      modify biofuel ;
      rename Zipcode=ZIP2;
run;
quit;

/* merge these three data sets by ZIP2 to create neighboring mill info for
each ZIP1*/
proc sort data=ts.ls80d;
      by ZIP2;
run;
proc sort data=ts.codedmillszipcode;
      by  ZIP2;
run;
proc sort data=ts.biofuel;
      by  ZIP2;
run;
data ts.ls80milltypes;
      merge ts.ls80d(in=inls80d) ts.codedmillszipcode ts.biofuel;
      by ZIP2;
      if inls80d;
run;

/* surrogate missing values */
data ts.ls80milltypes;
      set ts.ls80milltypes;
      if Bioref=. then Bioref=0;
      if  Milltotal=. then Milltotal=0;
      if Primary_mill=. then Primary_mill=0;
      if Secondary_mill=. then Secondary_mill=0;
      if Pulp_and_paper_mill=. then Pulp_and_paper_mill=0;
```

```
      if Other_Mill=. then Other_Mill=0;
run;
/* The data set TS.LS80MILLTYPES has 11,462,482 observations and 9
variables*/

/* rename ZIP1 as Zipcode for the merge step */
proc datasets library=ts ;
      modify ls80milltypes ;
      rename Zip1=Zipcode;
run;
quit;

/* convert Zipcodes to the corresponding ZCTAs by merging*/
proc sort data= ts.ls80milltypes;
      by Zipcode;
run;
proc sort data= ts.match;
      by Zipcode;
run;
data ts.ls80ZCTA (drop = Zipcode Zip2);
      merge ts.match(keep=Zipcode ZCTA City County_Name in=inmatch)
            ts.ls80milltypes(in=inls80milltypes);
      by Zipcode;
      if inmatch and inls80milltypes;
run;
/*The data set TS.LS80ZCTA has 11366640 observations and 10 variables.*/

/* create three sets of ZCTAs with neighboring mill info by the travel
radius*/
proc sql;
      create table ts.MillsIn40To80Miles as
      select ZCTA, avg(Drivingdistance) as DrivingDist_40To80,
                  sum(Milltotal) as Millall_40To80,
                  sum(Primary_mill) as Primary_mill_40To80,
                  sum(Secondary_mill) as Secondary_mill_40To80,
                  sum(Pulp_and_paper_mill) as Pulppaper_mill_40To80,
                  sum(Other_mill) as Other_mill_40To80,
                  sum(Bioref) as Bioref_40To80
                  from ts.ls80ZCTA
                  group by ZCTA
                  having avg(Drivingdistance)>=40 and
avg(Drivingdistance)<80;

      create table ts.MillsIn20To40Miles as
      select ZCTA, avg(Drivingdistance) as DrivingDist_20To40,
                  sum(Milltotal) as Millall_20To40,
                  sum(Primary_mill) as Primary_mill_20To40,
                  sum(Secondary_mill) as Secondary_mill_20To40,
                  sum(Pulp_and_paper_mill) as Pulppaper_mill_20To40,
                  sum(Other_mill) as Other_mill_20To40,
                  sum(Bioref) as Bioref_20To40
                  from ts.ls80ZCTA
                  group by ZCTA
```

```
                  having avg(Drivingdistance)>=20 and
avg(Drivingdistance)<40;

      create table ts.MillsIn0To20Miles as
      select ZCTA, avg(Drivingdistance) as DrivingDist_0To20,
                  sum(Milltotal) as Millall_0To20,
                  sum(Primary_mill) as Primary_mill_0To20,
                  sum(Secondary_mill) as Secondary_mill_0To20,
                  sum(Pulp_and_paper_mill) as Pulppaper_mill_0To20,
                  sum(Other_mill) as Other_mill_0To20,
                  sum(Bioref) as Bioref_0To20
                  from ts.ls80ZCTA
                  group by ZCTA
                  having avg(Drivingdistance)<20;
quit;

/* import data of predicted probabilities of ZCTAs as future siting locations
of Group 1 */
proc import out=ts.PredProbAllMills
      datafile= "D:\Nancy\Thesis\Decluster Sept.20\Predicted probability of
all mills.xls"
      DBMS=excel replace;
run;

/* import data of predicted probabilities of ZCTAs as future siting locations
of Group 2 */
proc import out=ts.PredProbPulpPaper
      datafile = "D:\Nancy\Thesis\Decluster Sept.20\predicted probability of
pulp and paper.xls"
      DBMS=excel replace;
run;

/* sort for merging */
proc sort data = ts.MillsIn40To80Miles;
      by ZCTA;
proc sort data = ts.MillsIn20To40Miles;
      by ZCTA;
proc sort data = ts.MillsIn0To20Miles;
      by ZCTA;
proc sort data=ts.PredProbAllMills;
by ZCTA;
proc sort data=ts.PredProbPulpPaper;
by ZCTA;

/* create final data set of ZCTAs with predicted probability and neighboring
mill info in Group 1*/
data ts.PredProbAllMillsAndNearMillNum;
      merge ts.MillsIn40To80Miles(drop=DrivingDist_40To80)
                  ts.MillsIn20To40Miles(drop=DrivingDist_20To40)
                  ts.MillsIn0To20Miles(drop=DrivingDist_0To20)
                  ts.PredProbAllMills(in=inPredProb);
      by ZCTA;
      if inPredProb;
run;
```

```
/* surrogate missing values */
data ts.PredProbAllMillsAndNearMillNum;
      set ts.PredProbAllMillsAndNearMillNum;
      if Millall_0To20=. then Millall_0To20=0;
      if Primary_mill_0To20=. then Primary_mill_0To20=0;
      if Secondary_mill_0To20=. then Secondary_mill_0To20=0;
      if Pulppaper_mill_0To20=. then Pulppaper_mill_0To20=0;
      if Other_mill_0To20=. then Other_mill_0To20=0;
      if Bioref_0To20=. then Bioref_0To20=0;
      if Millall_20To40=. then Millall_20To40=0;
      if Primary_mill_20To40=. then Primary_mill_20To40=0;
      if Secondary_mill_20To40=. then Secondary_mill_20To40=0;
      if Pulppaper_mill_20To40=. then Pulppaper_mill_20To40=0;
      if Other_mill_20To40=. then Other_mill_20To40=0;
      if Bioref_20To40=. then Bioref_20To40=0;
      if Millall_40To80=. then Millall_40To80=0;
      if Primary_mill_40To80=. then Primary_mill_40To80=0;
      if Secondary_mill_40To80=. then Secondary_mill_40To80=0;
      if Pulppaper_mill_40To80=. then Pulppaper_mill_40To80=0;
      if Other_mill_40To80=. then Other_mill_40To80=0;
      if Bioref_40To80=. then Bioref_40To80=0;
run;
/*The data set TS.PREDPROBALLMILLSANDNEARMILLNUM has 3982 observations and 25
variables.*/

/* create final data set of ZCTAs with predicted probability and neighboring
mill info in Group 2*/
data ts.PredProbPulpPaperAndNearMillNum;
      merge ts.MillsIn40To80Miles(drop=DrivingDist_40To80)
            ts.MillsIn20To40Miles(drop=DrivingDist_20To40)
            ts.MillsIn0To20Miles(drop=DrivingDist_0To20)
            ts.PredProbPulpPaper(in=inPredProb);
      by ZCTA;
      if inPredProb;
run;
/* surrogate missing values */
data ts.PredProbPulpPaperAndNearMillNum;
      set ts.PredProbPulpPaperAndNearMillNum;
      if Millall_0To20=. then Millall_0To20=0;
      if Primary_mill_0To20=. then Primary_mill_0To20=0;
      if Secondary_mill_0To20=. then Secondary_mill_0To20=0;
      if Pulppaper_mill_0To20=. then Pulppaper_mill_0To20=0;
      if Other_mill_0To20=. then Other_mill_0To20=0;
      if Bioref_0To20=. then Bioref_0To20=0;
      if Millall_20To40=. then Millall_20To40=0;
      if Primary_mill_20To40=. then Primary_mill_20To40=0;
      if Secondary_mill_20To40=. then Secondary_mill_20To40=0;
      if Pulppaper_mill_20To40=. then Pulppaper_mill_20To40=0;
      if Other_mill_20To40=. then Other_mill_20To40=0;
      if Bioref_20To40=. then Bioref_20To40=0;
      if Millall_40To80=. then Millall_40To80=0;
      if Primary_mill_40To80=. then Primary_mill_40To80=0;
      if Secondary_mill_40To80=. then Secondary_mill_40To80=0;
      if Pulppaper_mill_40To80=. then Pulppaper_mill_40To80=0;
```

```
      if Other_mill_40To80=. then Other_mill_40To80=0;
      if Bioref_40To80=. then Bioref_40To80=0;
run;
/*The data set TS.PREDPROBPULPPAPERANDNEARMILLNUM has 5878 observations and
25 variables*/

/** import the ZCTAs with existing mills for computing the neighboring mill
tolerance in Group 1 **/
proc import datafile='D:\Nancy\Thesis\Decluster Sept.20\total mill01.xls'
      out=ts.totalmill_1 DBMS=excel replace;
run;

/* attach the neighboring mill info to the ZCTAs with exising mills of Group
1 */
proc sort data=ts.totalmill_1;
      by ZCTA;
run;
data ts.RealAllMillsAndNearMillNum;
      merge ts.MillsIn40To80Miles(drop=DrivingDist_40To80)
            ts.MillsIn20To40Miles(drop=DrivingDist_20To40)
            ts.MillsIn0To20Miles(drop=DrivingDist_0To20)
            ts.totalmill_1(in=inTotalMill_1);
      by ZCTA;
      if inTotalMill_1;
run;

/* surrogate missing values */
data ts.RealAllMillsAndNearMillNum;
      set ts.RealAllMillsAndNearMillNum;
      if Millall_0To20=. then Millall_0To20=0;
      if Primary_mill_0To20=. then Primary_mill_0To20=0;
      if Secondary_mill_0To20=. then Secondary_mill_0To20=0;
      if Pulppaper_mill_0To20=. then Pulppaper_mill_0To20=0;
      if Other_mill_0To20=. then Other_mill_0To20=0;
      if Bioref_0To20=. then Bioref_0To20=0;
      if TotalMills_0To20=. then TotalMills_0To20=0;
      if Millall_20To40=. then Millall_20To40=0;
      if Primary_mill_20To40=. then Primary_mill_20To40=0;
      if Secondary_mill_20To40=. then Secondary_mill_20To40=0;
      if Pulppaper_mill_20To40=. then Pulppaper_mill_20To40=0;
      if Other_mill_20To40=. then Other_mill_20To40=0;
      if Bioref_20To40=. then Bioref_20To40=0;
      if TotalMills_20To40=. then TotalMills_20To40=0;
      if Millall_40To80=. then Millall_40To80=0;
      if Primary_mill_40To80=. then Primary_mill_40To80=0;
      if Secondary_mill_40To80=. then Secondary_mill_40To80=0;
      if Pulppaper_mill_40To80=. then Pulppaper_mill_40To80=0;
      if Other_mill_40To80=. then Other_mill_40To80=0;
      if Bioref_40To80=. then Bioref_40To80=0;
      if TotalMills_40To80=. then TotalMills_40To80=0;
run;
/*NOTE: The data set TS.REALALLMILLSANDNEARMILLNUM has 2523 observations and
24 variables.*/
```

```
/** import the ZCTAs with existing mills for computing the neighboring mill
tolerance in Group 2 **/
proc import datafile='D:\Nancy\Thesis\Decluster Sept.20\pulpandpaper01.xls'
      out=ts.pulppaper_1 DBMS=excel  replace;
run;

/* attach the neighboring mill info to the ZCTAs with existing mills of Group
1 */
proc sort data=ts.pulppaper_1;
by ZCTA;
data ts.RealPulpPaperAndNearMillNum;
      merge ts.MillsIn40To80Miles(drop=DrivingDist_40To80)
            ts.MillsIn20To40Miles(drop=DrivingDist_20To40)
            ts.MillsIn0To20Miles(drop=DrivingDist_0To20)
            ts.pulppaper_1(in=inPulpPaper_1);
      by ZCTA;
      if inPulpPaper_1;
run;

/* surrogate missing values */
data ts.RealPulpPaperAndNearMillNum;
      set ts.RealPulpPaperAndNearMillNum;
      if Millall_0To20=. then Millall_0To20=0;
      if Primary_mill_0To20=. then Primary_mill_0To20=0;
      if Secondary_mill_0To20=. then Secondary_mill_0To20=0;
      if Pulppaper_mill_0To20=. then Pulppaper_mill_0To20=0;
      if Other_mill_0To20=. then Other_mill_0To20=0;
      if Bioref_0To20=. then Bioref_0To20=0;
      if TotalMills_0To20=. then TotalMills_0To20=0;
      if Millall_20To40=. then Millall_20To40=0;
      if Primary_mill_20To40=. then Primary_mill_20To40=0;
      if Secondary_mill_20To40=. then Secondary_mill_20To40=0;
      if Pulppaper_mill_20To40=. then Pulppaper_mill_20To40=0;
      if Other_mill_20To40=. then Other_mill_20To40=0;
      if Bioref_20To40=. then Bioref_20To40=0;
      if TotalMills_20To40=. then TotalMills_20To40=0;
      if Millall_40To80=. then Millall_40To80=0;
      if Primary_mill_40To80=. then Primary_mill_40To80=0;
      if Secondary_mill_40To80=. then Secondary_mill_40To80=0;
      if Pulppaper_mill_40To80=. then Pulppaper_mill_40To80=0;
      if Other_mill_40To80=. then Other_mill_40To80=0;
      if Bioref_40To80=. then Bioref_40To80=0;
      if TotalMills_40To80=. then TotalMills_40To80=0;
run;
/*NOTE: The data set TS.REALPULPPAPERANDNEARMILLNUM has 191 observations and
24 variables.*/

proc sql;
/* compute the tolerance of neighboring mills for Group 1 */
create table ts.AllMillsTolerance as
      select max(Millall_40To80) as Millall_40To80Toler,
            max(Primary_mill_40To80) as Primary_mill_40To80Toler,
            max(Secondary_mill_40To80) as Secondary_mill_40To80Toler,
            max(Pulppaper_mill_40To80) as Pulppaper_mill_40To80Toler,
```

```
            max(Other_mill_40To80) as Other_mill_40To80Toler,
            max(Bioref_40To80) as Bioref_40To80Toler,
            max(Millall_20To40) as Millall_20To40Toler,
            max(Primary_mill_20To40) as Primary_mill_20To40Toler,
            max(Secondary_mill_20To40) as Secondary_mill_20To40Toler,
            max(Pulppaper_mill_20To40) as Pulppaper_mill_20To40Toler,
            max(Other_mill_20To40) as Other_mill_20To40Toler,
            max(Bioref_20To40) as Bioref_20To40Toler,
            max(Millall_0To20) as Millall_0To20Toler,
            max(Primary_mill_0To20) as Primary_mill_0To20Toler,
            max(Secondary_mill_0To20) as Secondary_mill_0To20Toler,
            max(Pulppaper_mill_0To20) as Pulppaper_mill_0To20Toler,
            max(Other_mill_0To20) as Other_mill_0To20Toler,
            max(Bioref_0To20) as Bioref_0To20Toler
      from ts.RealAllMillsAndNearMillNum;

/* compute the tolerance of neighboring mills for Group 2 */
create table ts.PulpPaperTolerance as
      select max(Millall_40To80) as Millall_40To80Toler,
            max(Primary_mill_40To80) as Primary_mill_40To80Toler,
            max(Secondary_mill_40To80) as Secondary_mill_40To80Toler,
            max(Pulppaper_mill_40To80) as Pulppaper_mill_40To80Toler,
            max(Other_mill_40To80) as Other_mill_40To80Toler,
            max(Bioref_40To80) as Bioref_40To80Toler,
            max(Millall_20To40) as Millall_20To40Toler,
            max(Primary_mill_20To40) as Primary_mill_20To40Toler,
            max(Secondary_mill_20To40) as Secondary_mill_20To40Toler,
            max(Pulppaper_mill_20To40) as Pulppaper_mill_20To40Toler,
            max(Other_mill_20To40) as Other_mill_20To40Toler,
            max(Bioref_20To40) as Bioref_20To40Toler,
            max(Millall_0To20) as Millall_0To20Toler,
            max(Primary_mill_0To20) as Primary_mill_0To20Toler,
            max(Secondary_mill_0To20) as Secondary_mill_0To20Toler,
            max(Pulppaper_mill_0To20) as Pulppaper_mill_0To20Toler,
            max(Other_mill_0To20) as Other_mill_0To20Toler,
            max(Bioref_0To20) as Bioref_0To20Toler
      from ts.RealPulpPaperAndNearMillNum;

/* attach the tolerance numbers to the ZCTAs as potential siting locations
for Group 1 */
create table ts.PulpPaperAdjustedProb as
      select * from ts.PredProbPulpPaperAndNearMillNum,
ts.PulpPaperTolerance;

/* attach the tolerance numbers to the ZCTAs as potential siting locations
for Group 2 */
create table ts.AllMillsAdjustedProb as
      select * from ts.PredProbAllMillsAndNearMillNum, ts.AllMillsTolerance;
quit;

/* compute the adjusted probability of ZCTAs as future siting locations */
/* based on the tolerance numbers for Group 1 */
data ts.AllMillsAdjustedProb (drop = i);
      set ts.AllMillsAdjustedProb;
```

```
      adjProb = prob_all;
      array millNumber(*) Primary_mill_40To80
                  Secondary_mill_40To80 Pulppaper_mill_40To80
                  Primary_mill_20To40 Secondary_mill_20To40
                  Pulppaper_mill_20To40 Primary_mill_0To20
                  Secondary_mill_0To20 Pulppaper_mill_0To20;
      array toler(*) Primary_mill_40To80Toler
                  Secondary_mill_40To80Toler
                  Pulppaper_mill_40To80Toler
                  Primary_mill_20To40Toler
                  Secondary_mill_20To40Toler
                  Pulppaper_mill_20To40Toler
                  Primary_mill_0To20Toler
                  Secondary_mill_0To20Toler
                  Pulppaper_mill_0To20Toler;
      do i = 1 to 9;
            adjProb = adjProb * exp(-log(2) * millNumber(i) / (toler(i)+1));
      end;
run;

/* show the top 25 future locations before de-cluster for Group 1*/
proc sort data = ts.AllMillsAdjustedProb;
      by descending prob_all;
run;
data ts.allmillsOriginaltop25(keep=ZCTA adjprob Prob_all);
      set ts.AllMillsAdjustedProb(obs=25);
run;

/* show the top 25 future locations after de-cluster for Group 1*/
proc sort data = ts.AllMillsAdjustedProb;
      by descending adjProb;
run;
data ts.allmillstop25(keep=ZCTA adjprob Prob_all);
      set ts.AllMillsAdjustedProb(obs=25);
run;

/* compute the adjusted probability of ZCTAs as future siting locations */
/* based on the tolerance numbers for Group 2 */
data ts.PulpPaperAdjustedProb (drop = i);
      set ts.PulpPaperAdjustedProb;
      adjProb = prob_pulp;
      array millNumber(*) Primary_mill_40To80
                  Secondary_mill_40To80 Pulppaper_mill_40To80
                  Primary_mill_20To40 Secondary_mill_20To40
                  Pulppaper_mill_20To40  Primary_mill_0To20
                  Secondary_mill_0To20 Pulppaper_mill_0To20;
      array toler(*) Primary_mill_40To80Toler
                  Secondary_mill_40To80Toler
                  Pulppaper_mill_40To80Toler
                  Primary_mill_20To40Toler
                  Secondary_mill_20To40Toler
                  Pulppaper_mill_20To40Toler
                  Primary_mill_0To20Toler
                  Secondary_mill_0To20Toler
```

```
                  Pulppaper_mill_0To20Toler;
      do i = 1 to 9;
            if i = 3 or i = 6 or i = 9 then
                  adjProb = adjProb * exp(-log(2) * millNumber(i) /
(toler(i)+1));
            else if millNumber(i) > toler(i) then
                  adjProb = adjProb * exp(-log(2) * millNumber(i) /
(toler(i)+1));
      end;
run;

/* show the top 25 future locations before de-cluster for Group 2 */
proc sort data = ts.PulpPaperAdjustedProb;
      by descending prob_pulp;
run;
data ts.pulppaperOriginaltop25(keep=ZCTA adjprob prob_pulp);
      set ts.PulpPaperAdjustedProb(obs=25);
run;

/* show the top 25 future locations after de-cluster for Group 2 */
proc sort data = ts.PulpPaperAdjustedProb;
      by descending adjProb;
run;
data ts.pulppapertop25(keep=ZCTA adjprob prob_pulp);
      set ts.PulpPaperAdjustedProb(obs=25);
run;
```

Vita

Xu Liu is a Graduate Research Assistant working under the direction of Professor Timothy Young in the Forest Products Center at the University of Tennessee, Knoxville. She is planning to graduate from the University of Tennessee with a Master of Science in Statistics in December 2009. She received a Master of Arts in Economics from Nankai University, Tianjin, China in 2006, a Bachelor's degree in Accounting from Dongbei University of Finance & Economics, Dalian, China in 2001 and graduated from Dalian Polytechnic University, Dalian, China with an Associate's degree in Industrial Foreign Trade in 2000. She worked for HyClone International Trade (Tianjin) Co., Ltd. as a business assistant from July 2000 to June 2001. Then, she became a business supervisor in the HyClone-PIERCE Beijing Representative Office, China and worked there until July 2003.