

To the Graduate Council:

I am submitting herewith a thesis written by Xia Huang entitled “Bayesian Logistic Regression Model for Siting 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. Zaretski

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate
School

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

Bayesian logistic regression models for siting biomass-using facilities

A Thesis

Presented for the

Master of Science Degree

The University of Tennessee, Knoxville

Xia Huang

December 2010

Dedication

This thesis is dedicated to my beloved parents and auntie, Helen Huang for their tremendous help, support and love.

Acknowledgements

I earnestly thank my advisor, Dr. Timothy M. Young, for his enduring assistance and time on both my Master's research and overall academic well-being. A special thank has to be addressed to my committee members, Dr. Frank M. Guess and Dr. Russell L. Zaretzki, who provided extensive support and understanding on my studying and on my personal life.

Great thanks go to Mr. James H. Perdue (U.S. Forest Service) for his professional guidance on the biomass definition and genius development of the public domain site www.BioSAT.net. The database I built in this thesis is an expansion of the website's database. Thanks to Ms. Sachiko Hurst (UT Office of Bioenergy Programs) for her great help with data extraction from SQL server.

I would further like to thank U.S. Forest Service employee and UT graduate student Mr. Andrew J. Hartsell, Research Associate Ms. Kerri E. Norris, and Research Associate Ms. Christy Pritchard for their support on my data collection and data management. I appreciate Post Doctoral Research Associate Dr. Nicolas André for his help with data generation and computation. I must note the kind assistance of Ms. Xu Liu, who built the woody biomass database and developed classical logistic regression models. I thank Graduate Research Assistants Mr. Zeng Yan and Mr. Dillon Carty for their time to review the thesis and for their continuous encouragement. I also thank Administrative Support Assistant Ms. Amanda Silk for her help and kindness. In addition, this research is funded by U.S. Forest Service, Southern Research Station and the Southeastern Regional Sun Grant Center at the University of Tennessee.

Abstract

Key sources of oil for western markets are located in complex geopolitical environments that increase economic and social risk. The amalgamation of economic, environmental, social and national security concerns for petroleum-based economies have created a renewed emphasis on alternative sources of energy which include biomass. The stability of sustainable biomass markets hinges on improved methods to predict and visualize business risk and cost to the supply chain.

This thesis develops Bayesian logistic regression models, with comparisons of classical maximum likelihood models, to quantify significant factors that influence the siting of biomass-using facilities and predict potential locations in the 13-state Southeastern United States for three types of biomass-using facilities. Group I combines all biomass-using mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants. Group II included pulp and paper mills, and biorefineries that use agricultural and wood residues. Group III included food processing mills and biorefineries that use agricultural and wood residues. The resolution of this research is the 5-digit ZIP Code Tabulation Area (ZCTA), and there are 9,416 ZCTAs in the 13-state Southeastern study region.

For both classical and Bayesian approaches, a training set of data was used plus a separate validation (hold out) set of data using a pseudo-random number-generating function in SAS[®] Enterprise Miner. Four predefined priors are constructed. Bayesian estimation assuming a Gaussian prior distribution provides the highest correct classification rate of 86.40% for Group I; Bayesian methods assuming the non-informative uniform prior has the highest correct

classification rate of 95.97% for Group II; and Bayesian methods assuming a Gaussian prior gives the highest correct classification rate of 92.67% for Group III. Given the comparative low sensitivity for Group II and Group III, a hybrid model that integrates classification trees and local Bayesian logistic regression was developed as part of this research to further improve the predictive power. The hybrid model increases the sensitivity of Group II from 58.54% to 64.40%, and improves both of the specificity and sensitivity significantly for Group III from 98.69% to 99.42% and 39.35% to 46.45%, respectively. Twenty-five optimal locations for the biomass-using facility groupings at the 5-digit ZCTA resolution, based upon the best fitted Bayesian logistic regression model and the hybrid model, are predicted and plotted for the 13-state Southeastern study region.

Table of Contents

Chapter	Page
Chapter 1 Introduction.....	1
Chapter 2 Literature Review	4
2.1 Bioenergy and Biofuels	4
2.1.1 Introduction to Bioenergy and Biofuels	4
2.1.2 Importance of Bioenergy and Biofuels.....	5
2.2 Biomass	7
2.2.1 Concepts of Biomass	7
2.2.2 Availability of Biomass	8
2.2.3 Benefits and Concerns of Biomass Utilization.....	10
2.3 Biomass-using Facilities Siting Models.....	11
2.4 Bayesian Logistic Regression Model.....	13
2.4.1 General Introduction of Bayesian Logistic Regression.....	13
2.4.2 Application History	16
Chapter 3 Data and Analysis Methods.....	18
3.1 Variable Explanation.....	18
3.1.1 Three Study Groups.....	18
3.1.2 The Response Variable.....	20
3.1.3 Explanatory Variables	24
3.2 Data Management	29
3.2.1 Data Sources	29
3.2.2 Biomass Estimation Using GIS	31
3.2.3 Transportation-related Data Generation	32
3.2.4 Data Standardization and Treatment of Missing Data.....	38
3.3 Bayesian Logistic Regression Model.....	39
3.3.1 Logistic Regression	39
3.3.2 Model Selection Methods and Criteria.....	42
3.3.3 Bayesian Inference	43

3.3.4 Classification Trees with Local Bayesian Logistic Regression.....	45
3.3.5 Model Assessment.....	47
3.3.6 Model Scoring and Interpretation.....	48
Chapter 4 Results and Discussion.....	49
4.1 Logistic Regression Model Selection.....	49
4.1.1 Group I Biomass-using Facilities	49
4.1.2 Group II Biomass-using Facilities.....	51
4.1.3 Group III Biomass-using Facilities.....	53
4.2 Maximum Likelihood Estimates and Bayesian Estimation	55
4.2.1 Group I Biomass-using Facilities	55
4.2.2 Group II Biomass-using Facilities.....	60
4.2.3 Group III Biomass-using Facilities.....	67
4.3 Classification Trees with Local Bayesian Logistic Regression	72
4.3.1 Group II Biomass-using Facilities.....	72
4.3.2 Group III Biomass-using Facilities.....	77
Chapter 5 Conclusions and Future Research.....	84
References.....	89
Appendices.....	102
A-1 ArcGIS codes for Biomass Estimation	103
A-2 Maps of Agricultural Residue Availability in 13 Southeastern U.S. states	108
A-3 A SAS macro for converting datasets from SAS to R and WinBUGS.....	109
A-4 WinBUGS and R codes for Bayesian inferences.....	118
A-5 Estimations of Bayesian inferences	131
Vita	137

List of Tables

Table	Page
Table 3-1 Explanatory variables for three groups of biomass-using facilities.....	25
Table 3-2 Combination value for variables at 5-digit ZCTAs and surrogating number for missing values	39
Table 4-1 AIC, BIC values and misclassification rates of the four models for Group I.....	49
Table 4-2 Classification Table from the validation dataset for Group I biomass-using facilities .	51
Table 4-3 AIC, BIC values and misclassification rates of the four models for Group II	52
Table 4-4 Classification Table from the validation dataset for Group II biomass-using facilities	53
Table 4-5 AIC, BIC values and misclassification rates of the four models for Group III.....	54
Table 4-6 Classification Table from the validation dataset for Group III biomass-using facilities	55
Table 4-7 Classification rates under maximum likelihood estimation and Bayesian methods for Group I biomass-using facilities.....	56
Table 4-8 Parameter estimates under Gaussian prior for Group I biomass-using facilities	58
Table 4-9 Classification rates under maximum likelihood estimation and Bayesian methods for Group II biomass-using facilities	62
Table 4-10 Parameter estimates under Uniform prior for Group II biomass-using facilities	64
Table 4-11 Classification rates under maximum likelihood estimation and Bayesian methods for Group III biomass-using facilities	68
Table 4-12 Parameter estimates under Gaussian prior for Group III biomass-using facilities.....	69
Table 4-13 Significant variables for five subsets of Group II biomass-using facilities.....	74

Table 4-14 Correct classification rates under maximum likelihood estimation and Bayesian methods for 5 subsets and the whole set of Group II biomass-using facilities.....	75
Table 4-15 Classification table by summing up classified values achieved from the “best” local model in each of five subsets for Group II biomass-using facilities.....	76
Table 4-16 Significant variables for three subsets of Group III biomass-using facilities	79
Table 4-17 Correct classification rates under maximum likelihood estimation and Bayesian methods for 3 subsets and the whole set of Group III biomass-using facilities	81
Table 4-18 Classification table by summing up classified values achieved from the “best” local model in each of three subsets for Group III biomass-using facilities	81
Table A-5.1 Parameter estimates under Uniform prior for Group I biomass-using facilities	131
Table A-5.2 Parameter estimates under Laplace prior for Group I biomass-using facilities	132
Table A-5.3 Parameter estimates under Cauchy prior for Group I biomass-using facilities	133
Table A-5.4 Parameter estimates under Gaussian prior for Group II biomass-using facilities ..	134
Table A-5.5 Parameter estimates under Laplace prior for Group II biomass-using facilities	134
Table A-5.6 Parameter estimates under Cauchy prior for Group II biomass-using facilities.....	135
Table A-5.7 Parameter estimates under Uniform prior for Group III biomass-using facilities..	135
Table A-5.8 Parameter estimates under Laplace prior for Group III biomass-using facilities...	136
Table A-5.9 Parameter estimates under Cauchy prior for Group III biomass-using facilities ...	136

List of Figures

Figure	Page
Figure 3-1 All biomass-using mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants in 13 Southeastern states	21
Figure 3-2 Pulp and paper mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants in 13 Southeastern states	22
Figure 3-3 Food processing mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants in 13 Southeastern state	23
Figure 3-4 Flow Chart of estimating agricultural residues at the level of 5-digit ZCTA	33
Figure 3-5 Agricultural residue quantity at the level of 5-digit ZCTA in 13 Southeastern states	34
Figure 3-6 Agricultural residue harvesting costs at the level of 5-digit ZCTA in 13 Southeastern states.....	35
Figure 3-7 Supply agricultural residues from neighboring ZCTAs to a demand ZCTA within 40- mile, 80-mile, 120-mile, and 160 mile haul driving distances.....	37
Figure 3-8 Four distinct models in SAS [®] Enterprise Miner	41
Figure 4-1 Cumulative and Non-Cumulative lift charts that assesses four models for Group I biomass-using facilities.....	50
Figure 4-2 Cumulative and Non-Cumulative lift charts that assesses four models for Group II biomass-using facilities.....	52
Figure 4-3 Cumulative and Non-Cumulative lift charts that assesses four models for Group III biomass-using facilities.....	54
Figure 4-4 Posterior distributions of parameters for Group I biomass-using facilities	59

Figure 4-5 Top 25 optimal locations for Group I biomass-using facilities in Southeastern 13 states.....	61
Figure 4-6 Posterior distributions of parameters for Group II biomass-using facilities.....	65
Figure 4-7 Top 25 optimal locations for Group II biomass-using facilities in Southeastern 13 states.....	66
Figure 4-8 Posterior distributions of parameters for Group III biomass-using facilities.....	70
Figure 4-9 Top 25 optimal locations for Group III biomass-using facilities in Southeastern 13 states.....	71
Figure 4-10 The classification tree build by JMP® for Group II biomass-using facilities	73
Figure 4-11 Top 25 optimal locations based upon the hybrid model for Group II biomass-using facilities in Southeastern 13 states	78
Figure 4-12 The classification tree build by JMP® for Group III biomass-using facilities	79
Figure 4-13 Top 25 optimal locations based upon the hybrid model for Group III biomass-using facilities in Southeastern 13 states	83
Figure A-2.1 Agricultural residue quantity and harvest costs at the level of 5-digit ZCTA in 13 Southeastern states	108

Chapter 1 Introduction

Sustainable economic growth requires safe resources of energy for the industrial production. Today's most frequently used industrial energy, petroleum, is neither sustainable, because of depleting oil reserves, nor environmentally friendly. Concerns over environmental health and national security accelerate the emergence towards the development of renewable energy. While various alternative energy resources, such as wind power, solar, hydropower, biomass, as well as nuclear fission and fusion, successfully support economic development, biofuels derived from biomass resources offer a promising solution for substitution away from petroleum-based energy.

Research and development work in the emerging field of “biomass and bioenergy” are most notable in the United States and Europe (e.g., Kamm *et al.* 1998, Wright 2006, Galik 2009). Significant industrial developments have been in the United States since 2000. The United States is expected to produce 36 million gallons of biofuels by 2020, which represents at least 25% of liquid fuels from the new bio-based product industry. China, Brazil, and India have also recognized the importance of implementing bioenergy, and have established several national programs in bioenergy research, development, and deployment. China targets 10% (30 billion watts) of its electricity generation from bioenergy by 2020.

A biorefinery, which requires a large and constant supply of biomass, is a facility that integrates biomass conversion processes and equipment to produce fuels, power, or value-added chemicals. Biomass for use in the biorefinery includes any organic matters that contain stored energy from the sun. Four main biomass classifications exist – wood or agricultural residues,

municipal solid waste, landfill gas and biogas, and alcohol fuels (Calle *et al.* 2007). In this thesis, the focus is wood or agricultural residues biomass. The goal of this research is to identify and quantify influential factors on siting biorefineries and biofuels plants that use woody or agricultural-derived biomass. Five specific objectives were developed to achieve this goal:

- 1) Develop an expanded database from the BioSAT model (Young *et al.* 2009) and previous work from Liu (2009), which includes socio-demographic variables, agricultural residue availability variables, and transportation-related variables;
- 2) Improve the data quality and consistency from the use of GIS and spatial analysis;
- 3) Develop logistic regression models and apply Bayesian inference to identify variables influencing locations for biorefineries and biofuels using wood or agricultural residues;
- 4) Compare Bayesian logistic regression models with classical logistic regression models;
- 5) Predict sites for potential biorefineries and biofuels that would use wood or agricultural residues.

The remainder of this thesis is organized into four chapters of literature review, data and analysis methods, results and discussion, conclusions, and future research. Chapter Two provides a literature review with a general introduction to bioenergy and biofuels, discussions of its importance, and a review of biomass feasibility. Attention is given to previous research related to the siting model of biomass-using facilities, with a review of the development and application of Bayesian logistic regression. Chapter Three summarizes data acquisition, processing, and management. This chapter also summarizes model construction, selection, and assessment. Chapter Four reports model results for the three predefined groups of biomass-using

facilities. Bayesian logistic regression models are examined for predictive power through comparative studies of maximum likelihood estimates and Bayesian inference. Significant factors in the 13-state region are explained in detail for the three predefined groups. The top 25 potential locations in the 13-state region for each group of biomass-using facilities are presented and mapped. Chapter Five synthesizes the findings from this study and provides suggestions for future research.

Chapter 2 Literature Review

2.1 Bioenergy and Biofuels

2.1.1 Introduction to Bioenergy and Biofuels

Development of alternative energy source is of major importance in the United States as high consumption of petroleum-based energy is an infeasible long term energy solution. According to Energy Information Administration (EIA) reports, 84% of U.S. primary energy consumption in 2008 was from fossil fuels. The United States began to import energy in 1950s, and the net imported energy has increased 16-fold by 2008, which accounted for 26% of all energy consumed (Annual Energy Review 2008). Conversion of biomass offers potential as one such energy source that is renewable, abundant, and environmentally friendly. According to U.S. Department of Energy (2010), “bioenergy uses renewable biomass resources to produce an array of energy related products including electricity, liquid, solid, and gaseous fuels, heat, chemicals, and other materials”. Today, bioenergy ranks second (to hydropower) in renewable U.S. primary energy production and accounts for three percent of the primary energy production in the United States (U.S. Department of Energy 2010). Bioenergy could effectively contribute to reducing the U.S. dependence on foreign oil and fossil fuels.

The renewable nature of biofuels as an energy source is an important asset for the future potential from a security standpoint. Biofuels, as the name implies, are fuels made from biomass resources through biochemical (fermentation of sugar to alcohol, and anaerobic digestion or fermentation) or thermochemical processes (gasification, pyrolysis, liquefaction). Broadly, biofuels include ethanol, methanol, biodiesel, biocrude, and methane (Parlack *et al.* 2005),

covering solid biomass, i.e., wood, agricultural waste, energy crops and dried manure; liquid fuels, such as bioalcohols, biodiesel and methanol; and various biogases. Strictly, biofuels consist of liquid and gas fuels commonly used for automotive transportation (Fulton *et al.* 2004). Biofuels have many advantages, including sustainability, reduction of greenhouse gas emissions, and guaranteed supply of raw materials (Reijnders 2006), that are increasingly receiving public, scientific and legislative attention. Currently, biofuel industries are expanding rapidly in Europe, Asia, South America, and the United States (Byrne *et al.* 1996, Puhan *et al.* 2005, Soccol *et al.* 2005, Amigun *et al.* 2008). Wright (2006) discussed the worldwide biomass energy development opportunities, and summarized the main sources of bioenergy. She found that “biofuel feedstocks include sugar from sugarcane in Brazil, starch from maize grain in the U.S., and oil seeds (soy or rapeseed) for biodiesel in the U.S., EU, and Brazil.” Holm-Nielsen *et al.* (2006) summarized total area and areas of interest for biomass production in the world, and discussed biomass utilization in the energy and industry from a future perspective.

2.1.2 Importance of Bioenergy and Biofuels

Bioenergy is an attractive energy source for a number of reasons. First, it is renewable as long as it is properly managed. Bioenergy provides the opportunity for increased local, regional, and national energy self-sufficiency across the globe. It reduces the dependency on foreign oil and enhances national energy security. Increasing the use of bioenergy also leads to improved economic development, especially in rural areas, since it attracts investment in new business opportunities for small-and medium-sized enterprises in the field of biofuel production,

preparation, transportation, trade and use. Producing and using more biofuels results in fewer greenhouse gas emissions than petroleum fuel counterparts.

The growing seriousness of the global energy problem and associated concerns over environmental pollution, economic health, and national security are substantially increasing the importance of the development of bioenergy and biofuels. The U.S. federal government has set the goal of tripling U.S. use of bioenergy and biofuels by the year 2010 (U.S. Department of Energy 2010). Meeting the goal could create \$15-20 billion a year in new income for farmers and rural America, and reduce annual greenhouse gas emissions. As a substitute for fossil fuels, bioenergy and biofuels have attracted much recent attention among researchers. Biomass Research and Development Board (2009) provided an extensive discussion on economic drivers and environmental implications of increasing feedstock production for bioenergy and biofuels in the United States. Milbrandt (2005) gave a geographic perspective on the availability of biomass resources in the United States. Brechbill and Tyner (2008) used harvest switchgrass and corn stover to compute the up-to-date economic costs of biomass collection, transportation, and supply to Indiana cellulosic and electric utility facilities, and also estimated the greenhouse gas emissions reduction from using biomass instead of coal. Galik *et al.* (2009) analyzed three U.S. southern states aggregate bioenergy potential and the potential supply cost of woody biomass. Hoskinson *et al.* (2007) evaluated the ethanol conversion yields of four corn stover harvest scenarios in Iowa. Thompson *et al.* (2010) studied the U.S. biodiesel use mandate and biodiesel feedstock markets and suggested a hierarchy of price effects for industrial and feed use.

Increasing bioenergy research and development takes the lead in providing the technology for a bio-based economy in the 21st century (Singh *et al.* 2003). The U.S.

Department of Agriculture and Energy (2009) announced grants for more than \$24 million in research and development of technologies to produce biofuels, bioenergy and high-value biobased products. Mckendry (2002) provided a general overview of the main biomass conversion technologies with specific regard to the production of a fuel suitable for spark ignition gas engines. Ptasiński *et al.* (2007) compared different types of biofuels with reference to their gasification efficiency. Demirbas (2007) described the combustion and conversion technologies used for the generation of electricity and heat, and discussed the progress and recent trends in biofuels.

2.2 Biomass

2.2.1 Concepts of Biomass

In general, the definition of biomass is: 1) the total mass or number of living organisms in a particular unit area or volume of habitat (Merriam-Webster 2010); 2) organic matter that has stored solar energy through the process of photosynthesis (U.S. Department of Energy 2010); 3) biological material derived from living organism that can be used as a source of energy or for its chemical composition. By combining the second and third definition, biomass is defined in this thesis as any biological material derived from living organism storing solar energy through the process of photosynthesis that can be used as a source of energy or for its chemical composition. Biomass is available on a renewable or recurring basis (Utah Department of Natural Resources 2010), including wood or agricultural residues, municipal solid waste, landfill gas and biogas, and alcohol fuels (Calle *et al.* 2007). In this thesis we focus on the first biomass source of wood and agricultural residues.

Wood residues are forest-derived biomass which includes small diameter trees, tops, limbs, branches and other woody parts that would not be included in a conventional timber harvest (Massachusetts Sustainable Forest Bioenergy Initiative 2008). Wood residues can be obtained from logging operations, industry forest product manufacturers, fuel treatment thinning, and urban wood waste. Woody biomass-using facilities include five different types of mills: primary wood processing mills, secondary wood processing mills, pulp and paper mills, other mills, and woody biomass-using bioenergy and biofuels plants (Liu 2009).

Agricultural residues are agriculture-derived biomass that remains in the field after harvest. The most common residues include corn stover (stalks, leaves, and/or cobs), and straw associated with wheat, rice, barley, or oat production. Five specific types of agricultural residues are studied in this research: Barley Straw, Corn Stover, Sorghum Straw, Oat Straw, and Wheat Straw. Because of their immediate availability, agricultural residues are expected to play an early role in the development of the cellulosic ethanol industry. In addition, agricultural residues are a good replacement of virgin wood fiber for the construction material and pulp and paper industries (Hayes 1998).

2.2.2 Availability of Biomass

Biomass resources are in general very widespread and abundant on the Earth's surface. According to U.S. Department of Agriculture and U.S. Department of Energy (2005), over 1.3 billion dry tons per year of biomass potential in the United States could be available for biofuels production to meet more than one-third of the current demand for transportation fuels. About

368 million dry tons of sustainably removable biomass could be annually produced on forestlands, and about 998 million dry tons per year, besides meeting food, feed, and export demands, could come from agricultural lands.

Many studies have been undertaken to assess the possible contribution of biomass to the future global energy supply. Shell International Petroleum Company, Inc. (1995) laid out scenarios in which biomass resources would provide one-half to two-thirds as much energy worldwide as fossil fuels do by 2025. Fischer and Schrattenholzer (2001) conducted a bottom-up calculation by using land use model with complementary data to estimate potential biomass supply for energy from 1990 to 2050. Smeets *et al.* (2007) examined the underlying factors, i.e., population growth, demand for wood and crop yields, per capita consumption of food, natural forest growth, and wood production from plantations, to determine and calculate global and regional bioenergy production potentials derived from biomass resources by 2050.

However, estimates of the global bioenergy production potential vary from 33 to 1135 EJ/yr¹ (Hoogwijk 2002). Bernes *et al.* (2003) provided a detailed analysis and comparison on 17 studies on global biomass production potentials. He pointed out that the major reason for the estimation differences resulted from the fact that most studies were either demand or resource focused which ignored or failed to model demand and supply interactions. Vries *et al.* (2007) included geographical (land use), techno-economic (scale, labor cost), and institutional (policy regime, legislation) factors explicitly considering the interactions between demand and supply to evaluate the potential availability of biomass energy at a global level for the period 2000-2050.

¹ 1EJ equals to 10¹⁸Joules

2.2.3 Benefits and Concerns of Biomass Utilization

Harvesting biomass presents both opportunities and challenges. Bartuska (2006) summarized the positive attributes of biomass utilization that contribute to the environment and economy. She pointed out, “biomass utilization can reduce forest management costs, help mitigate climate change, reduce risks to life and property, and help provide a secure, competitive energy source. Shifting to a homegrown, renewable energy economy provides opportunities for growth and expansion, especially for rural communities as these renewable feedstocks are directly connected to the land, primarily agricultural, and forestry lands”. Groscurth *et al.* (2000) conducted a comprehensive analysis of the economic and environmental performance of the energy use of biomass for selected existing facilities throughout the European Union. They claimed that the appropriately organized use of biomass for energy had significant environmental advantages compared to the use of fossil fuels. Hill *et al.* (2006) evaluated environmental, economic and energetic costs of ethanol from corn grain and biodiesel from soybeans, and concluded that ethanol from corn grain and biodiesel from soybeans could provide much more useful energy than the fossil energy, reduce several major air pollutants, and had minimal impacts on human and environmental health. However, Hill *et al.* (2006) also pointed out that the high production costs made biomass for the use of energy unprofitable without subsidies even though they have advantages coming from low agricultural inputs and efficient conversion of feedstocks to fuel.

Despite the merits of biomass utilization stated above, there are some challenges pose by biomass harvesting and utilization for energy. Caputo (2009) summarized three main problems. First, the potential for an increased demand of biomass would drive the level of biomass harvesting to an unsustainable level. The negative consequences may create problems with

biodiversity, soil productivity, water conservation, and invasive species infestation. The cost of biomass associated with harvesting, transporting, storing, and utilizing the materials often exceeded its value on the energy market. That would make its price uncompetitive to fossil fuels. Moreover, the equipment, conversion technologies, and logistic processes were under development for the cost-effective biomass. Caputo (2009) finally emphasized the role of federal policies on ensuring sustainability of biomass harvesting and improving its economic feasibility.

2.3 Biomass-using Facilities Siting Models

A critical factor in ensuring the viability of biomass-using bioenergy and biofuels plants is to identify economically favorable sites. Considering environmental impacts, economic influence, political incentives, transportation-related factors, and availability of feedstocks and labor for biomass-using facilities, developing siting models to examine possible influential factors and to identify potential economically favorable sites may be complex. Significant research efforts have focused on this topic.

Sperling (1984) presented a generalized analytical framework which embodied a disaggregate microscale approach to identify and specify critical factors for determining the attractiveness of biomass fuel plants. This disaggregated microscale approach integrated site-specific considerations and included five crucial issues: feedstock supply, fuel distribution, fuel demand, co-product demand, and feedstock processing, which exert the most strongly influence on the site selection and size determination of biomass fuel plants. The systematic framework

could be applied in any areas with abundant biomass. The results provided insights into formulating and analyzing public policies and actions.

Young *et al.* (1991) developed a deterministic model with the incorporation of spatial analyses using Geographic Information System (GIS) as a site selection tool to estimate the economic availability of woody biomass for energy production in the Southeastern United States. The model included market and non-market factors, and found Northeast Florida, Southern Georgia, Southern Alabama, and the Coastal Plain of South Carolina remained relatively low cost for energy production from woody biomass, while the South Delta of Louisiana, Kentucky, West Virginia, and the mountain regions of Tennessee and Virginia were the highest cost areas.

Graham *et al.* (2000) constructed a regional-scale GIS-based modeling system to estimate potential biomass supplies from energy crops and assess optimal locations for siting biomass-using facilities. This system considered the growing regions of energy crops, the spatial variability in their yield, and transportation costs associated with acquiring the feedstock needed for an energy facility to estimate potential switchgrass costs and supplies in 11 U.S. states. They concluded that transportation costs increased with growing facility demand; Iowa, North Dakota, and South Dakota remained the lowest costs, while South Carolina, Missouri, Georgia, and Alabama had the highest costs. This system provided quantitative economic and environmental information on potential biomass supplies to federal or state policy makers and agencies, thus improving bioenergy policy decision making.

Leduc *et al.* (2008) established a linear mixed integer programming model to determine the optimal geographic locations and sizes of biofuels plants in Austria. The model analyzed the energy chain of biofuel production, which included biomass harvesting and transportation, biofuel production, transportation and distribution, and calculated the total costs of this energy

chain. With the minimization of the total costs, three optimal geographical locations and sizes of biofuel plants in Austria were determined. The developed model provided a useful tool for the biomass-based planning process.

Liu (2009) developed logistic regression models to improve the understanding of the key factors that influence the locations of two groups of existing wood-using bioenergy and biofuels plants, and other traditional woody biomass-using plants in the 13 Southeastern U.S. states. The model included socio-economic, transportation-related, and woody biomass availability variables to identify the statistically significant factors on siting bioenergy and biofuels manufacturing facilities. Twenty-five optimal locations for each study group at a 5-digit ZCTA level in the Southeastern U.S. states were predicted from the logistic regression models. This study provided a new insight for the application of logistic models for optimal siting of woody biomass-using facilities.

2.4 Bayesian Logistic Regression Model

2.4.1 General Introduction of Bayesian Logistic Regression

Bayesian logistic regression, which applies Bayesian inference, has the formulation of a logistic equation and includes both continuous and categorical explanatory variables. The logistic equation transforms the categorical response variables into logarithmic forms, thus making the forms of the coefficients of explanatory variables consistent with other linear models. The general form of the logistic equation is:

$$\log it [\theta(x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

Instead of using maximum likelihood methods in classical logistic regression, Bayesian logistic regression applies Bayesian inference, a useful approach specifying the probability distribution for explanatory variables, to estimate parameters β . Bayesian Inference starts with formulating a prior probability distribution over the unknown parameters β , which summarizes a set of beliefs of knowledge in hand before any observations are taken, and uses Bayes' Rule that encapsulates the core of Bayesian inference to obtain a posterior probability distribution for these parameters β . Mathematically, Bayes' Rule represents how the conditional probability of observed data D given parameters β relates to the converse conditional probability of parameters β given observed data D :

$$p(\beta | D) = \frac{p(\beta, D)}{p(D)} = \frac{p(\beta)p(D | \beta)}{p(D)} \quad (2)$$

where $p(\beta, D)$ is a joint probability distribution for parameters β and observed data D ; $p(\beta)$ is a prior probability for parameters β ; $p(\beta | D)$ is a posterior probability for parameters β ; $p(D | \beta)$ is the likelihood function; and $p(D)$ is the probability distribution of observed data D .

Markov chain Monte Carlo (MCMC) simulation is a general computational method in Bayesian inference to obtain a sequence of random samples from a probability distribution. This method is based on drawing values of parameters β from approximate distributions, and then correcting those draws to better approximate the target posterior distribution, $p(\beta | D)$. The samples are drawn sequentially with the distribution of the sampled draws depending on the last drawn value, and these draws form a Markov chain. The key to this method's success is not the

Markov property but the approximate distributions that can be improved at each step in the simulation thus finally converging to the target posterior distribution (Gelman *et al* 2004). A plethora of literature exists on Bayesian inference and Gibbs sampling or Metropolis-Hastings Algorithm for sampling from a probability distribution. In fact, these two algorithms are special examples of Markov chain Monte Carlo simulation. Gibbs sampling generates a sequence of samples from a joint probability distribution of two or more random variables, and Metropolis-Hastings Algorithm draws samples from any probability distribution, only requiring a calculated function proportional to a proposal generating density.

Three main differences exist between Bayesian inference and classical statistics. First, the idea of prior probability does not exist in classical statistics, but it constitutes the basis of Bayesian inference. Moreover, the overall approach of making inference and interpretation of Bayesian inference and classical statistics are slightly different. Classical statistics fits distribution on the observed data and draw estimations of parameters from that. Bayesian inference starts with fitting distribution on the parameters, and derives updated posterior distribution of parameters from the prior beliefs. Furthermore, confidence intervals are constructed in classical statistics, while Bayesian inference uses credibility intervals. By controlling the Type I (or alpha) error level, a confidence level is determined and the confidence interval of parameters based on that confidence level can then be constructed. Since the posterior probability of parameters is derived in Bayesian inference, the credibility interval of parameters can be obtained to have a natural interpretation in terms of probabilities. Genkin *et al.* (2007) summarized two advantages of Bayesian inference over classical statistics that Bayesian inference can effectively avoid the overfitting problem, which is commonly encountered in the fields of gene expression and linguistics, caused by a limited number of observations with a large

size of variables; Bayesian inference, which does not require matrix inverse calculation, significantly increases computational efficiency and reduces the expense.

2.4.2 Application History

The term “Bayesian” referred to Tomas Bayes (1763) who first brought an early, incomplete mathematical idea of a problem in Bayesian inference. Ironically, Bayes had little impact on the development of Bayesian statistics; it was Pierre-Simon Laplace (1774) who pioneered and popularized Bayesian inference and used it to approach problems in celestial mechanics, medical statistics, reliability and jurisprudence. Laplace (1774) also summarized the way of using uniform prior distribution in Bayesian inference by stating the “principle of insufficient reason”, and introduced primitive versions of conjugate priors. The middle of the 20th century was an era of further development of Laplace’s ideas and a lot of research contributed to the construction of prior distributions and derived posterior distributions (Jeffreys 1939, Savage 1954, Jaynes 1957, and many others).

It was not until the discovery of Markov chain Monte Carlo simulation in the 1980s that solved computational problems that Bayesian inference experienced a dramatic growth of research and applications. Box (1980) reviewed the complementary roles of Bayesian in the model building and robustness process of the predictive distribution and of the posterior distribution. Smith and Spiegelhalter (1980) studied the forms of prior global and local prior specification to examine their respective roles as choice criteria among alternative linear models. West *et al.* (1985) used conjugate prior and posterior distributions for the exponential family

parameters in the development of dynamic Bayesian models for nonlinear, non-normal time series and regression problems in order to obtain a tractable sequential analysis with predictive distributions available in a standard form.

After 1990, Bayesian inference was widely accepted and used. Instead of asking what is Bayesian inference, the focus of analysts was on how to apply Bayesian inference. Because of the unique characteristics, i.e., easily accommodate information from previous studies, effectively set credibility intervals and efficiently avoid overfitting data, Bayesian logistic regression models are frequently used in fields as diverse as computational linguistics, public health, fishery, habitat suitability, and epidemiology. Xu and Akella (2008) proposed a Bayesian logistic regression algorithm incorporating relevance feedback information to actively select documents for user evaluation, and demonstrated the effectiveness of their approaches in the use of several TREC datasets. Marshall *et al.* (1994) applied Bayesian logistic model for risk assessment in coronary artery bypass grafting and conducted comparative studies of various statistical techniques showing that Bayesian model was among those with the highest predictive power. Hilborn *et al.* (1994) applied Bayesian estimation and decision analysis for an age-structured model using biomass survey data to assess the probability of fishery stock size. Smith (2010) discussed the advantages and challenges of Bayesian approaches applying in spatial searching and fishing location choices. Xu *et al.* (2010) constructed a two-stage model by combining a facility location model with Bayesian networks to identify optimal sensors locations and to infer the probability of the occurrence of a contamination event and the possible contamination source.

Chapter 3 Data and Analysis Methods

3.1 Variable Explanation

3.1.1 Three Study Groups

Biorefineries and bioenergy plants are defined as facilities that integrate biomass conversion processes and equipment to produce fuels, power, or value-added chemicals from biomass. Four main types of biomass include wood or agricultural residues, municipal solid waste, landfill gas and biogas, and alcohol fuels. Only twenty-nine wood-using bioenergy/biofuels plants and thirteen biorefineries that use agricultural residues are located within the 13-state Southeastern study region. Because a large number of traditional biomass-using facilities (i.e. primary wood processing mills, secondary wood processing mills, pulp and paper mills, and food processing mills) exist and similar factors are likely to affect the attractiveness and suitability of a site for both traditional and biomass-using plants, we consider three potential definitions for good locations for biorefineries and bioenergy/biofuels plants (at a 5-digit ZCTA² resolution):

Group I: All biomass-using mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants;

Group II: Pulp and paper mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants;

Group III: Food processing mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants.

² ZCTA is the abbreviation of U.S. Census Bureau ZIP Code Tabulation Areas

Group I includes primary wood processing mills, secondary wood processing mills, pulp and paper mills, wood-using bioenergy and biofuel plants, and biorefineries using agricultural residues. It follows a more modern planning view of total biomass-using management. Group II and Group III provide a more focused definition, and both contain a smaller sample of observations. As defined by Perlack *et al.* (2005), primary wood processing mills convert roundwood into other products and include sawmills, medium density fiberboard (MDF), oriented strand board (OSB), particleboard, plywood, veneer post, pole, piling, dealer, yard, energy and wood chips. Secondary mills utilize the products of primary mills. Examples include millwork, containers and pallets, buildings, furniture, flooring, paper and paper products. Secondary wood processing mills in this paper include not only mills processing the above products but also those that produce planed wood products, remanufactured wood products, trusses, moldings, kiln dried products, treated wood products, plants, decking, and siding. Pulp and paper mills are included in the Group I biomass-using facilities. Here we give pulp and paper mills special focus because traditional wood pulps and agricultural residues can be raw materials for the manufacture of paper (Wong 1996). Food processing mills are inventory centers storing harvested crops and can be possible sites to transform agricultural residues into other energy forms for consumption.

The locations (5-digit ZCTAs) of all Groups I, II, and III biomass-using facilities are displayed in Figures 3-1, 3-2, and 3-3, respectively. As shown in the figures, the geographic dispersion of all three groups is different. Each state has Group I mills, but Oklahoma, Texas, and Florida have a smaller quantity of Group I mills compared to the large volume of mills in other states. Figure 2 illustrates that Georgia has the highest concentration of Group II mills relative to other states, and Oklahoma does not have any Group II mills. Figure 3 shows that

each state has Group III mills, and most Group III mills are located along the Mississippi river and coastal ports.

3.1.2 The Response Variable

Three separate response variables are considered for modeling and ranking potential refinery sites for the above three predefined groups. The first response variable, $y_{i1} = 1$, if the i th 5-digit ZCTA has a biomass-using facility as defined in Group I. The second variable, $y_{i2} = 1$, is defined similarly based on the existence of a Group II. Given a Group III mill located in the i th 5-digit ZCTA, the third response variable will be defined as 1, $y_{i3} = 1$. Some 5-digit ZCTAs are not suitable for building biomass-using facilities due to their geographic and/or economic characteristics. If the i th 5-digit ZCTA has no land, no living trees, or is in a large metropolitan area, this 5-digit ZCTA is regarded as a “non-probable” location for biomass-using facilities in all three groups, and thus we code the response variable for the “non-probable” 5-digit ZCTAs as $y_i = 0$.

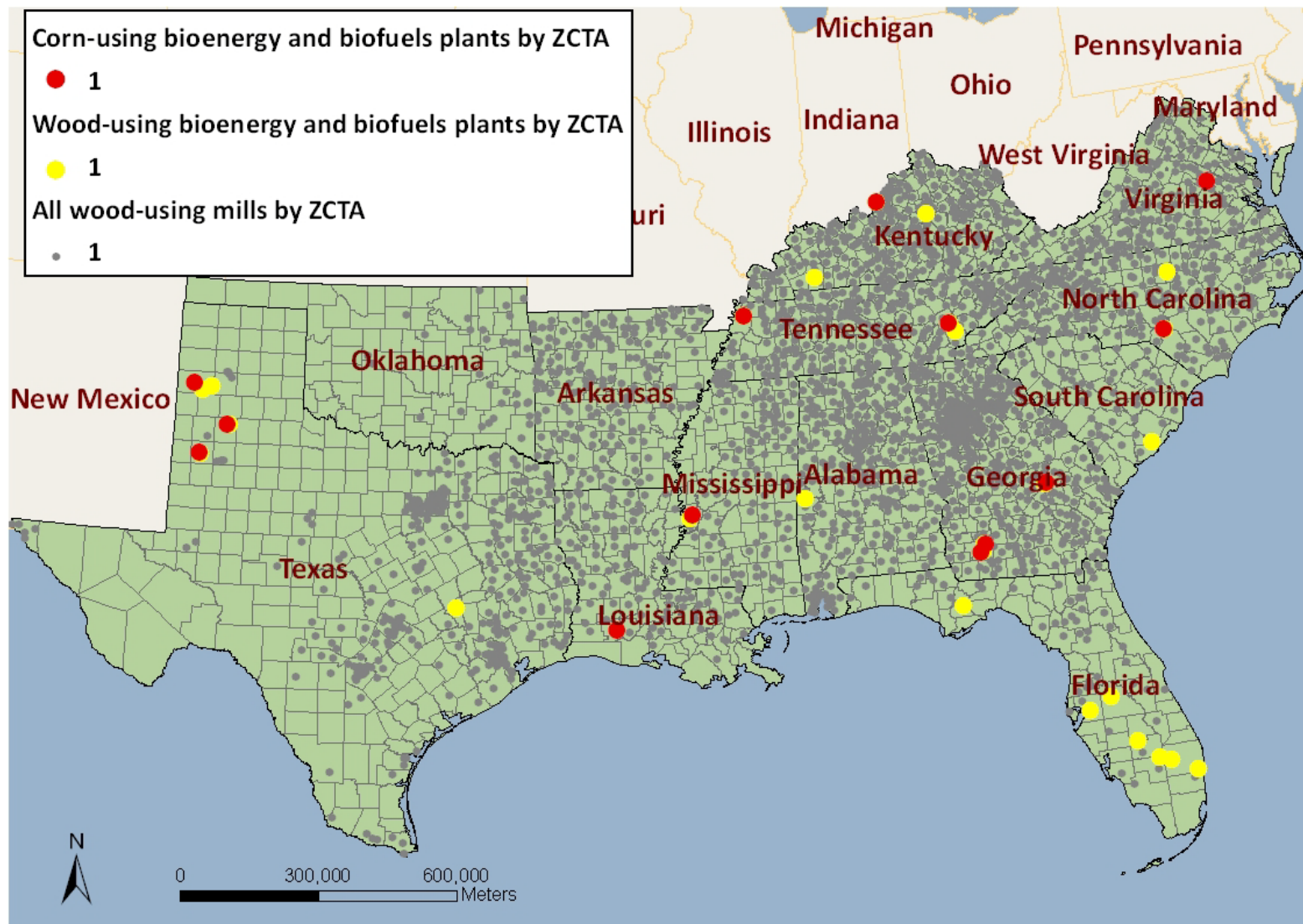


Figure 3-1 All biomass-using mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants in 13 Southeastern states

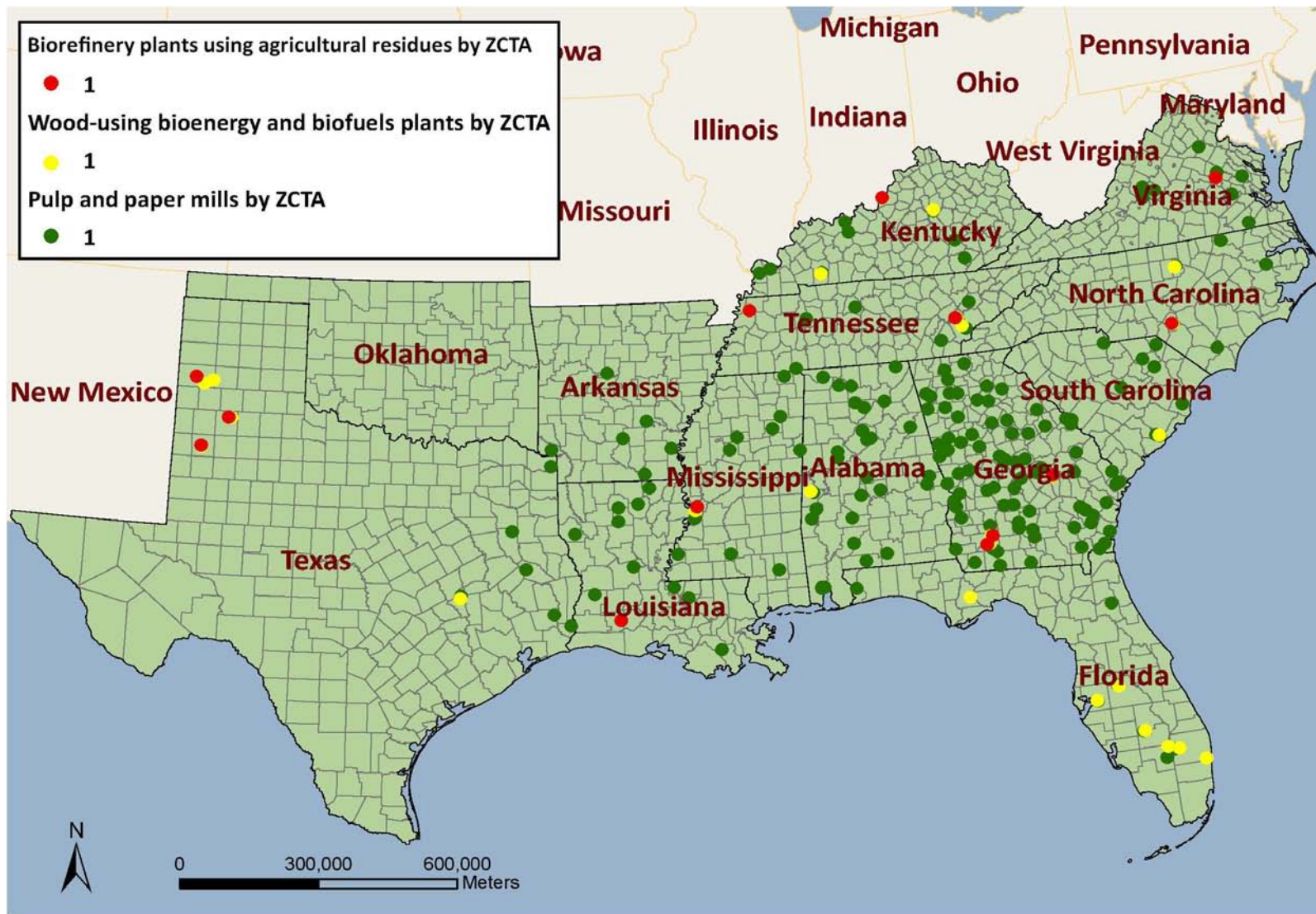


Figure 3-2 Pulp and paper mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants in 13 Southeastern states

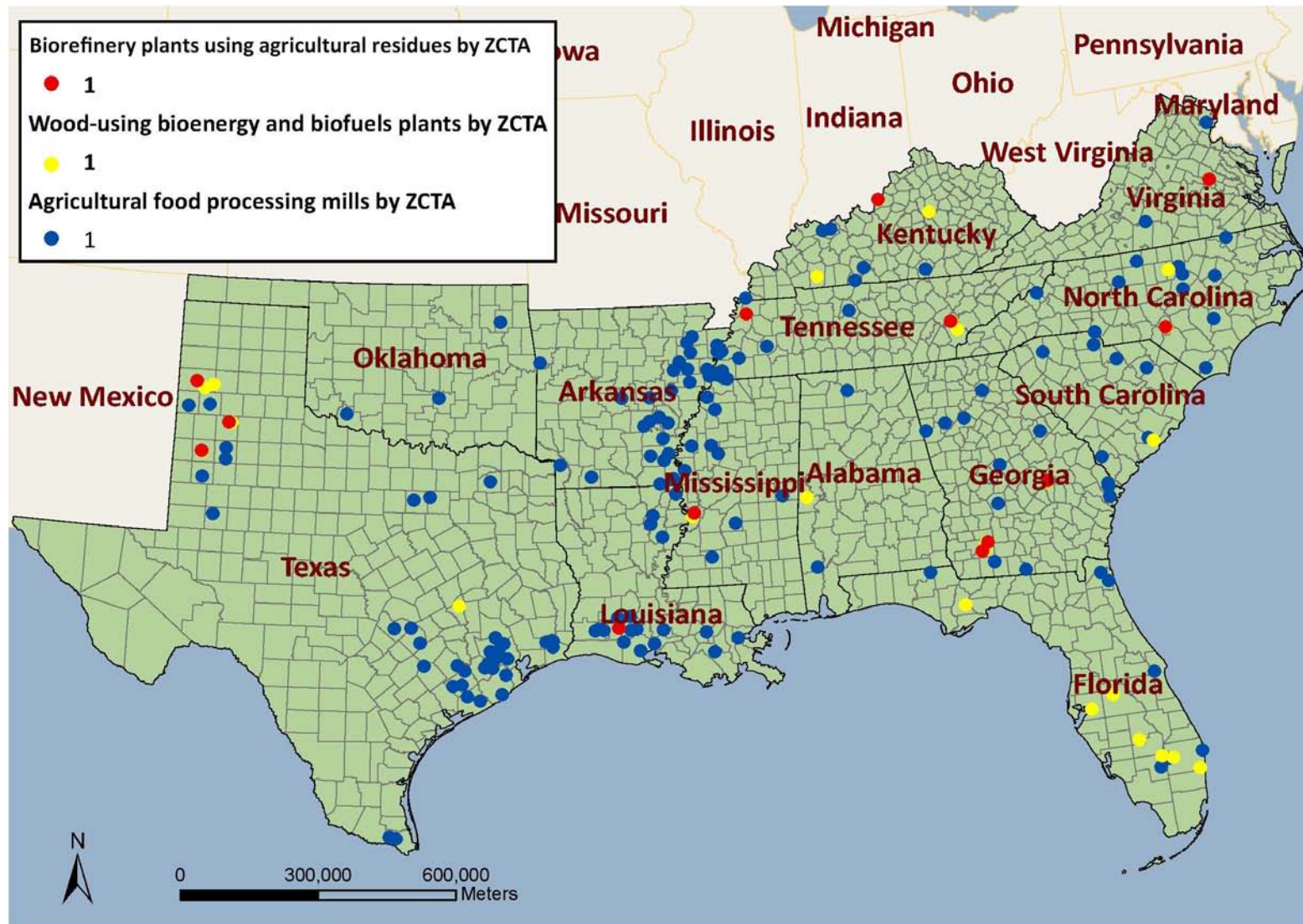


Figure 3-3 Food processing mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants in 13 Southeastern state

3.1.3 Explanatory Variables

Three groups of explanatory variables are categorized as economic/social variables, biomass availability variables, and transportation-related variables. A total 119 explanatory variables were used in this study (see Table 1 for a detailed explanation of units, collection levels, and variable types). For example, “Population Density” is a continuous variable standing for population density in the unit of people per square mile in each 5-digit ZCTA. Models of the Group I and Group III biomass-using facilities use the first 116 explanatory variables in Table 3-1. Models of Group II biomass-using facilities include the last additional three explanatory variables in Table 3-1 for the consideration of their potential impacts on the locations of Group II biomass-using facilities. The three explanatory variables are continuous and represent the number of primary wood processing mills, the number of secondary wood processing mills, and the number of other mills in each 5-digit ZCTA, labeled as “Primary_Mill_No”, “Secondary_Mill_No”, and “Other_Mill_No” respectively.

Table 3-1 Explanatory variables for three groups of biomass-using facilities

Variable name	Variable Type	Collection level	Unit	Explanation
Population	Continuous	5-digit ZCTA	People	Population in each 5-digit ZCTA
Population_Density	Continuous	5-digit ZCTA	People/miles	Population density in each 5-digit ZCTA
Employment	Continuous	5-digit ZCTA	People	Employed person in all industries in each 5-digit ZCTA
Median_Family_Income	Continuous	5-digit ZCTA	Dollar	Median of family income in 1999 in each 5-digit ZCTA
Income_Index	Continuous	5-digit ZCTA	Dollar/people	Median of family income in 1999 per employed person in each 5-digit ZCTA
Sqmiwater	Continuous	5-digit ZCTA	Mile ²	Water area within 5-digit ZCTAs
Barley_RES_Qty	Continuous	County	Dry tons	Agricultural materials left in an agricultural field after the crop has been harvested. Residues in every county are from Barley Straw, Corn Stover for silage and for grain, Oat Straw, Sorghum Straw, Wheat All Straw, and Wheat Winter All Straw.
Corn_RES_Qty				
Oat_RES_Qty				
Sorghum_RES_Qty				
WheatAll_RES_Qty				
WheatWinter_RES_Qty				
Barley_RES_HC	Continuous	County	Dollar/dry tons	Harvesting costs of agricultural residues in each county. Agricultural residues are from Barley Straw, Corn Stover for silage and for grain, Oat Straw, Sorghum Straw, Wheat All Straw, and Wheat Winter All Straw.
Corn_RES_HC				
Oat_RES_HC				
Sorghum_RES_HC				
WheatAll_RES_HC				
WheatWinter_RES_HC				
Barley_40_MCost	Continuous	ZIP Code	Dollar/mile	Marginal trucking cost of total Barley Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Barley_80_MCost				
Barley_120_MCost				
Barley_160_MCost				
Corn_40_MCost	Continuous	ZIP Code	Dollar/mile	Marginal trucking cost of total Corn Stover residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Corn_80_MCost				
Corn_120_MCost				
Corn_160_MCost				
Oat_40_MCost	Continuous	ZIP Code	Dollar/mile	Marginal trucking cost of total Oat Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Oat_80_MCost				
Oat_120_MCost				
Oat_160_MCost				

Table 3-1 (Continued)

Variable name	Variable Type	Collection level	Unit	Explanation
Sorghum_40_MCost	Continuous	ZIP Code	Dollar/ mile	Marginal trucking cost of total Sorghum Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Sorghum_80_MCost				
Sorghum_120_MCost				
Sorghum_160_MCost				
WheatAll_40_MCost	Continuous	ZIP Code	Dollar/ mile	Marginal trucking cost of total Wheat All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatAll_80_MCost				
Wheat All_120_MCost				
Wheat All_160_MCost				
WheatWinter_40_MCost	Continuous	ZIP Code	Dollar/ mile	Marginal trucking cost of total Wheat Winter All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatWinter_80_MCost				
WheatWinter_120_MCost				
WheatWinter_160_MCost				
Barley_40_TCost	Continuous	ZIP Code	Dollar/ mile	Total trucking cost of total Barley Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Barley_80_TCost				
Barley_120_TCost				
Barley_160_TCost				
Corn_40_TCost	Continuous	ZIP Code	Dollar/ mile	Total trucking cost of total Corn Stover residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Corn_80_TCost				
Corn_120_TCost				
Corn_160_TCost				
Oat_40_TCost	Continuous	ZIP Code	Dollar/ mile	Total trucking cost of total Oat Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Oat_80_TCost				
Oat_120_TCost				
Oat_160_TCost				
Sorghum_40_TCost	Continuous	ZIP Code	Dollar/ mile	Total trucking cost of total Sorghum Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Sorghum_80_TCost				
Sorghum_120_TCost				
Sorghum_160_TCost				
WheatAll_40_TCost	Continuous	ZIP Code	Dollar/ mile	Total trucking cost of total Wheat All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatAll_80_TCost				
WheatAll_120_TCost				
WheatAll_160_TCost				

Table 3-1 (Continued)

Variable name	Variable Type	Collection level	Unit	Explanation
WheatWinter_40_TCost	Continuous	ZIP Code	Dollar/ mile	Total trucking cost of total Wheat Winter All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatWinter_80_TCost				
WheatWinter_120_TCost				
WheatWinter_160_TCost				
Barley_40_ACost	Continuous	ZIP Code	Dollar/ mile	Average trucking cost of total Barley Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Barley_80_ACost				
Barley_120_ACost				
Barley_160_ACost				
Corn_40_ACost	Continuous	ZIP Code	Dollar/ mile	Average trucking cost of total Corn Stover residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Corn_80_ACost				
Corn_120_ACost				
Corn_160_ACost				
Oat_40_ACost	Continuous	ZIP Code	Dollar/ mile	Average trucking cost of total Oat Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Oat_80_ACost				
Oat_120_ACost				
Oat_160_ACost				
Sorghum_40_ACost	Continuous	ZIP Code	Dollar/ mile	Average trucking cost of total Sorghum Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Sorghum_80_ACost				
Sorghum_120_ACost				
Sorghum_160_ACost				
WheatAll_40_ACost	Continuous	ZIP Code	Dollar/ mile	Average trucking cost of total Wheat All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatAll_80_ACost				
WheatAll_120_ACost				
WheatAll_160_ACost				
WheatWinter_40_ACost	Continuous	ZIP Code	Dollar/ mile	Average trucking cost of total Wheat Winter All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatWinter_80_ACost				
WheatWinter_120_ACost				
WheatWinter_160_ACost				
Barley_40_TQty	Continuous	ZIP Code	Dollar/ mile	Cumulative quantity of Barley Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Barley_80_TQty				
Barley_120_TQty				
Barley_160_TQty				

Table 3-1 (Continued)

Variable name	Variable Type	Collection level	Unit	Explanation
Corn_40_ TQty	Continuous	ZIP Code	Dollar/ mile	Cumulative quantity of Corn Stover residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Corn_80_ TQty				
Corn_120_ TQty				
Corn_160_ TQty				
Oat_40_ TQty	Continuous	ZIP Code	Dollar/ mile	Cumulative quantity of Oat Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Oat_80_ TQty				
Oat_120_ TQty				
Oat_160_ TQty				
Sorghum_40_ TQty	Continuous	ZIP Code	Dollar/ mile	Cumulative quantity of Sorghum Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
Sorghum_80_ TQty				
Sorghum_120_ TQty				
Sorghum_160_ TQty				
WheatAll_40_ TQty	Continuous	ZIP Code	Dollar/ mile	Cumulative quantity of Wheat All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatAll_80_ TQty				
WheatAll_120_ TQty				
WheatAll_160_ TQty				
WheatWinter_40_ TQty	Continuous	ZIP Code	Dollar/ mile	Cumulative quantity of Wheat Winter All Straw residues within a 40-mile, 80-mile, 120-mile, and 160-mile haul distances.
WheatWinter_80_ TQty				
WheatWinter_120_ TQty				
WheatWinter_160_ TQty				
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 number means that the ZIP Code has more railroad access.
No_ports	Continuous	5-digit ZCTA	Port	Number of water ports in each 5-digit ZCTA
Primary_Mill_No	Continuous	5-digit ZCTA	Mill	Number of primary wood processing mills in each 5-digit ZCTA
Secondary_Mill_No	Continuous	5-digit ZCTA	Mill	Number of secondary wood processing mills in each 5-digit ZCTA

Table 3-1 (Continued)

Variable name	Variable Type	Collection level	Unit	Explanation
Other_Mill_No	Continuous	5-digit ZCTA	Mill	Number of other mills in each 5-digit ZCTA. Other mills include forestry companies, logging mills, and companies that provide equipment and supplies, such as logging machine rental companies (Perlack <i>et al.</i> 2005).

3.2 Data Management

3.2.1 Data Sources

This research involves large volumes of data collected from various sources, including U.S. Census Bureau (2000), U.S. Forest Service (Perlack *et al.* 2005), U.S. National Land Cover Dataset (2000), various internet sources (2009), railroad companies (Burlington Northern Santa Fe Railway, CSX Corporation, Inc., Norfolk Southern System, and Union Pacific), and other transportation firms³. Another important data source is the BioSAT model (see www.BioSAT.net). The BioSAT database includes measurements of marginal cost, average cost, and total cost of delivered total mill residues within 40-mile, 80-mile, 120-mile, and 160-mile haul distances, and total quantity of available mill residues within these distances. These records are summarized at the level of U.S. Census Bureau 5-digit ZIP Code tabulation area (5-digit ZCTA).

³ Personal communication by Liu (2009): 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.

Various computer software packages are used for data generation, standardization, verification, and combination. SAS[®] 9.1 Base and SAS[®] PROC SQL are used for organizing data from various sources, for merging data sets containing different explanatory variables, for standardizing data sets to be easily analyzed, and for verifying data quality. JMP[®] 8.0.1 and Microsoft[®] Excel 2007 are used as supplementary tools for data management and verification. MATLAB[®] and ArcGIS[®] are employed to estimate agricultural residue quantity at the level of 5-digit ZCTAs, to calculate the driving time and driving distances of 5-digit ZCTA pairs within 40-mile, 80-mile, 120-mile, and 160-mile haul distances, which are used to compute the cost and quantity data in the BioSAT SQL server database, and to produce the final maps.

Depending upon the sources, data were collected at the ZIP Code, ZCTA, City, or County level. Data sets with the same level of resolution are merged directly. Data sets with different levels of resolution are merged based on the corresponding geographical relationship between the hierarchical structure of ZIP Code, ZCTA, City, and County. Missing values are surrogated after the data merging. The final data set in the study is at the level of 5-digit ZCTAs. To maintain acceptable data quality and consistency, 5-digit ZCTA level data follows the guidelines of the U.S. Census Bureau.

The whole data is first randomly partitioned into two sets: 60% used for classical approach and 40% used for Bayesian approach. For both the classical and Bayesian approaches, a training set of data is used plus a separate validation (hold out) set of data using a pseudo-random number-generating function in SAS[®] Enterprise Miner. The training data are used to develop the models while the validation data are used to evaluate the model performance.

3.2.2 Biomass Estimation Using GIS

Raw agricultural residue quantity data were collected at the county level (Table 3-1). To ensure data quality and maintain a consistent level, agricultural residues in 5-digit ZCTAs are estimated both statistically and geographically using Geographic Information System (GIS). We use national land cover information data set (U.S. National Land Cover Dataset 2000) and digital maps to identify arable land areas for crops in each 5-digit ZCTA, and then proportionally calculate agricultural residues quantity and derive the harvesting costs for each 5-digit ZCTA.

As shown in Figure 3-4, the map layer of agricultural residue quantity (USDA NASS 2008) in the county level is first overlapped with the 5-digit ZCTA boundary map layer. By applying the “Contain” spatial function in ArcGIS®, 5-digit ZCTAs with agricultural residue quantity that locate in the counties are selected. Implementing “Identity” function, each agricultural residue county is split into multiple area parts via the 5-digit ZCTA boundary shape, and each area part is identified by a unique identifier for its belonging 5-digit ZCTA. Through overlapping each area part with the land cover information layer, the number of pixels in all land cover classes is extracted for each area part based on its geographical shape. Cultivated crops, a land cover class, is chosen in the analysis because all types of agricultural crops are grown on this type of arable land (Pimentel *et al.* 1981). By aggregating the cultivate crop pixels in the unit of county, a pixel ratio for each area part is calculated and the agricultural residue quantity in every area part can be derived by this pixel ratio. So far, each area part has been attached a proportional agricultural residue quantity of its belonging county. Through the identifier in each area part, multiple area parts can be assembled if they belong to the same 5-digit ZCTA and share the same identifier. Finally, all agricultural residue quantity of area parts are summed up for every 5-digit ZCTA and the harvesting costs can be derived for each 5-digit ZCTA.

Figure 3-5 and Figure 3-6 illustrate the geographical concentration of aggregated agricultural residue (Barley Straw, Corn Stover for Silage and for Grain, Oat Straw, Sorghum Straw, Wheat All Straw, and Wheat Winter All Straw) quantity in yellow and harvesting costs in blue at the level of 5-digit ZCTA in 13 Southeastern states. Darker color represents more residue quantity and higher harvesting costs. Note that Alabama, Florida, Georgia, and Louisiana have the lowest concentration of agricultural residues compared to other states. The highest harvesting costs are from the middle Texas region, and the largest quantity of agricultural residue is from northwestern Oklahoma and middle Alabama.

3.2.3 Transportation-related Data Generation

In order to gather sufficient feedstocks to satisfy the demand of a biorefinery, it is necessary to collect residues from neighboring 5-digit ZCTAs within certain trucking distances, and transport these residues to the demand 5-digit ZCTA where the potential biorefinery is located. Transportation cost, which represents a substantial portion of biomass total cost, is essential for evaluating the economic feasibility of bioenergy. Variables like marginal cost, average cost, and total cost can be generated based on the BioSAT model (Young *et al* 2009). The SQL server database for BioSAT contains 82 million transportation cost records for 33 Eastern states. This database contains real driving time and driving distance for every pair of ZCTAs in the 13 Southeastern states for up to a 200-mile one-way haul distance. This database was created by first finding the list of neighboring ZCTAs within a required driving distance (e.g. 40 miles, 80 miles, 120 miles, and 160 miles) of each ZCTA, and then calculating the values of marginal cost, average cost, total cost, and total quantity for the demand ZCTA.

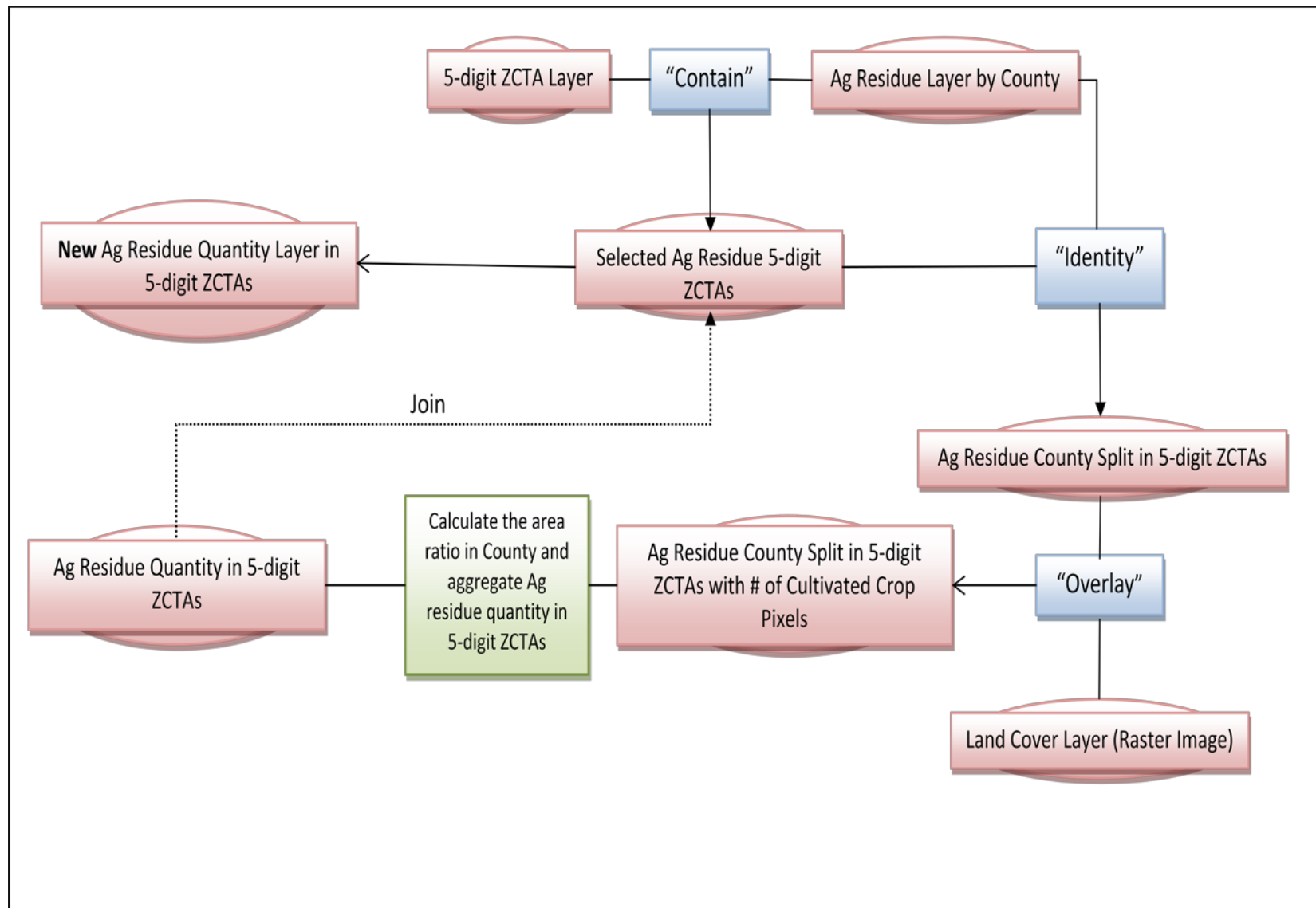


Figure 3-4 Flow Chart of estimating agricultural residues at the level of 5-digit ZCTA

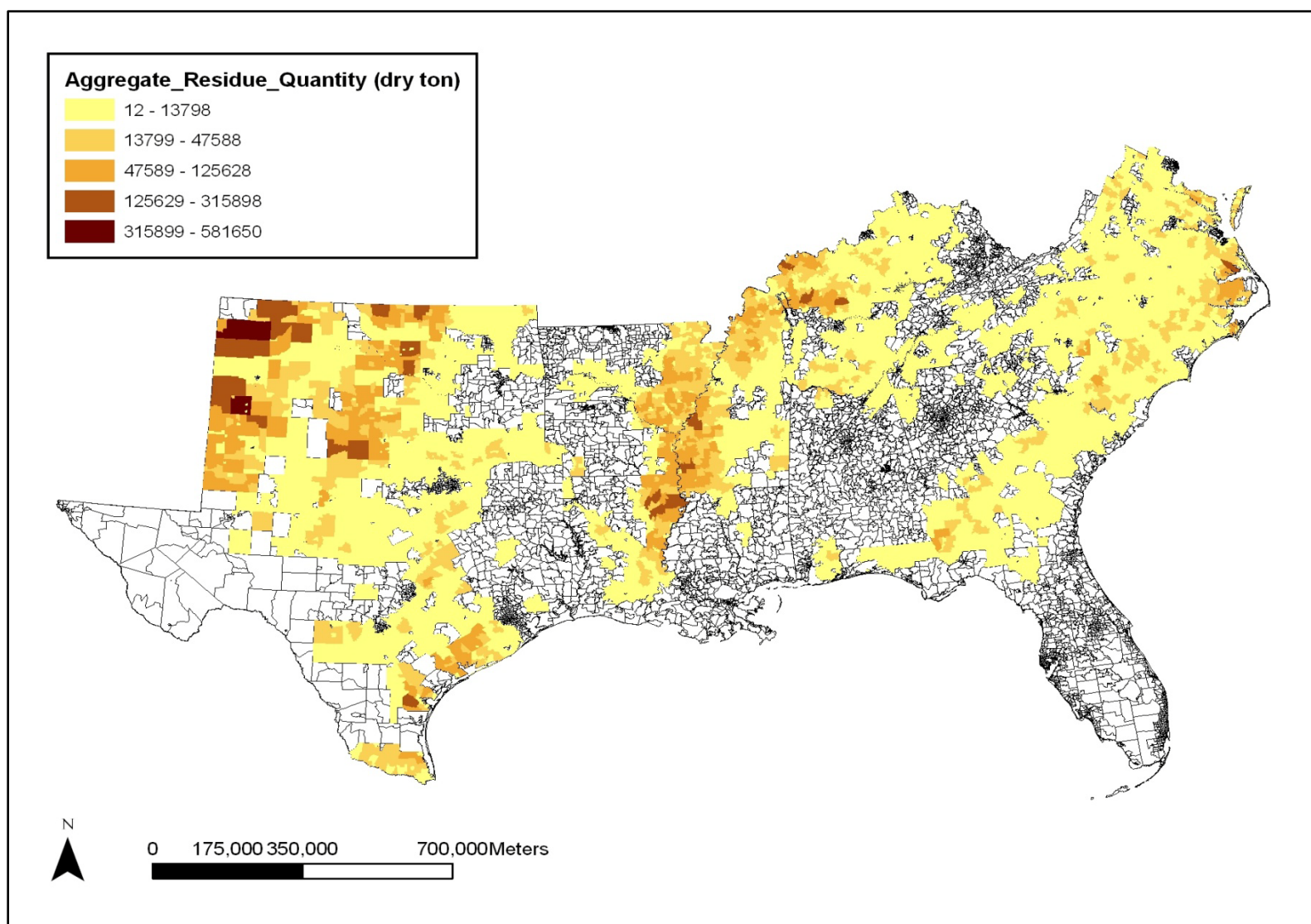


Figure 3-5 Agricultural residue quantity at the level of 5-digit ZCTA in 13 Southeastern states

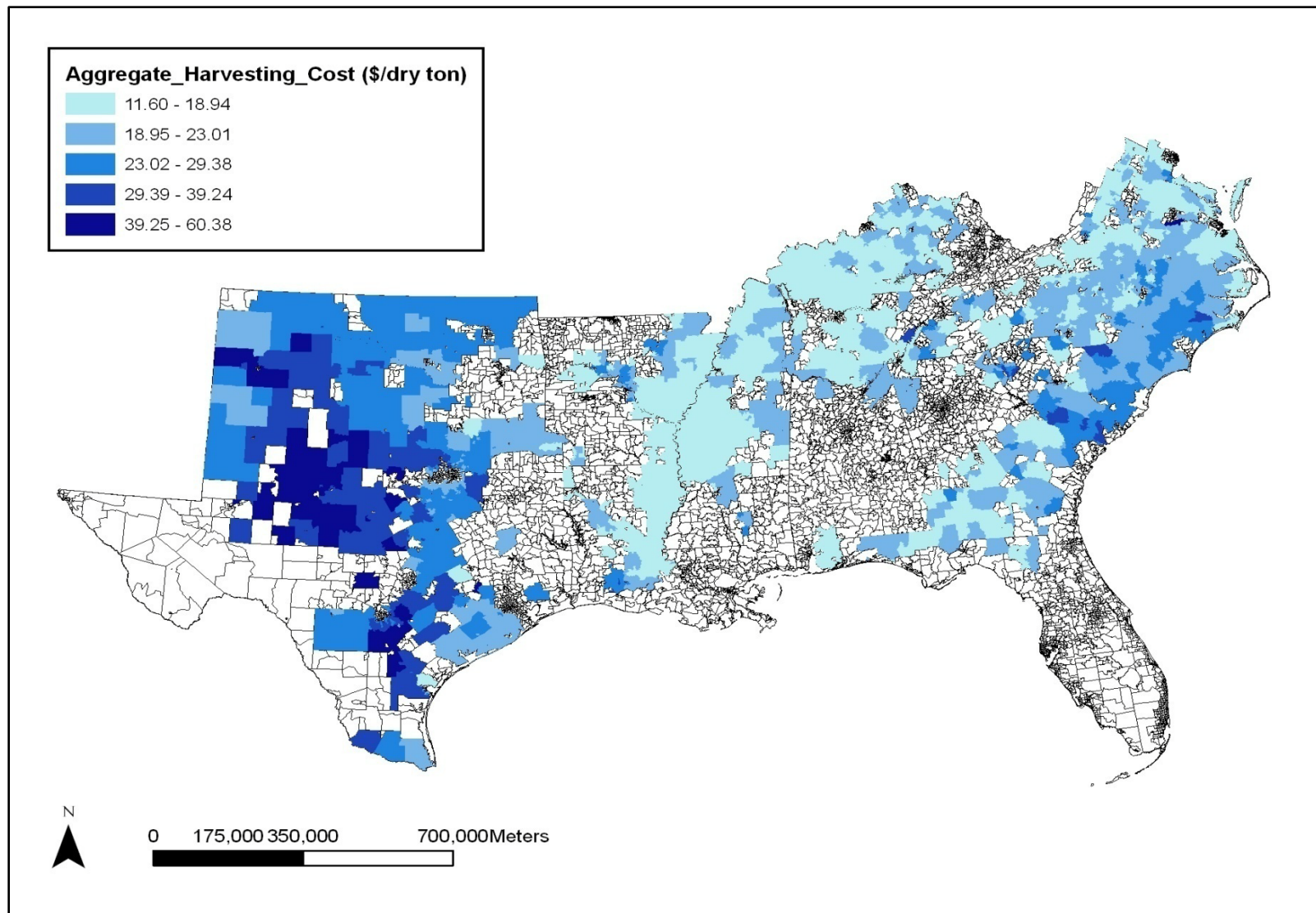


Figure 3-6 Agricultural residue harvesting costs at the level of 5-digit ZCTA in 13 Southeastern states

Since calculating the driving distance between each possible pair of ZCTAs within any possible bio-supply basin is time consuming and computationally intensive, setting a sphere distance from a given ZCTA to another one helps to efficiently estimate this metric. Sphere distance is expressed as in equation (3), which uses the latitude and longitude of every ZCTA pair (Moritz 2000 and Wang 2008),

$$D = \sqrt{(Md\phi)^2 + (N \cos \phi d\lambda)^2} \quad (3)$$

where ϕ is mean latitude; $d\phi$ is difference in latitude; $d\lambda$ is difference of longitude in radius; M is Earth's radius of curvature in the north-south meridian at ϕ ; and N is radius of curvature in the prime normal to M at ϕ .

The following procedure is used to locate potential neighboring ZCTAs that have a sphere distance of no more than the required driving distance (e.g. 40 miles, 80 miles, 120 miles, and 160 miles). Micorsoft[®] MapPoint 2006 is then used to estimate the real driving time and distance between these potential neighboring ZCTAs and the demand ZCTAs. Those potential ZCTAs, whose real driving distances are no more than the required distance (e.g. 40 miles, 80 miles, 120 miles, and 160 miles) are reserved as the nearest neighboring ZCTAs for the given ZCTA. This procedure repeats until all neighboring ZCTAs are found. The marginal cost, average cost, total cost, and total quantity from each supply ZCTA for a given demand ZCTA within the required bio-basin distance is generated. Figure 3-7 shows an example of geographic distributions of supply agricultural residues from neighboring ZCTAs to a demand ZCTA (75013) within 40-mile, 80-mile, 120-mile and 160-mile haul distances. Darker color here represents larger agricultural residue quantities from supply neighboring ZCTAs that can transport residues to the demand ZCTA within the required driving distance.

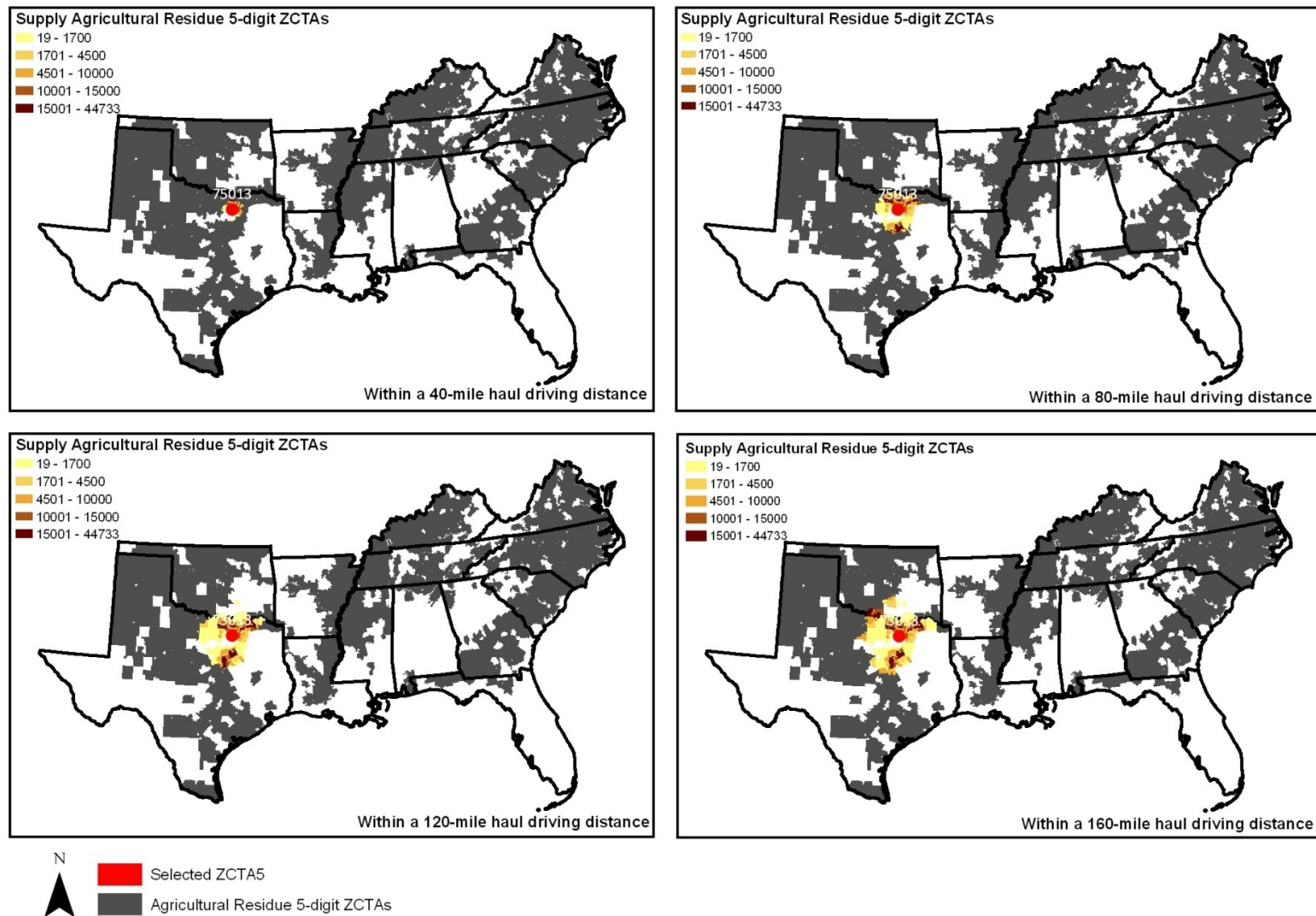


Figure 3-7 Supply agricultural residues from neighboring ZCTAs to a demand ZCTA within 40-mile, 80-mile, 120-mile, and 160 mile haul driving distances

3.2.4 Data Standardization and Treatment of Missing Data

Since different data sources have different scales and resolution, all raw data were first standardized. The standardization equation is given in equation (4) below where each individual observation is subtracted from the mean value and this difference is divided by the standard deviation. All standardized values are used in the thesis analysis.

$$x_{i,\cdot} = \frac{x_{i,j} - \text{mean}(\sum_{j=1}^n x_{i,j})}{\text{std dev}(\sum_{j=1}^n x_{i,j})} \quad (4)$$

A large number of missing data are observed in agricultural availability variables. For instance, certain 5-digit ZCTAs do not have agricultural residue quantities, or some particular 5-digit ZCTAs are inaccessible to others. To avoid ambiguous meaning for missing data, a “0” is assigned to missing data for quantity-related variables, e.g., Population, Sqmiwater, Employment, Population_Density, Median_Family_Income, Income_Index, and Residue_Qty variables for types of agricultural residues. When merging transportation-related variables at a 5-digit ZCTA level, several different methods shown in Table 3-2 are used in accordance with their distinctive features. Marginal cost is the change in total cost that arises when one additional unit of agricultural residue quantity is supplied to the demand 5-digit ZCTA. The minimum marginal cost for a supply ZCTA is the first selected supply ZCTA.

Table 3-2 Combination value for variables at 5-digit ZCTAs and surrogating number for missing values

Variables	Combination Value	Surrogating Number for Missing Values
Marginal Cost	Minimum value	9999
Average Cost	Average value	9999
Total Cost	Average value	9999
Total Quantity	Average value	0

3.3 Bayesian Logistic Regression Model

3.3.1 Logistic Regression

Logistic regression models the relationship between a two-level categorical response variable (binary response) and the explanatory variables which can be continuous or categorical (nominal/ordinal). Suppose that the response variable y is binary and the outcomes are coded $y=1$ or $y=0$ with respective probability p and $1-p$. For a particular subject, the logistic regression equation is expressed as

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (5)$$

where $x_1, x_2, \dots, x_n$ are the subject's feature measurements on a group of explanatory variables, and $\beta_1, \beta_2, \dots, \beta_n$ are parameter estimates for explanatory variables. With the *logit* function, it relates the explanatory variables to the probability of an outcome $y=1$.

There are two advantages of logistic regression models. First, logistic regression is a flexible instrument to analyze a mixed set of continuous and categorical (nominal/ordinal) data

(Hosmer and Lemeshow 1989). In logistic regression, continuous variables are easily included in the model, while categorical variables can also be accommodated as dummy or product variables. In contrast, other models like loglinear models are restricted to categorical data relying on nominal and ordinal variables (Agresti 1990). Moreover, logistic regression has no limitations on the distribution of explanatory variables. It's not necessary for explanatory variables to be normally distributed, to be linear related, nor to have equal variance within each group. Since the response variable is either "1" or "0" with probability of p and $1-p$ respectively, the function of an explanatory variable on the response variable is not linear. Instead, logistic regression uses a logarithm function on the odds ratio $\frac{p}{1-p}$ to transform a range of response results to be real numbers.

The stepwise logistic regression can be implemented for the data set of biorefineries using wood or agricultural residues to determine automatically which explanatory variables play a significant role in investigating the quality of a potential site. A variable transformation procedure can be used to preprocess the data set and may improve the predictive ability of the logistic regression. Liu (2009) built up four distinctive models to reflect different ways of employing these two procedures (Figure 3-4):

Model 1: Logistic regression without variable selection and variable transformation;

Model 2: Logistic regression with stepwise variable selection;

Model 3: Logistic regression with variable transformation;

Model 4: Logistic regression with both variable selection and variable transformation.

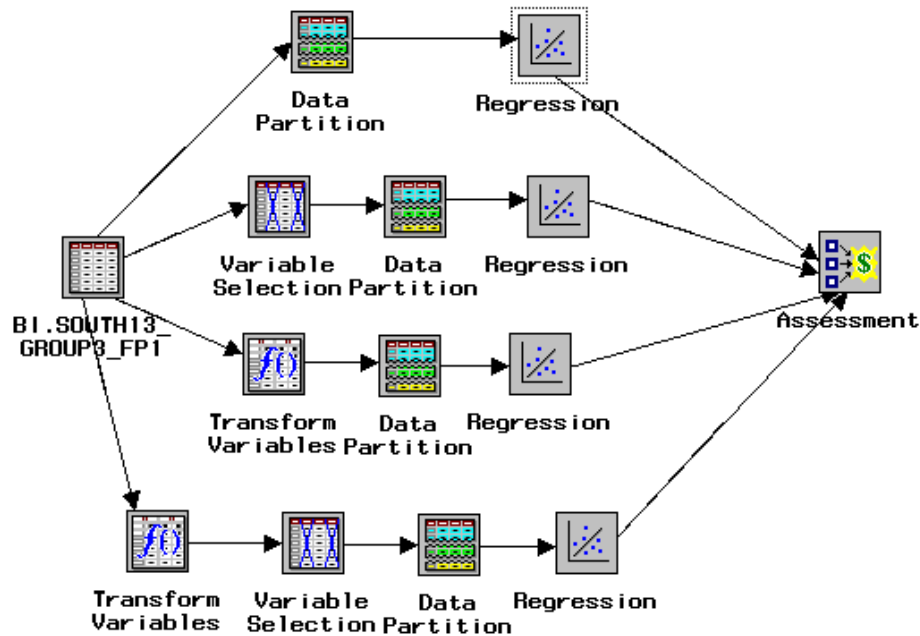


Figure 3-8 Four distinct models in SAS[®] Enterprise Miner

The full model included in the variable selectin contains a full set of explanatory variables and all two-way interactions. In the variable selection procedure, the explanatory variables and two-way interactions that do not have a R-square improvement of 0.0005 to the response variable are removed.

In each model, the data set is partitioned into two parts: 60% of the data is randomly selected as the training set and the rest 40% of the data is the validation set. The training data set is used to develop the stepwise logistic regression models; and the validation data is used to evaluate the model performance by constructing the classification table.

3.3.2 Model Selection Methods and Criteria

The goal of selecting a good logistic regression model is to correctly predict an outcome using the most parsimonious model (Liu 2009). Three common model selection methods can be used to balance the goodness-of-fit with simplicity. Forward selection adds explanatory variables sequentially until further addition do not improve the fit. Backward elimination begins with a complex model and sequentially remove variables. Stepwise selection is a combination of forward selection and backward elimination, allowing variables to enter or exit the model. Although the above three methods help to build up a simple model that fits adequately to the current data set, they are open for overfitting the data at the same time.

Several criteria can help protect against the danger of overfitting that emerges from model selection procedure. The criteria used in this study are the Akaike information criterion (AIC) and Bayesian information criterion (BIC), which are defined as:

$$\begin{aligned} AIC &= -2\ln(L_m) + 2m \\ BIC &= -2L_m + m\ln(n) \end{aligned} \tag{6}$$

where m is the number of parameters in the model, L_m is the maximized *log-likelihood* of the model, and n is the sample size. Both AIC and BIC take into account not only the statistical goodness-of-fit, but also a penalty that is an increasing function of the number of estimated parameters, which discourages overfitting the model. The BIC penalizes free parameters more strongly than does the AIC. The optimal model is the one that tends to have its fitted values closest to the true outcome probabilities. This is the model that minimizes AIC or BIC criterion.

3.3.3 Bayesian Inference

Bayesian logistic regression extends logistic regression into a Bayesian framework (Xu and Akella 2008). Instead of using maximum likelihood methods in logistic regression, it applies Bayesian inference, a useful approach specifying the probability distribution for the underlying categorical or continuous variables, to estimate parameters β . Bayesian inference, which allows ready incorporation of prior beliefs and the combination of such beliefs with statistical data, is well suited for representing the uncertainties in the value of explanatory variables (Jaakkola and Jordan 1996). In practice, Bayesian inference expresses the uncertainty of parameters in terms of probability distributions named as prior probability distributions, and derives posterior estimate probability distributions from it. For example, by expressing the uncertainty in parameter vector β for a given model M as the posterior probability distribution $p(\beta | M, D)$, where D is the observed data, we have

$$p(y = 1 | x, M, D) = \int_{\beta} p(y = 1, \beta | x, M, D) d\beta = \int_{\beta} p(y = 1 | x, \beta, M) p(\beta | M, D) d\beta \quad (7)$$

where

$$p(y = 1 | x, \beta, M) = \frac{1}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)} \quad (8)$$

The key of Bayesian inference is to choose the parametric family for prior probability distributions. Two categories are suggested: non-informative prior distributions and informative prior distributions. A Non-informative prior distribution expresses vague or general information about a parameter. The common non-informative prior distribution is uniform distribution,

which assigns equal probabilities to all value possibilities, and always yields similar results as classical statistics. Thus, Bayesian and classical statistics are not exclusive and they are overlapped to some extent. In fact, classical approaches are approximate Bayesian using certain priors. An informative prior distribution reflects specific and definite information about a parameter. If both of prior and posterior distributions belong to the same family, the prior distribution is called conjugate prior distribution, which is a special case in informative prior distributions. Four different prior distributions selected from non-informative and informative prior distributions are constructed on parameters β :

Prior 1: Uniform prior distribution $p(\beta) \propto \text{constant}$ (9)

Prior 2: Gaussian prior distribution $p(\beta | u, \sigma^2) \propto \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(\beta - u)^2}{2\sigma^2})$ (10)

Prior 3: Laplace prior distribution $p(\beta | u, \lambda) \propto \frac{\lambda}{2} \exp(-\lambda|\beta - u|)$

(11)

Prior 4: Cauchy prior distribution $p(\beta | a, b) \propto \frac{1}{\pi b [1 + (\frac{\beta - a}{b})^2]}$ (12)

Using the maximum likelihood estimates as our prior beliefs, the four predefined prior probability distributions of parameter β are built up. The posterior probability distribution under each prior is obtained by conducting the Markov chain Monte Carlo simulation. The expected value of the posterior probability distribution of parameter β will be treated as regression coefficients of Bayesian logistic models.

Two statistical software packages are used for Bayesian inference analysis. WinBUGS[®] 3.0.2 (Windows Bayesian Inference Using Gibbs Sampling) and R[®] 2.10.0 provide a convenient environment to conduct a Markov chain Monte Carlo simulation of a random walk in the space of parameters β , which converges to a stationary distribution approximating the joint distribution. In each analysis, one independent chain is run for 10,000 iterations, and convergence is assessed by visual inspection of the chain. In addition to providing a posterior estimate of the mean $\hat{\beta}$, the stationary distribution also provides a posterior standard deviation, the median and a 95% credibility interval of $\hat{\beta}$ via the estimated covariance for β .

3.3.4 Classification Trees with Local Bayesian Logistic Regression

Classification trees (CTs) are widely used in machine learning for looking at hierarchical models of relationship between explanatory variables and a categorical response variable. CTs are non-parametric models with very high expressive power which provide a symbolic representation for easy interpretation by humans (White and Sifneos 2002). The idea behind CTs is to use variables to partition the whole dataset into several homogeneous subsets with respect to the same class. The CT has a hierarchical tree structure in which a node denotes a splitting variable, the branches of a node denote value or value ranges of the correspondingly splitting variable, and a leaf denotes a class. The CT construction is build through a binary recursive partitioning process starting from the whole dataset. Each partitioning is based upon a split value of some variables. For example, $[\leq 10,000, > 10,000]$ is a possible split value of the population variable that implies partitioning a given study group in two (population) subsets accordingly. In each recursive partitioning iteration, the aim is to find a variable, along with its

split value, that can result in subsets having maximal homogeneity in their classes. The impurity measure of classes in a given set is the entropy of the set with respect to the classes. In this research, the classification tree approach is only used to identify subsets of study groups and will not be used to directly predict outcomes. JMP[®]8.0.1 was selected to construct CTs in this research, which searches for the splitting variables, by minimizing the residual log-likelihood chi-square reported as G^2 , to maximally discriminate between classes.

Inspired from the hybrid model proposed by Abu-Hanna and Keizer (2003) that integrates classification trees with local logistic regression to improve the predictive performance, this research experiments the hybrid model, but applies Bayesian inference instead of using maximum likelihood methods for the estimates in local logistic regressions. The strategy underlying the hybrid model is first to use CTs partitioning the whole datasets in several manageable subsets (the number of subsets is no more than 10), and then to fit a local logistic regression in each subset to get a probability estimate of siting biomass-using facilities. In the second step, we executed the four predefined modeling approaches on each subset to examine the statistically significant variables, and used Bayesian inference assuming four above different priors to get regression estimates of significant variables. Along with using entropy as the splitting criterion to get healthy subsets, we also controlled the number of observations in each subset greater than 50 to avoid overfitting data.

3.3.5 Model Assessment

3.3.5.1 Lift Charts

Cumulative gains and lift charts are visual aids for assessing and measuring model predictive power. Lift charts, which seek to compare the ratio between the results obtained with and without a predictive model, measure the effectiveness of that predictive model. Lift values, which represents the percentage of predicted values, are calculated as the percentage of true positive responses over the percentage of samples drawn from the raw data set. Graphically, lift charts consisting of a lift curve and a baseline are plotted by putting lift values on the y-axis and the percentage of samples drawn from the raw data set on the x-axis. When comparing different models, the greater the area between the lift curve and the baseline, the better the model is. Both cumulative and non-cumulative lift charts compare models at a decile level.

3.3.5.2 Classification Tables

The classification table is useful to summarize the predictive power of a binary regression model (Agresti 2007). This cross classifies the binary outcomes of y with the model's prediction $\hat{y} = 0$ or $\hat{y} = 1$ under a cutoff π_0 . The prediction $\hat{y} = 1$ when $\hat{\pi}_i > \pi_0$, and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$. The two useful summaries of predictive power are sensitivity $P(\hat{y} = 1 | y = 1)$ and specificity $P(\hat{y} = 0 | y = 0)$. Sensitivity measures the proportion of actual observations of $y = 1$ that is correctly identified as $\hat{y} = 1$. Specificity measures the proportion of actual observations of $y = 0$ that is correctly identified as $\hat{y} = 0$. Then, the overall proportion of correct classification is a weighted average of the sensitivity and specificity as the equation shown.

$$\begin{aligned}
P(\text{correct classification}) &= P(\hat{y}=1 \text{ and } y=1) + P(y=0 \text{ and } y=0) \\
&= P(\hat{y}=1 | y=1)P(y=1) + P(y=0 | y=0)P(y=0)
\end{aligned}
\tag{13}$$

3.3.6 Model Scoring and Interpretation

Given a specific response variable and a set of explanatory variables, the fitted Bayesian logistic regression models provide an estimated probability or odds that a 5-digit ZCTA with a particular set of explanatory variables will contain a biomass-using facility. However, when applying the models outside the learning data set, the interpretation does not hold due to the retrospective way of building the sample. Instead, we use the estimated probability simply as a score which will tell us how similar the 5-digit ZCTA is to a 5-digit ZCTA in the learning sample that had a high probability of having a facility. This score can then be used to “rank” orders from those that are most similar to the 5-digit ZCTAs with existing facilities to those that are least like 5-digit ZCTAs with existing facilities. The score should not be interpreted as a predictive estimate of the probability that a particular 5-digit ZCTA will contain a biomass-using facility or biorefinery in the future.

Chapter 4 Results and Discussion

4.1 Logistic Regression Model Selection

4.1.1 Group I Biomass-using Facilities

Using the learning sample of 4,853 observed 5-digit ZCTAs across the 13 Southeastern states, the four modeling approaches described in Subsection 3.3 are executed for the response defined using Group I biomass-using facilities. The AIC, BIC values and misclassification rates of these four models are summarized in Table 4-1.

As shown in Table 4-1, Model 3 has the lowest AIC and BIC score, and the lowest misclassification rate in the training test sample. The lift charts of all four models are also compared, but no significant difference is found (Figure 4-1).

Table 4-1 AIC, BIC values and misclassification rates of the four models for Group I

Model Name	AIC Value	BIC Value	Misclassification Rate
Model 1 Logistic regression without variable selection and variable transformation	2130.002210	2267.463904	0.142170
Model 2 Logistic regression with stepwise variable selection	2181.507218	2289.085935	0.144574
Model 3 Logistic regression with variable transformation	2052.163225	2171.695133	0.130495
Model 4 Logistic regression with both variable selection and variable transformation	2082.748632	2208.257136	0.130838

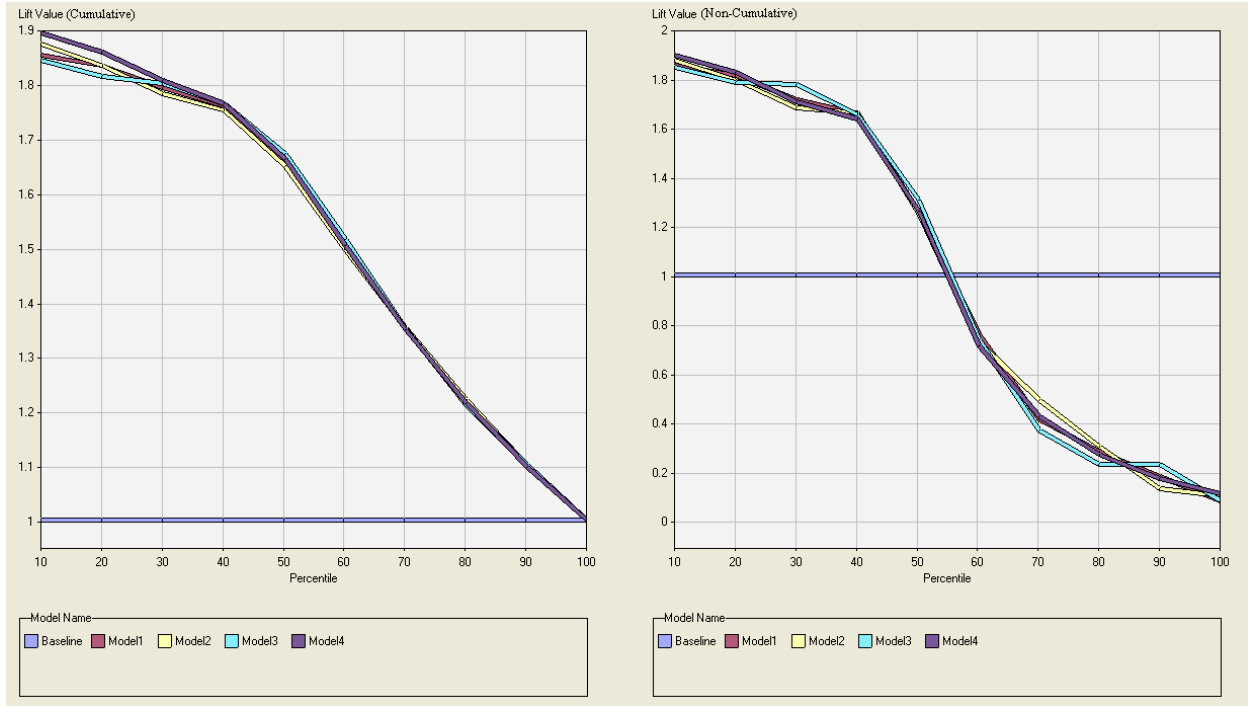


Figure 4-1 Cumulative and Non-Cumulative lift charts that assesses four models for Group I biomass-using facilities

The classification table from the validation dataset of Model 3 for Group I is shown in Table 4-2, which verifies Model 3 has a very good predictive power for the siting locations of the Group I biomass-using facilities. The sensitivity of Model 3 is $P(\hat{y} = 1 | y = 1) = 88.42\%$, and the specificity is $P(\hat{y} = 0 | y = 0) = 79.59\%$. The overall proportion of correct classification is $P(\hat{y} = 1 | y = 1)P(y = 1) + P(\hat{y} = 0 | y = 0)P(y = 0) = 84.19\%$.

Table 4-2 Classification Table from the validation dataset for Group I biomass-using facilities

Predicted Value Actual Value	0	1	Total
0	741 (79.59%)	190	931 (47.96%)
1	117	893 (88.42%)	1010 (52.04%)
Total	858	1083	1941

4.1.2 Group II Biomass-using Facilities

The same four modeling approaches using the learning sample of 2955 observed 5-digit ZCTAs are executed for the responses defined for Group II biomass-using facilities. The AIC, BIC values and misclassification rates are summarized in the Table 4-3. Among the models, Model 1 has the lowest AIC score, and nearly lowest BIC score. Model 1 also has the lowest misclassification rate in the training data set. The lift charts still cannot differentiate the four models (Figure 4-2). Therefore, we can conclude that Model 1, the pure stepwise logistic regression model, is best.

The classification table from the validation dataset of Model 1 for Group II is shown in Table 4-4 with the sensitivity $P(\hat{y} = 1 | y = 1) = 52.32\%$ and the specificity $P(\hat{y} = 0 | y = 0) = 98.54\%$. The overall proportion of correct classification is $P(\hat{y} = 1 | y = 1)P(y = 1) + P(\hat{y} = 0 | y = 0)P(y = 0) = 95.18\%$.

Table 4-3 AIC, BIC values and misclassification rates of the four models for Group II

Model Name	AIC Value	BIC Value	Misclassification Rate
Model 1 Logistic regression without variable selection and variable transformation	359.048718	441.255143	0.035533
Model 2 Logistic regression with stepwise variable selection	415.482278	464.806133	0.038353
Model 3 Logistic regression with variable transformation	381.734297	431.058151	0.036661
Model 4 Logistic regression with both variable selection and variable transformation	384.675263	444.959974	0.039481

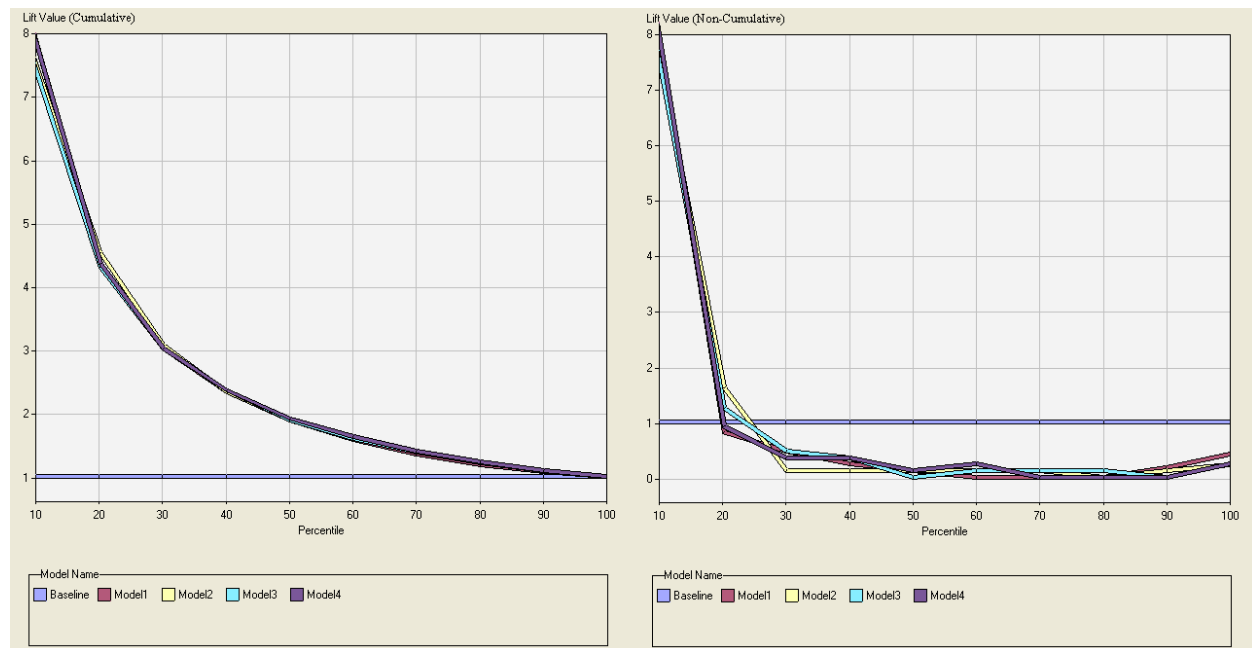


Figure 4-2 Cumulative and Non-Cumulative lift charts that assesses four models for Group II biomass-using facilities

Table 4-4 Classification Table from the validation dataset for Group II biomass-using facilities

Predicted Value Actual Value	0	1	Total
0	1080 (98.54%)	16	1096 (92.72%)
1	41	45 (52.32%)	86 (7.28%)
Total	1121	61	1182

4.1.3 Group III Biomass-using Facilities

Repeatedly executing the four modeling approaches for the responses defined for Group III biomass-using facilities, we can see that Model 3 obtains the lowest AIC and BIC scores, but has the highest misclassification rate for the training data set (Table 4-5). The lift charts shown in Figure 4-3 still fail to effectively differentiate the four models. Considering the limited number of responses, 155 in total, that has biomass-using facilities located in responding 5-digit ZCTAs in Group III, we conclude that Model 3, a stepwise logistic regression model with variable transformation, is best.

The classification table from the validation dataset of Model 3 for Group III is shown in Table 4-6 with the sensitivity $P(\hat{y} = 1 | y = 1) = 29.82\%$ and the specificity $P(\hat{y} = 0 | y = 0) = 98.38\%$. The overall proportion of correct classification is $P(\hat{y} = 1 | y = 1)P(y = 1) + P(\hat{y} = 0 | y = 0)P(y = 0) = 91.98\%$. Although the sensitivity is low, the high overall correct classification rate still makes Model 3 powerful for prediction.

Table 4-5 AIC, BIC values and misclassification rates of the four models for Group III

Model Name	AIC Value	BIC Value	Misclassification Rate
Model 1 Logistic regression without variable selection and variable transformation	405.968155	449.348303	0.073144
Model 2 Logistic regression with stepwise variable selection	427.187813	460.927927	0.076419
Model 3 Logistic regression with variable transformation	400.766161	439.326292	0.077511
Model 4 Logistic regression with both variable selection and variable transformation	391.193720	434.573867	0.0764192

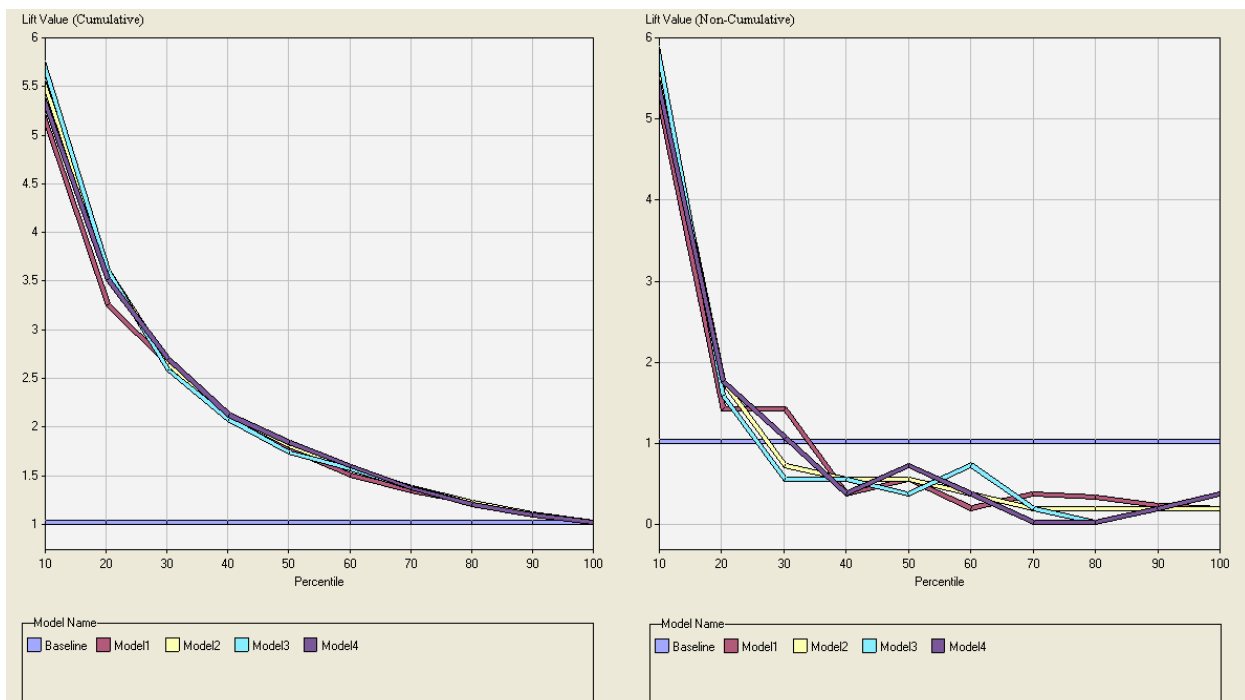


Figure 4-3 Cumulative and Non-Cumulative lift charts that assesses four models for Group III biomass-using facilities

Table 4-6 Classification Table from the validation dataset for Group III biomass-using facilities

Predicted Value Actual Value	0	1	Total
0	545 (98.38%)	9	554 (90.67%)
1	40	17 (29.82%)	57 (9.33%)
Total	585	26	611

4.2 Maximum Likelihood Estimates and Bayesian Estimation

Maximum likelihood estimates (MLE) of parameters show that population density (negative influence) is the common statistically significant factor (p-value < 0.0001) for all three groups. There are several additional significant factors for each group. Bayesian estimation, which predefines four different prior probability distributions for parameter β , effectively improves the predictive power of logistic regression models. Twenty-five optimal locations in the Southeastern states (5-digit ZCTAs) are then predicted respectively.

4.2.1 Group I Biomass-using Facilities

Eighteen out of 116 explanatory variables are statistically significant (p-values < 0.05) in the selected best Model 3 for the predefined Group I, which are Median_Family_Income, Sorghum_RES_Qty, WheatWinter_RES_HC, Population, Population_Density, Barley_120_TQty, Corn_80_TCost, Corn_120_TQty, Oat_160_MCost, Oat_160_ACost, Sorghum_160_TQty, WheatAll_40_TCost, WheatAll_120_TQty, WheatAll_160_TCost,

WheatAll_160_MCost, and WheatWinter_160_TQty. RailroadAvailability is significant based on the Type 3 test⁴ (p-value =0.0007).

Using the maximum likelihood estimates as our prior beliefs to build up the four predefined prior probability distributions (Uniform prior, Gaussian prior, Laplace prior and Cauchy prior) of parameter β , the posterior probability distribution under each prior is obtained after running the Markov chain Monte Carlo simulation 10,000 times. We then use the expected value of the posterior probability distribution of parameter β as regression coefficients. In order to assess the predictive power of the model, the misclassification table for maximum likelihood estimation and Bayesian methods is summarized in Table 4-7. Gaussian prior for β has the highest correct classification rate followed by uniform prior for β , implying these two Bayesian methods have a higher predictive power than maximum likelihood estimation does.

Table 4-7 Classification rates under maximum likelihood estimation and Bayesian methods for Group I biomass-using facilities

Estimation Method		Specificity $P(\hat{y} = 0 y = 0)$	Sensitivity $P(\hat{y} = 1 y = 1)$	Correct Classification Rate
Maximum Likelihood Estimation		$\frac{1943}{2328} = 83.46\%$	$\frac{2223}{2525} = 88.04\%$	85.84%
Bayesian Inference	Uniform Prior	$\frac{1980}{2328} = 85.05\%$	$\frac{2201}{2525} = 87.17\%$	86.15%
	Gaussian Prior	$\frac{1990}{2328} = 85.48\%$	$\frac{2203}{2525} = 87.25\%$	86.40%
	Laplace Prior	$\frac{1960}{2328} = 84.19\%$	$\frac{2191}{2525} = 86.77\%$	85.53%
	Cauchy Prior	$\frac{1968}{2328} = 84.54\%$	$\frac{2191}{2525} = 86.77\%$	85.70%

⁴ 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).

By setting Gaussian prior distribution $p(\beta | u, \sigma^2)$, where u is the maximum likelihood estimate and σ^2 equals to 1 for each parameter, the posterior parameter estimates are summarized in Table 4-8. The estimates show that population, cumulative total quantity of corn stover residues within a 160-mile haul distance (Corn_160_TQty), average harvesting costs of oat straw residues within a 160-mile haul distance (Oat_160_ACost), total harvesting costs of wheat all straw residues within a 40-mile and 160-mile haul distance (WheatAll_40_TCost and WheatAll_160_TCost), and cumulative total quantity of wheat all straw residues within a 160-mile haul distance have positive coefficients, i.e., the relative score of this location increases as these variables increase. The rest 12 significant variables have negative coefficients implying that the appeal of a 5-digit ZCTA as a potential location rises as the decrease of these variables. The resulting posterior distribution for each parameter, which follows an approximate normal distribution, is shown in Figure 4-4. Interestingly, the posterior standard deviation and 95% credibility interval show considerably lower values for Median_Family_Income, WheatWinter_RES_HC, Corn_80_TCost, and WheatAll_40_TCost compared to others.

Table 4-8 Parameter estimates under Gaussian prior for Group I biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	1.182	0.589	0.547	1.157	1.912
Median_Family_Income	-0.443	0.058	-0.536	-0.442	-0.350
RailRoadAvailability 1	0.562	0.193	0.253	0.555	0.873
RailRoadAvailability 2	0.302	0.129	0.099	0.304	0.514
Sorghum_RES_Qty	-0.883	0.214	-1.294	-0.879	-0.497
WheatWinter_RES_HC	-0.056	0.018	-0.087	-0.056	-0.027
Population	1.062	0.130	0.817	1.060	1.306
Population_Density	-2.394	0.143	-2.665	-2.390	-2.133
Barley_120_TQty	-0.853	0.179	-1.127	-0.847	-0.583
Corn_80_TCost	-0.084	0.020	-0.119	-0.084	-0.049
Corn_120_TQty	-3.473	0.400	-4.259	-3.469	-2.703
Corn_160_TQty	5.864	0.398	5.101	5.857	6.626
Oat_160_MCost	-0.555	0.190	-0.868	-0.570	-0.221
Oat_160_ACost	0.715	0.202	0.360	0.735	1.052
Sorghum_160_TQty	-1.997	0.336	-2.282	-1.989	-1.691
WheatAll_40_TCost	0.125	0.022	0.093	0.125	0.159
WheatAll_120_TQty	2.121	0.524	1.293	2.134	2.967
WheatAll_160_TCost	0.649	0.254	0.186	0.643	1.167
WheatAll_160_MCost	-1.112	0.295	-1.711	-1.102	-0.579
WheatWinter_160_TQty	-5.311	0.498	-6.306	-5.296	-4.429

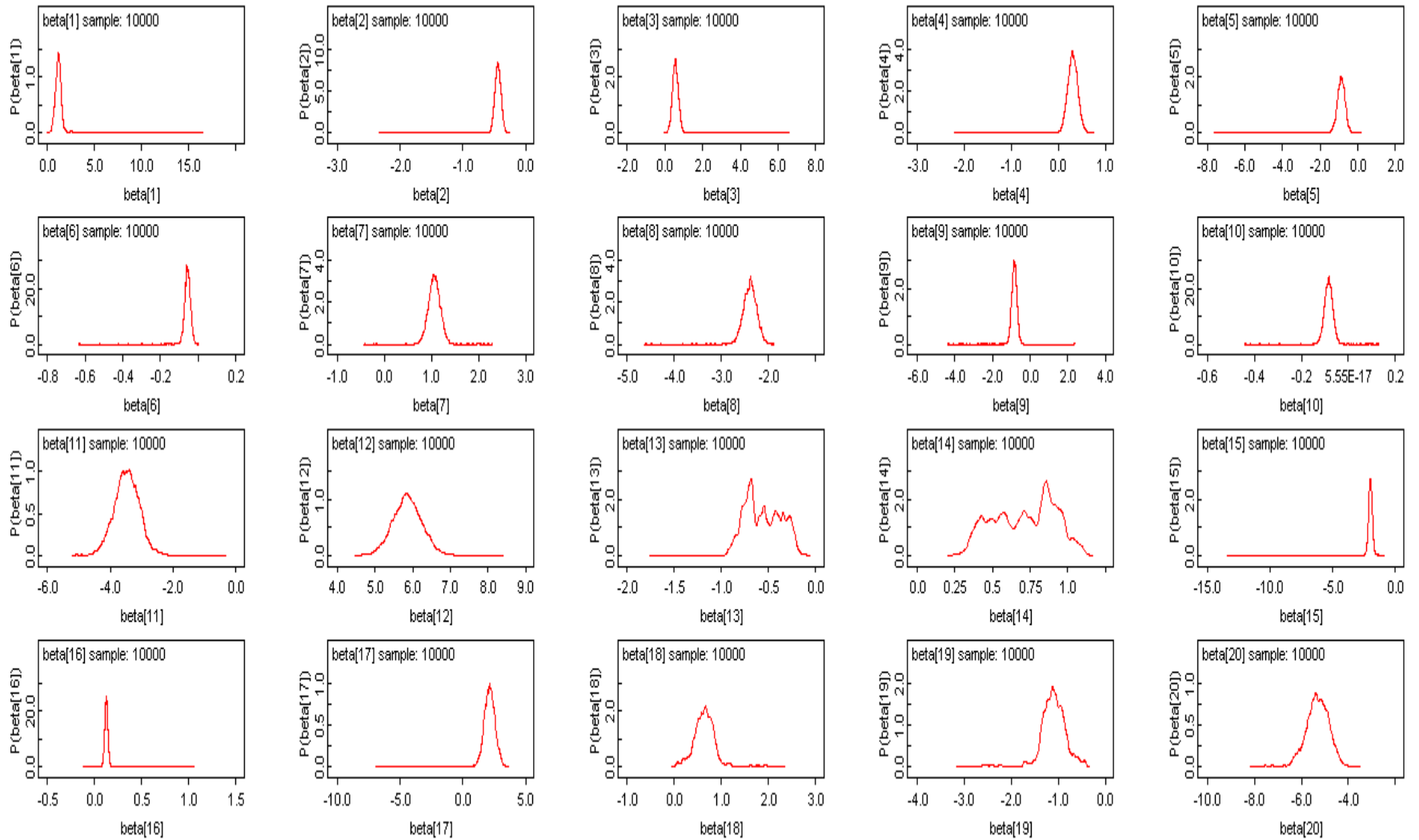


Figure 4-4 Posterior distributions of parameters for Group I biomass-using facilities

Based on the selected best Model 3 and estimates under Gaussian prior, the 25 potential locations (5-digit ZCTAs) with the highest scores as Group I facilities in the 13 states are predicted and plotted in Figure 4-5. There are ten possible locations in Mississippi, eight in Tennessee, six in Kentucky and one in Arkansas. Note in Figure 1 that there are many existing Group I biomass-using facilities clustered in Mississippi, Tennessee, and Kentucky, and thus a higher possibility for future sites located in these three states based upon our model. Therefore, the number of optimal locations in these three states is more than other states.

4.2.2 Group II Biomass-using Facilities

Fourteen out of 119 explanatory variables are statistically significant (p -values < 0.05) in the selected Model 1 for the predefined Group II, including Population, Population_Density, Primary_Mill_No, Secondary_Mill_No, Corn_RES_HC, Barley_120_ACost, Barley_160_TCost, Corn_160_TCost, Oat_160_TCost, Sorghum_80_TQty, Sorghum_160_TQty, WheatAll_120_MCost, WheatAll_160_TQty, and WheatWinter_160_TQty. Notice that the number of primary and secondary mills within 5-digit ZCTAs have statistically significant effects on the score of this 5-digit ZCTA as a potential location verifying the assumption that pulp and paper mills prefer a location with some primary and secondary mills since they can use the residues from the primary and secondary mills.

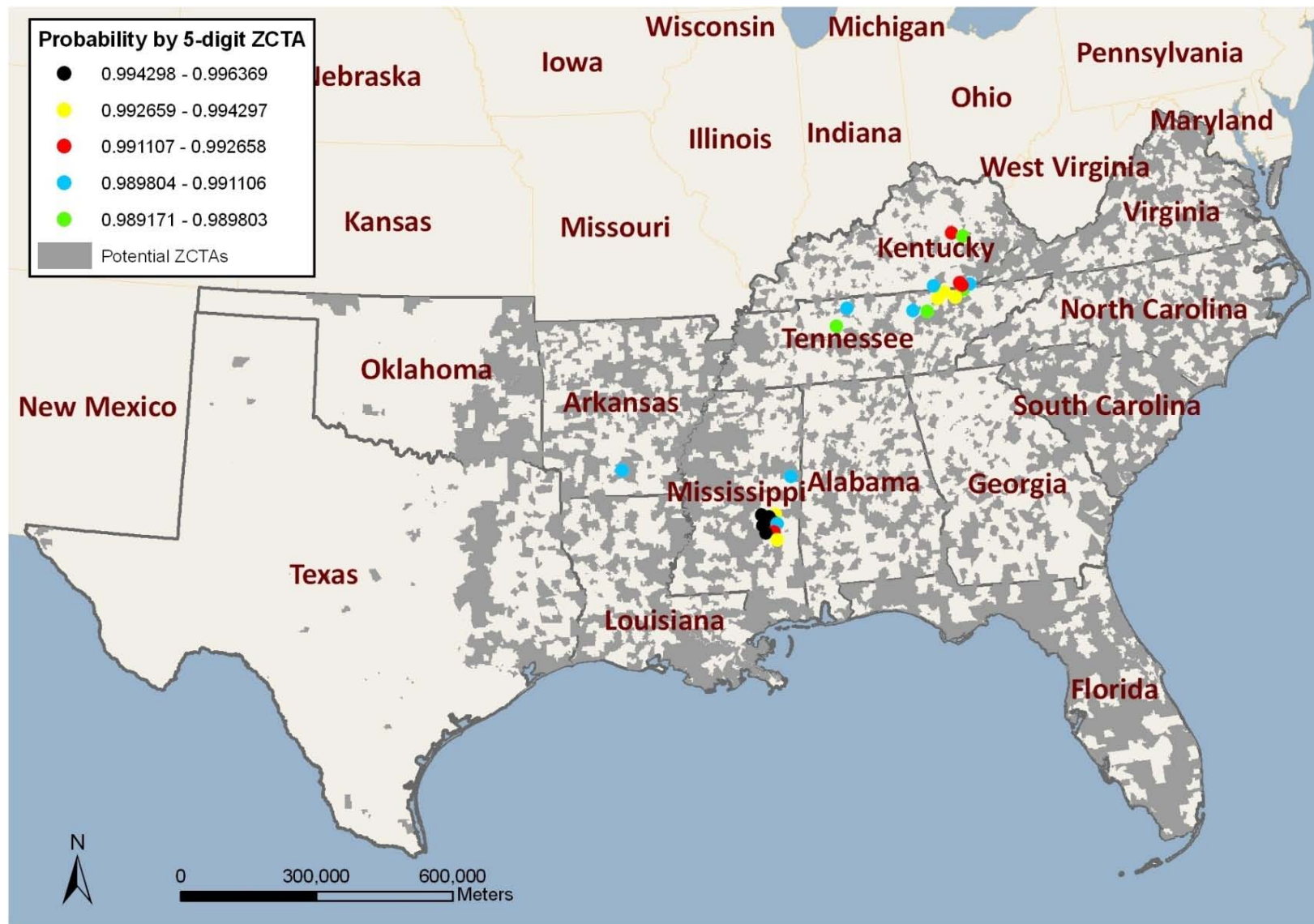


Figure 4-5 Top 25 optimal locations for Group I biomass-using facilities in Southeastern 13 states

Maximum likelihood estimates provide useful prior knowledge to construct our four different prior probability distributions of parameter β in the following step. After performing the Markov chain Monte Carlo simulation 10,000 times, we obtain the posterior probability distributions of parameter β under each prior and use the expected values of the posterior distributions as parameter coefficients. The misclassification table for maximum likelihood estimation and Bayesian methods is summarized in Table 4-9 to assess the predictive power of the model under different estimation methods. Non-informative uniform prior for β , which holds almost the same specificity but increases the sensitivity lot, has the highest correct classification rate followed by maximum likelihood estimation. It indicates that assuming a uniform prior for the parameter β fits the data better than maximum likelihood estimation, implying an improvement of the predictive power by applying this Bayesian method.

Table 4-9 Classification rates under maximum likelihood estimation and Bayesian methods for Group II biomass-using facilities

Estimation Method		Specificity $P(\hat{y} = 0 y = 0)$	Sensitivity $P(\hat{y} = 1 y = 1)$	Correct Classification Rate
Maximum Likelihood Estimation		$\frac{2732}{2764} = 98.84\%$	$\frac{103}{191} = 53.93\%$	95.94%
Bayesian Inference	Uniform Prior	$\frac{2724}{2764} = 98.55\%$	$\frac{112}{191} = 58.64\%$	95.97%
	Gaussian Prior	$\frac{2720}{2764} = 98.40\%$	$\frac{105}{191} = 54.97\%$	95.59%
	Laplace Prior	$\frac{2761}{2764} = 99.89\%$	$\frac{19}{191} = 9.95\%$	94.08%
	Cauchy Prior	$\frac{2757}{2764} = 99.75\%$	$\frac{39}{191} = 20.42\%$	94.63%

By setting uniform prior distribution $p(\beta)$ within a range, which contains the maximum likelihood estimate values, the posterior parameter estimates are summarized in Table 4-10. The estimates show that the number of primary and secondary mills in 5-digit ZCTAs (Primary_Mill_No and Secondary_Mill_No), the total cost of barley straw, corn stover, and oat straw residues within a 160-mile haul distance (Baryle_160_TCost, Corn_160_TCost, and Oat_160_TCost), and cumulative total quantities of wheat all straw residues within a 160-mile haul distance (WheatAll_160_TQty) have positive coefficients, implying the relative attractiveness of a 5-digit ZCTA as a potential location increases as these variables increase. The rest 8 significant variables have negative coefficients, i.e., the score of this location increases as these variables decrease. The resulting posterior distribution for each parameter is shown in Figure 4-6. Noticeably, the posterior probability distributions of parameters Oat_160_TCost, WheatAll_120_MCost, and WheatWinter_160_TQty are approximately normal, and the posterior distributions of Population_Density and WheatAll_160_TQty still keep uniform distributions.

Based on the selected best Model 1 and estimates under uniform prior, the 25 potential locations (5-digit ZCTAs) with the highest scores as Group II facilities in the 13 states are predicted and plotted in Figure 4-7. There are fifteen possible locations in Kentucky, four in Texas, two in Tennessee and Arkansas, and one in North Carolina and Georgia, respectively. Kentucky, Texas, Tennessee, Arkansas, and North Carolina have comparatively lower competition from existing Group II facilities than Georgia, and thus gains more potential locations.

Table 4-10 Parameter estimates under Uniform prior for Group II biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-5.498	0.1528	-5.849	-5.461	-5.307
Corn_RES_HC	-0.00014	0.00019	-0.00019	-0.00015	-0.0001
Population	-0.5202	0.05062	-0.6319	-0.5107	-0.4527
Population_Density	-1.605	0.05681	-1.695	-1.608	-1.506
Primary_Mill_No	1.376	0.1013	1.212	1.37	1.576
Barley_120_ACost	-0.00106	0.0001	-0.00128	-0.00104	-0.001
Barley_160_TCost	0.0013	0.00011	0.0012	0.00128	0.00152
Corn_160_TCost	1.187	0.0756	1.028	1.1990	1.295
Oat_160_TCost	0.00021	0.00003	0.00016	0.00021	0.00027
Sorghum_80_TQty	-3.347	0.04632	-3.467	-3.333	-3.301
Sorghum_160_TQty	-1.557	0.05723	-1.708	-1.54	-1.501
WheatAll_120_MCost	-0.00017	0.00006	-0.00028	-0.00017	-0.00006
WheatAll_160_TQty	7.664	0.1961	7.324	7.67	7.983
WheatWinter_160_TQty	-8.955	0.245	-9.456	-8.941	-8.546
Secondary_Mill_No	0.3058	0.02869	0.2323	0.3132	0.339

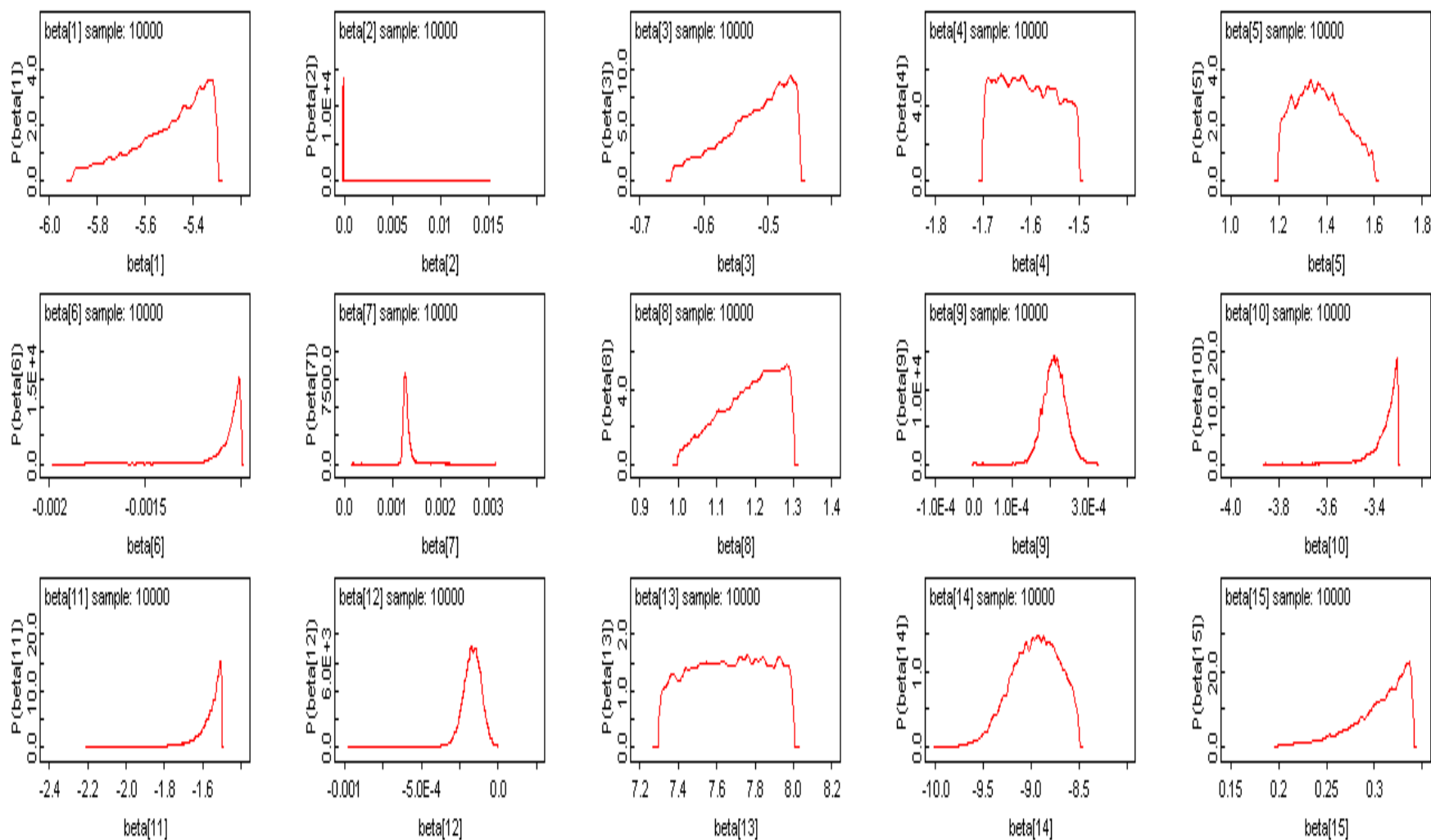


Figure 4-6 Posterior distributions of parameters for Group II biomass-using facilities

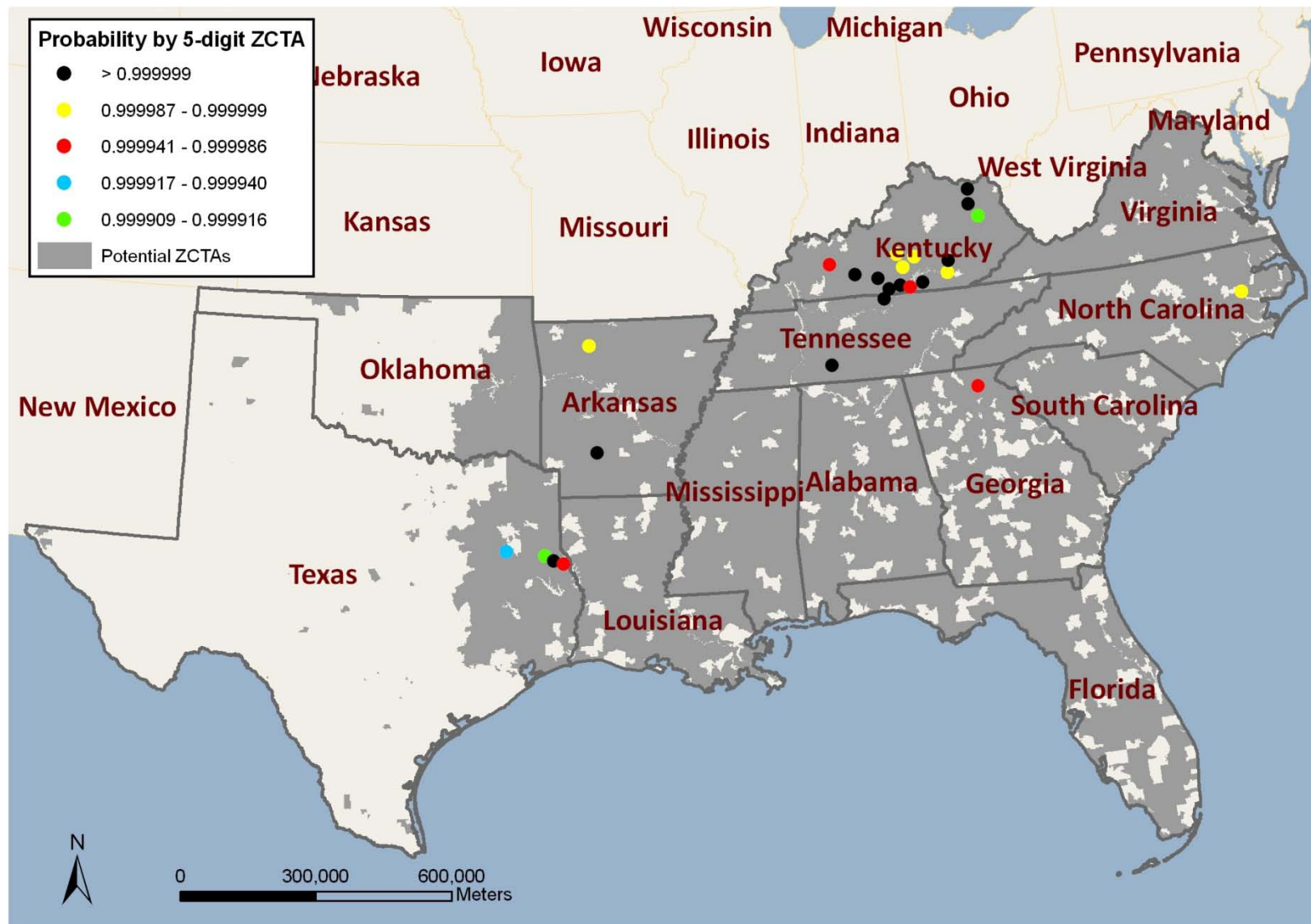


Figure 4-7 Top 25 optimal locations for Group II biomass-using facilities in Southeastern 13 states

4.2.3 Group III Biomass-using Facilities

Seven out of 116 explanatory variables are statistically significant (p-values <0.05) in the selected Model 3 for the predefined Group III, including Population_Density, Sqmiwater, Sorghum_RES_Qty, WheatWinter_RES_HC, Oat_40_ACost, Oat_120_TCost, and WheatWinter_80_TCost. The variable, water area within 5-digit ZCTAs (Sqmiwater), has a statistically significant effect on the score of this 5-digit ZCTA as a potential location, which verifies our observation from Figure 3 that most Group III facilities locate near water.

To build up the four predefined prior probability distributions in the use of maximum likelihood estimates as prior beliefs, the posterior probability distributions of parameter β under each prior is derived after performing the Markov chain Monte Carlo simulation 10,000 times. The expected values of posterior probability distributions of parameter β are treated as regression coefficients. The misclassification table for maximum likelihood estimates and Bayesian methods is summarized in Table 4-11 assessing the predictive power of the model under different estimation methods. Bayesian inference assuming four different priors has a higher correct classification rate than maximum likelihood estimation does. Notice that Bayesian methods assuming Gaussian prior, Laplace prior, and Cauchy prior improve both the sensitivity and specificity rates in the misclassification table, indicating that they have a better predictive power. Particularly, Gaussian prior, which has the highest sensitivity and specificity rates, results a highest correction rate of 92.67%, and thus fits the data best compared with other methods.

Table 4-11 Classification rates under maximum likelihood estimation and Bayesian methods for Group III biomass-using facilities

Estimation Method		Specificity $P(\hat{y} = 0 y = 0)$	Sensitivity $P(\hat{y} = 1 y = 1)$	Correct Classification Rate
Maximum Likelihood Estimation		$\frac{1351}{1372} = 98.47\%$	$\frac{56}{155} = 36.13\%$	92.14%
Bayesian Inference	Uniform Prior	$\frac{1354}{1372} = 98.69\%$	$\frac{56}{155} = 36.13\%$	92.34%
	Gaussian Prior	$\frac{1354}{1372} = 98.69\%$	$\frac{61}{155} = 39.35\%$	92.67%
	Laplace Prior	$\frac{1354}{1372} = 98.69\%$	$\frac{60}{155} = 38.71\%$	92.60%
	Cauchy Prior	$\frac{1354}{2764} = 98.69\%$	$\frac{59}{191} = 38.06\%$	92.54%

By setting Gaussian prior distribution $p(\beta | u, \sigma^2)$, where u is the maximum likelihood estimate and σ^2 equals to 1 for each parameter, the posterior parameter estimates are summarized in Table 4-12. The estimates show that water area within 5-digit ZCTAs, Sorghum residues quantities in each 5-digit ZCTA and average trucking cost of oat straw residues within a 40-mile haul driving distance have positive impacts, implying the relative suitability of a 5-digit ZCTA as a potential location increases as these variables increase. The rest 4 significant variables have negative impacts, i.e., the score of this location increases as these variables decrease. The derived posterior distribution for each parameter is shown in Figure 4-8. Note that Oat_120_TCost, WheatWinter_80_TCost, Oat_40_ACost, and WheatWinter_RES_HC have a longer tail, but lower standard deviation and narrower 95% creditability intervals than other significant explanatory variables.

Table 4-12 Parameter estimates under Gaussian prior for Group III biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-1.329	0.3541	-1.985	-1.315	-0.7098
Population_Density	-1.556	0.2873	-2.134	-1.553	-1.001
WheatWinter_80_TCost	-0.1359	0.041	-0.2175	-0.1351	-0.0587
Oat_120_TCost	-0.09914	0.03923	-0.1683	-0.09802	-0.03071
Oat_40_ACost	0.2406	0.0613	0.1447	0.2393	0.3393
Sorghum_RES_Qty	1.436	0.3407	0.8057	1.422	2.141
Sqmiwater	2.416	0.4015	1.634	2.415	3.203
WheatWinter_RES_HC	-0.1967	0.0335	-0.2555	-0.1962	-0.1394

Based on the selected best Model 3 and estimates under Gaussian prior, the 25 potential locations (5-digit ZCTAs) with the highest scores as Group III facilities in the 13 states are predicted and plotted in Figure 14. There are eleven possible locations in Texas, four in Florida and Louisiana, three in North Carolina, two in Arkansas, and one in South Carolina. As shown in Figure 4-9, the possible locations with higher probability are clustered along Mississippi river and coastal ports, which is consistent with existing Group III facilities in Figure 3-3. In addition, Texas has a high concentration of agricultural residues (shown in Figure 3-5(a)), thus attracting more potential locations.

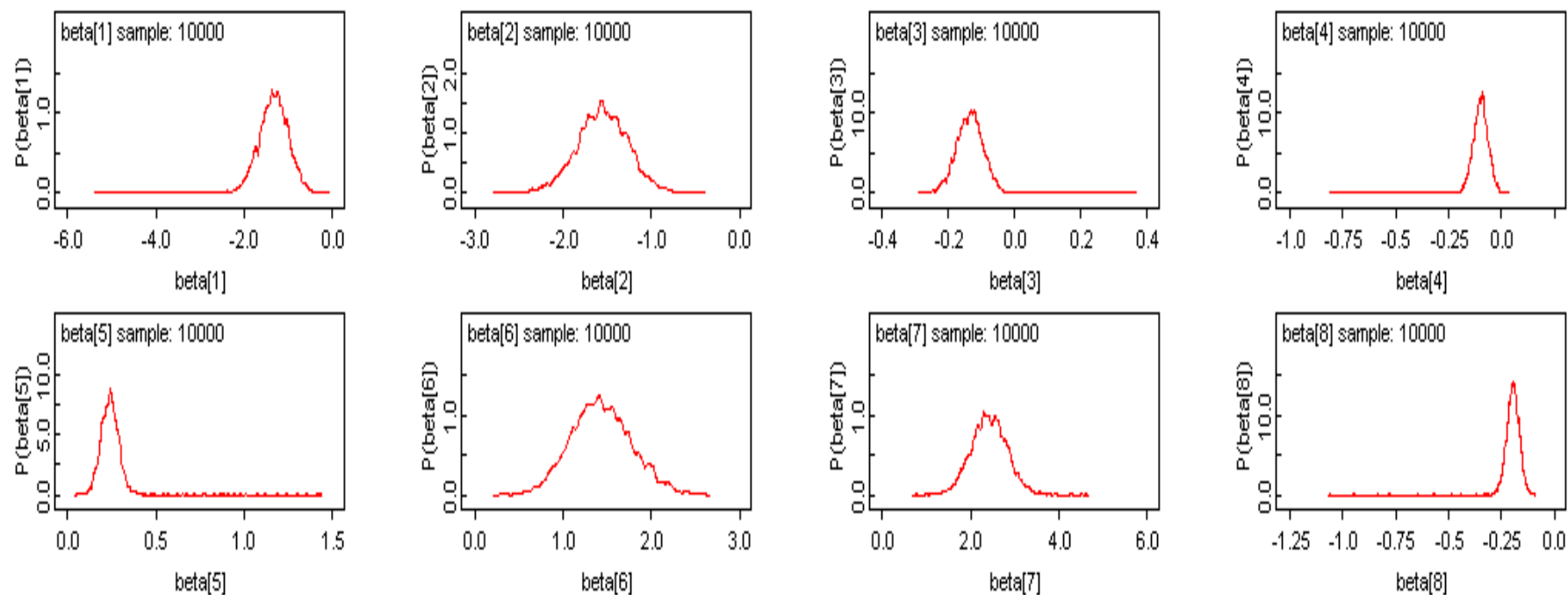


Figure 4-8 Posterior distributions of parameters for Group III biomass-using facilities

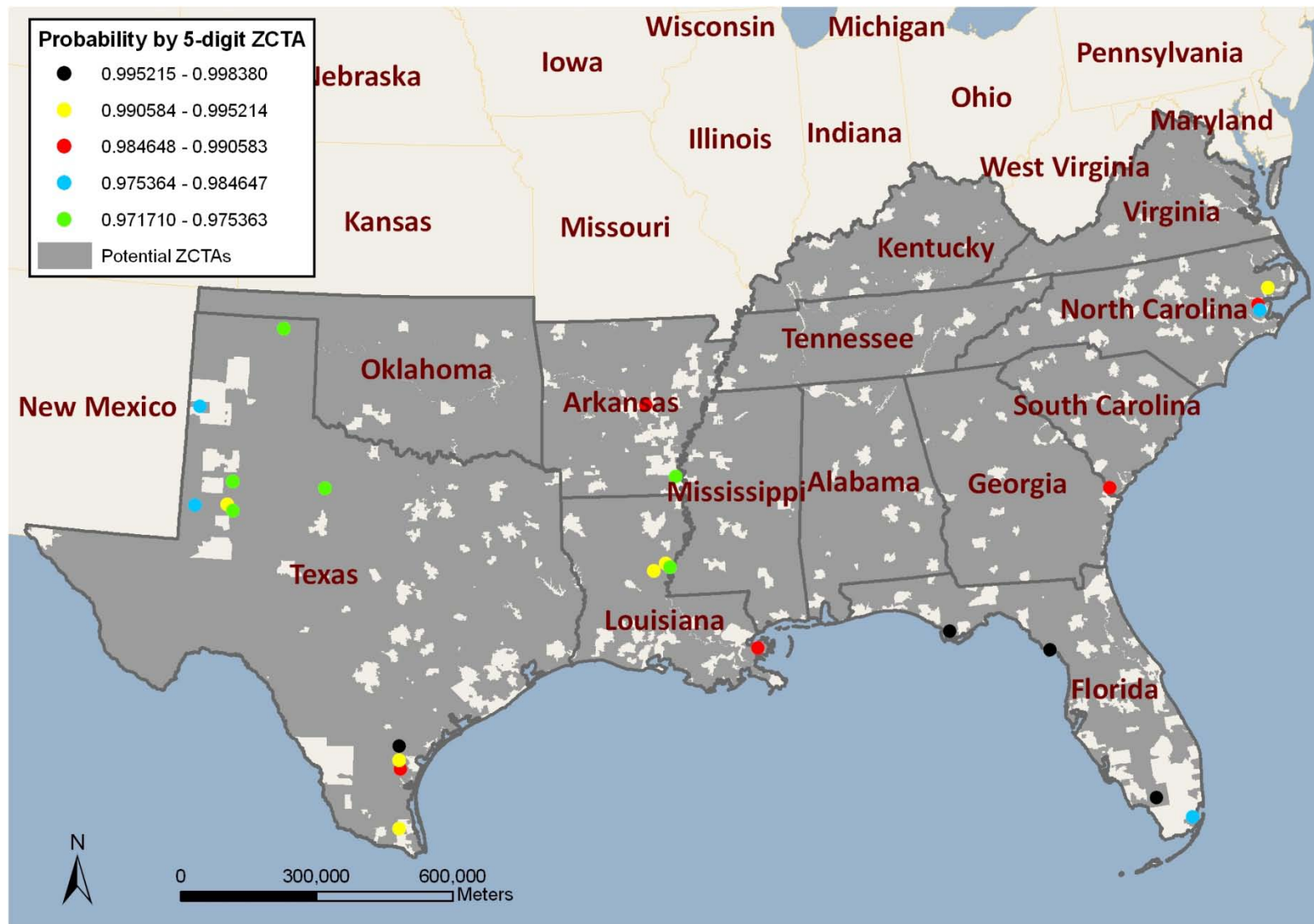


Figure 4-9 Top 25 optimal locations for Group III biomass-using facilities in Southeastern 13 states

4.3 Classification Trees with Local Bayesian Logistic Regression

Since the sensitivity for Group II and Group III biomass-using facilities reported in the section 4.1 is comparatively low, we then performed the hybrid model that integrates classification trees and local Bayesian logistic regression to further improve the predictive power for these two groups in this section. The results show that the hybrid model effectively increases the sensitivity of Group II biomass-using facilities from 58.54% to 64.40%, and has a higher correct classification rate than the developed best fitted Bayesian logistic regression model in section 4.2. For Group III biomass-using facilities, both of the specificity and sensitivity improves significantly, which are 99.42% and 46.45% respectively. Twenty-five locations with the highest scores in the Southeastern states (5-digit ZCTAs), based upon the hybrid model, are then predicted and mapped.

4.3.1 Group II Biomass-using Facilities

Five subsets of Group II biomass-using facilities are identified by the classification tree in Figure 4-10. Of all 2955 5-digit ZCTAs, 191 ZCTAs have existing mills and 2764 ZCTAs are not suitable for building any biomass-using facilities. Three significant variables, Primary_Mill_No, Population_Density, and Sorghum_160_TQty, are used to partition the whole data set. Under close scrutiny over the CT, Subset 1 has a comparatively larger number of 5-digit ZCTAs with existing mills than “non-probable” ZCTAs, while the other four Subsets mainly contain “non-probable” 5-digit ZCTAs.

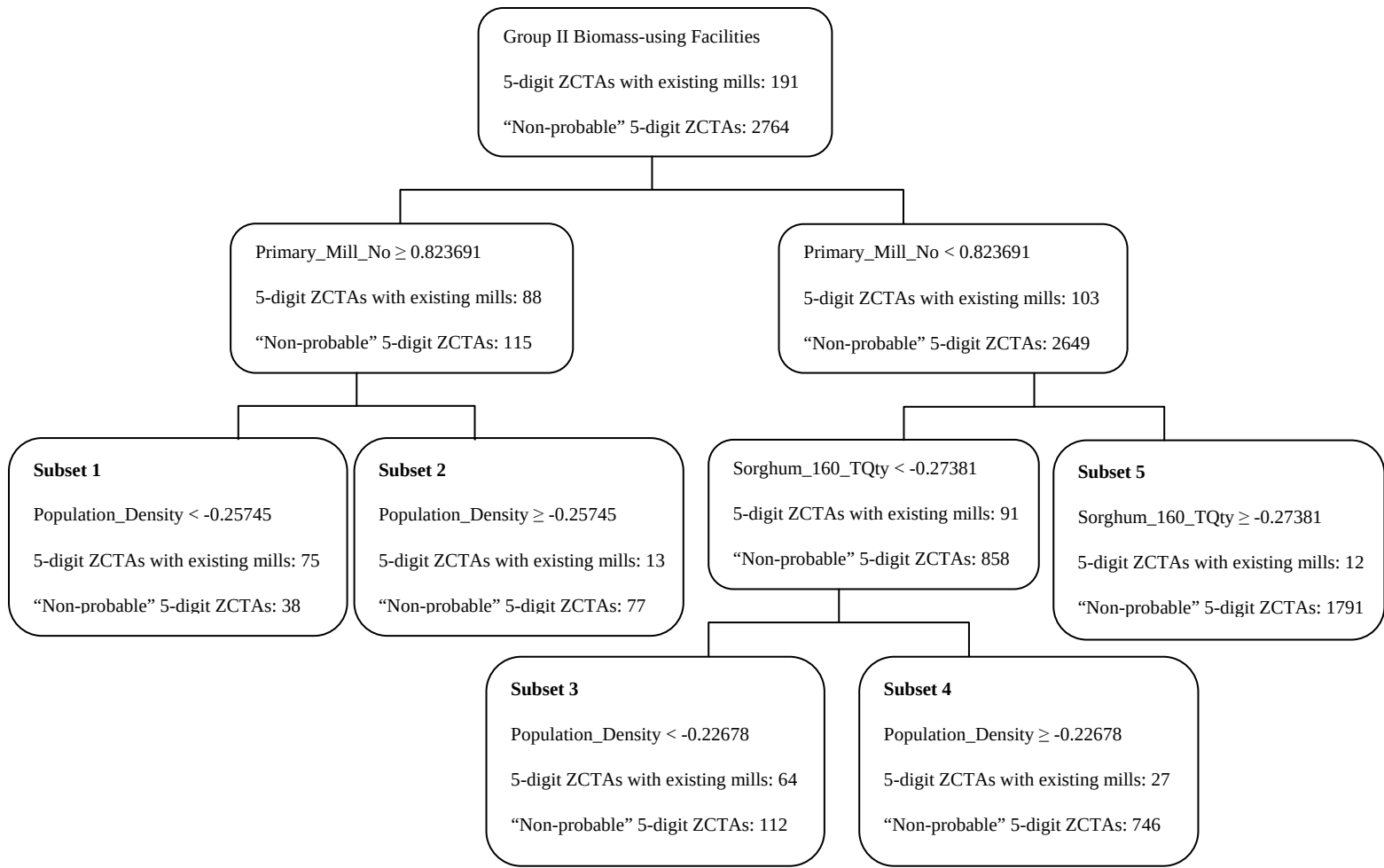


Figure 4-10 The classification tree build by JMP® for Group II biomass-using facilities

Table 4-13 Significant variables for five subsets of Group II biomass-using facilities

Significant Estimates	Whole Set				
	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
Population			0.0008		
Sorghum_80_TQty	0.0093				
Secondary_Mill_No		0.0106	0.0417	< .0001	
Railroad_Availability_1	0.2607*				
Railroad_Availability_2	0.0500				
Corn_160_TQty	0.0143				
Sorghum_120_ACost		0.0226			
Sqmiwater		0.0198			
WheatAll_RES_HC		< .0001			
Population_Density			0.0001		
WheatAll_160_TCost			0.0001		
No_ports					< .0001

* RailroadAvailability is significant based on the Type 3 test (p-value =0.0417)

By executing the four predefined modeling approaches on each subset, the model with the lowest AIC or BIC score is selected for determining statistically significant variables. Table 4-13 summarizes all statistically significant variables (p value is less than 0.05) for the five subsets. Subset 1 has four statistically significant variables, which are completely different from other four subsets. Secondary_Mill_No is a common significant variable for Subset 2, 3, and 4. Along with the significant splitting variable, Primary_Mill_No shown in Figure 4-10, it can lead to the same finding as discussed in the section 4.2.2 that pulp and paper mills prefer a location with some primary and secondary mills nearby since they could use the residues from them.

To build up the four predefined prior probability distributions (recall Uniform prior, Gaussian prior, Laplace prior, and Cauchy prior) for each subset, the maximum likelihood estimates achieved from the classical logistic regression are used for prior knowledge. The posterior probabilities distributions under each prior for the above significant variables are then derived and the expected values of posterior probability distributions are treated as regression coefficients. In order to assess the predictive performance of each subset in comparison with the

whole set, the misclassification table for maximum likelihood estimation and Bayesian methods is given in Table 4-14. Bayesian method assuming a Gaussian prior has a higher correct classification rate of 88.50% for the Subset 1 of Group II, and Bayesian method assuming a Uniform prior has exactly same correct classification rates of 85.56% and 78.98 as maximum likelihood estimation does for the Subset 2 and 3 for Group II, respectively. For the Subset 4 and 5, both maximum likelihood estimation and Bayesian methods, which assume the four predefined priors, give the same correct classification rates, since only “non-probable” ZCTAs are detected from their subsets. It is also noted that the correct classification rates of 96.51% and 99.33% respectively for Subset 4 and 5 are higher than that of the whole set, indicating the local models fit the data better for these two subsets than only one “best” model for the whole set does.

Table 4-14 Correct classification rates under maximum likelihood estimation and Bayesian methods for 5 subsets and the whole set of Group II biomass-using facilities

Estimation Method		Subset 1 (113)	Subset 2 (90)	Subset 3 (176)	Subset 4 (773)	Subset 5 (1803)	Whole Set (2955)
Maximum Likelihood Estimation		87.61%	85.56%	78.98%	96.51%	99.33%	95.94%
Bayesian Inference	Uniform Prior	87.61%	85.56%	78.98%	96.51%	99.33%	95.97%
	Gaussian Prior	88.50%	77.78%	63.63%	96.51%	99.33%	95.59%
	Laplace Prior	84.07%	23.33%	63.63%	96.51%	99.33%	94.08%
	Cauchy Prior	86.73%	58.89%	77.84%	96.51%	99.33%	94.63%

Table 4-15 Classification table by summing up classified values achieved from the “best” local model in each of five subsets for Group II biomass-using facilities

Predicted Value Actual Value	0	1	Total
0	2730 (98.77%)	34	2764
1	68	123 (64.40%)	191
Total	2798	157	2955

To assess the performance behavior of the hybrid model compared with the one best fitting model in Section 4.2, a classification table, which sums up classified values achieved from the “best” local model in each of five subsets for Group II biomass-using facilities, is constructed in Table 4-15. Bayesian method assuming a Gaussian prior gives the highest correct classification rate of 88.50% for the Subset 1, and it correctly classifies 31 out of 38 “non-probable” 5-digit ZCTAs and 69 out of 75 ZCTAs with existing biorefinery mills. For the Subset 2, either Bayesian method assuming a Uniform prior or maximum likelihood estimation has a highest correct classification rate of 85.56%, by which 70 out of 77 “non-probable” ZCTAs and 7 out of 13 ZCTAs with existing biorefinery mills are correctly classified. Through constructing the classification table for the other 3 subsets, a summarized correctly classified value for the “non-probable” 5-digit ZCTAs can be achieved by adding up the corresponding value from each of the five subsets, which is $31 + 70 + 92 + 746 + 1791 = 2730$. Following the same steps, another summarized correctly classified value for the ZCTAs with existing biorefinery mills is gained, which is $69 + 7 + 47 + 0 + 0 = 123$. Then, an overall proportion of correct classification is $P(\hat{y} = 1 | y = 1)P(y = 1) + P(\hat{y} = 0 | y = 0)P(y = 0) = 96.55\%$, where the sensitivity is $P(\hat{y} = 1 | y = 1) = 66.40\%$, and the specificity is $P(\hat{y} = 0 | y = 0) =$

98.77%. In comparison with the overall proportion of correct classification derived from only one “best” model in Section 4.2.2 (recall Table 4-9), the hybrid model increases the correct classification rate from 95.97% to 96.55%, mostly attributing to the improved sensitivity from 58.54% to 64.40%.

Based on the hybrid model, which comprises 5 different local models with the highest correct classification rates for the five subsets, a total of 25 potential locations (5-digit ZCTAs) with the highest ranking scores as Group II facilities in the 13 states are predicted and plotted in Figure 4-11. There are twelve possible locations in North Carolina, five in Tennessee and Virginia, two in South Carolina, and one in Kentucky.

4.3.2 Group III Biomass-using Facilities

Three subsets of Group III biomass-using facilities are partitioned from the whole set by the classification tree approach in Figure 4-12. Of all 1527 5-digit ZCTAs in the whole set, 155 ZCTAs locate existing mills and 1372 ZCTAs are not suitable for siting any biomass-using facilities. Two significant variables, WheatAll_RES_Qty and Population_Density, are used for identifying the subsets. Subset 1 has more 5-digit ZCTAs with existing mills than “non-probable” ZCTAs, while Subset 2 contains 10 ZCTAs with existing mills and 46 “non-probable” ZCTAs. Subset 3 is mainly consist of “non-probable” 5-digit ZCTAs.

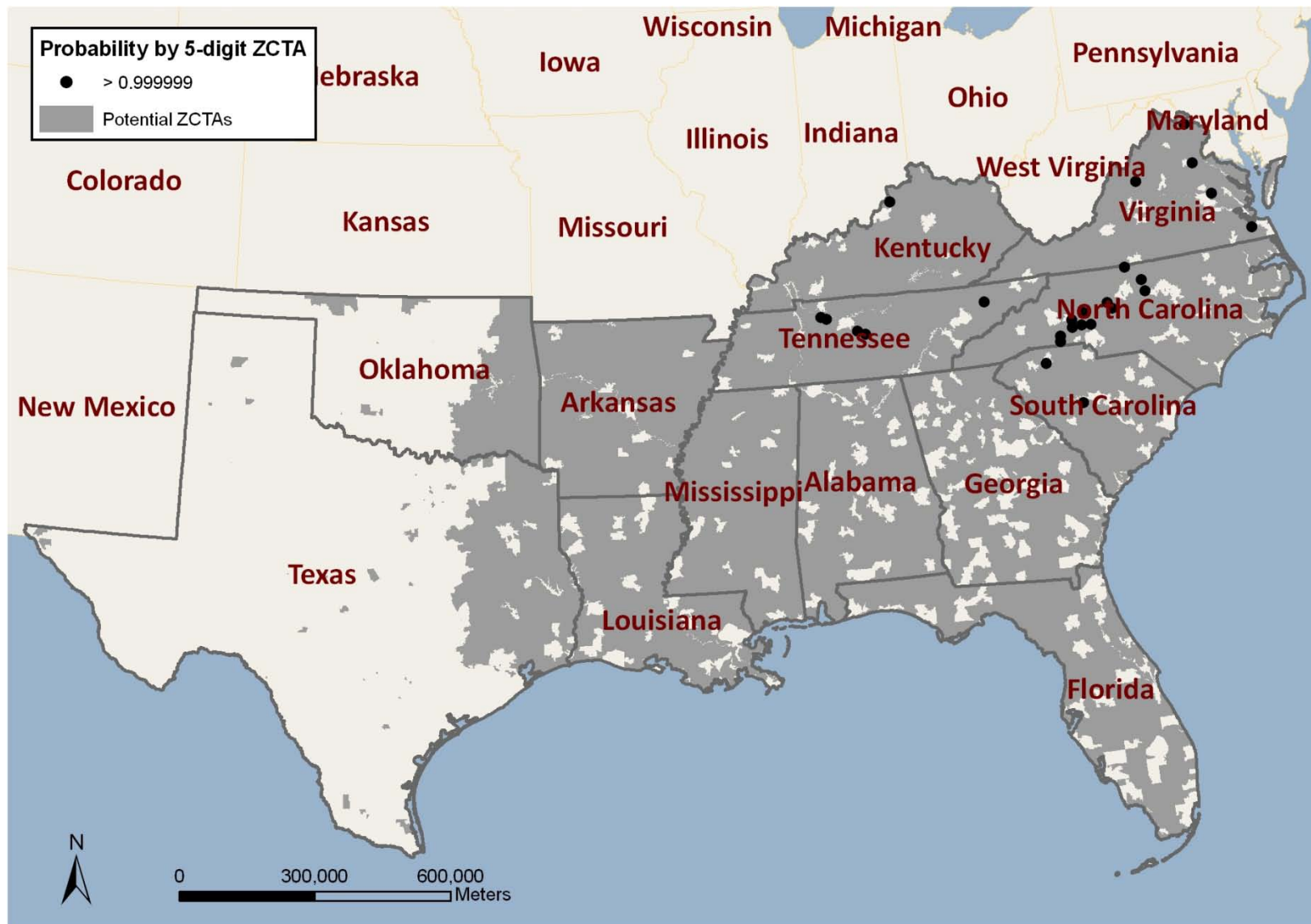


Figure 4-11 Top 25 optimal locations based upon the hybrid model for Group II biomass-using facilities in Southeastern 13 states

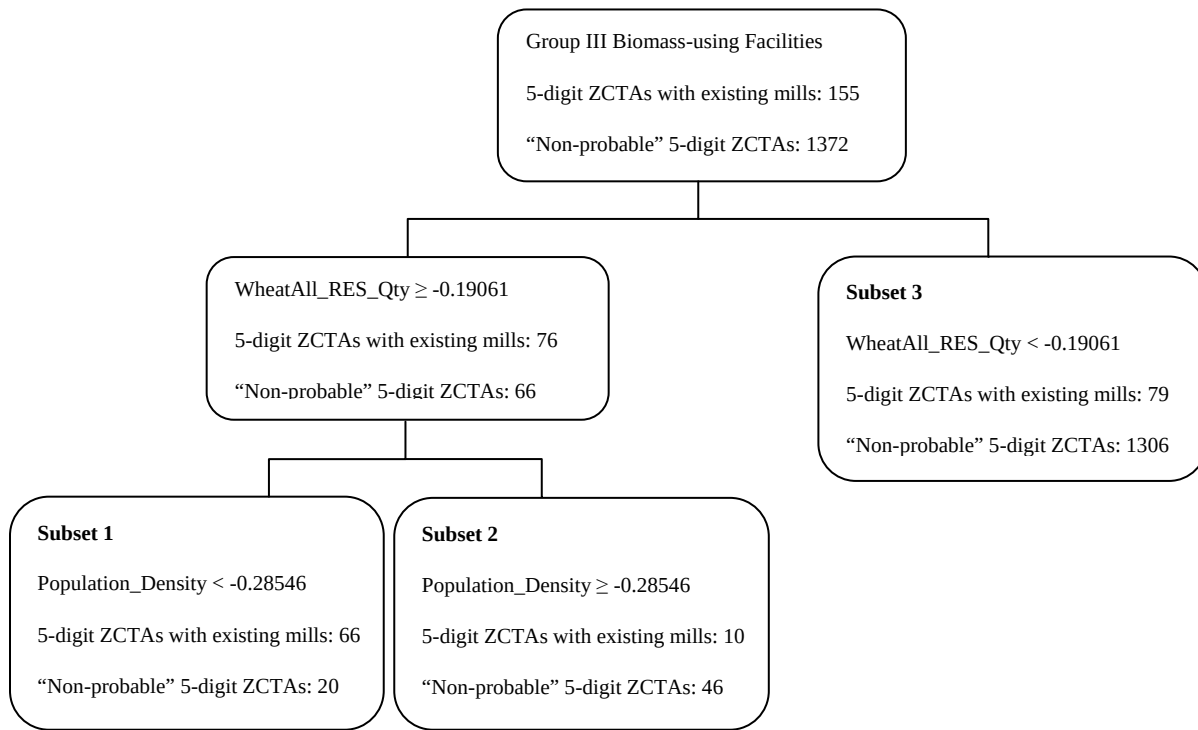


Figure 4-12 The classification tree build by JMP® for Group III biomass-using facilities

Table 4-16 Significant variables for three subsets of Group III biomass-using facilities

Significant Estimates	Whole Set		
	Subset 1	Subset 2	Subset 3
Population_Density			< .0001
Sorghum_RES_Qty		0.0459	
Sqmiwater			< .0001
WheatWinter_RES_HC			
Population	< .0001		
Barley_160_TQty	0.0160		
WheatAll_40_MCost	0.0082		
WheatAll_160_TCost	< .0001		0.0146

Using the same approach as for Group II, statistically significant variables (p value is less than 0.05) for the three subsets determined by the model with the lowest AIC or BIC score are summarized in Table 4-16. For each subset, the statistically significant variables are unique. It is noted that the Sqwiwater variable is significant under Subset 3, which indicates the same association discussed in the section 4.2.3 between the locations of existing Group III biomass-using facilities and water areas.

In the use of the maximum likelihood estimates gained from the classical logistic regression as prior beliefs, the posterior probability distributions under four predefined prior probability distributions for each subset can be obtained and the expected values of the posterior distribution are used as regression coefficients. A misclassification table for maximum likelihood estimation and Bayesian methods is constructed in Table 4-17 to inspect the predictive performance of the local models. For the Subset1, both maximum likelihood estimation and Bayesian methods assuming any one of the four predefined priors provide the same correct classification rates. Bayesian method assuming a Uniform or Cauchy prior has an exactly same correct classification rate of 85.17% as maximum likelihood estimation does for the Subset 2. For the Subset 3, Bayesian method assuming a Uniform prior has the highest correct classification rate of 94.73%, which is also higher than the best correct classification rate of 92.67% for the whole set. This result implies an improvement of predictive power with the local model for the Subset 3 in comparison with the one best model for the whole set.

Table 4-17 Correct classification rates under maximum likelihood estimation and Bayesian methods for 3 subsets and the whole set of Group III biomass-using facilities

Estimation Method		Subset 1 (86)	Subset 2 (56)	Subset 3 (1385)	Whole Set (1527)
Maximum Likelihood Estimation		88.37%	85.71%	94.51%	92.14%
Bayesian Inference	Uniform Prior	88.37%	85.71%	94.73%	92.34%
	Gaussian Prior	88.37%	82.14%	70.61%	92.67%
	Laplace Prior	88.37%	82.14%	94.30%	92.60%
	Cauchy Prior	88.37%	85.71%	94.30%	92.54%

Table 4-18 Classification table by summing up classified values achieved from the “best” local model in each of three subsets for Group III biomass-using facilities

Predicted Value Actual Value	0	1	Total
0	1364 (99.42%)	8	1372
1	83	72 (46.45%)	155
Total	1447	80	1527

A classification table by summing up classified values achieved from the “best” local model in each of the three subsets for Group III biomass-using facilities is constructed in Table 4-18 to view the predictive power of the hybrid model. For the Subset 1, either Bayesian method assuming any one of the four predefined priors or maximum likelihood estimation provides the same classification rate of 88.37% with the specificity of $\frac{13}{20} = 65\%$ and the sensitivity of

$\frac{63}{66} = 95.45\%$. For the Subset 2, Bayesian method assuming a Uniform or Cauchy prior has an exactly same correct classification rate as maximum likelihood estimation does, which correctly classifies all 46 “non-probable” ZCTAs and 2 out of ten ZCTAs with existing biorefinery mills. Bayesian method assuming a Uniform prior gives the highest correct classification rate for the Subset 3, by which 1305 out of 1306 “non-probable” ZCTAs and 7 out of 79 ZCTAs with existing biorefinery mills are correctly classified. Through adding up the corresponding value from each of three subsets, a summarized correctly classified value of the “non-probable” 5-digit ZCTAs is obtained as $13 + 46 + 1305 = 1364$, and another summarized correctly classified value for ZCTAs with existing mills is achieved as $63 + 2 + 7 = 72$. Then, an overall proportion of correct classification is $P(\hat{y} = 1 | y = 1)P(y = 1) + P(\hat{y} = 0 | y = 0)P(y = 0) = 94.04\%$, where the sensitivity is $P(\hat{y} = 1 | y = 1) = 46.45\%$, and the specificity is $P(\hat{y} = 0 | y = 0) = 99.42\%$. Comparing the overall proportion of correct classification derived from only one “best” model in Section 4.2.3 (recall Table 4-11), the hybrid model increases the correct classification rate from 92.67% to 94.04%, where both the specificity and sensitivity have an improvement.

Based on the hybrid model consist of 5 different local models with the highest correct classification rates for the three subsets, a total of 25 potential locations (5-digit ZCTAs) with the highest ranking scores as Group III facilities in the 13 states are predicted and plotted in Figure 4-13. There are nine possible locations in Tennessee, six in Louisiana, four in Texas, two in South Carolina and one in Oklahoma, Mississippi, Arkansas, and Alabama.

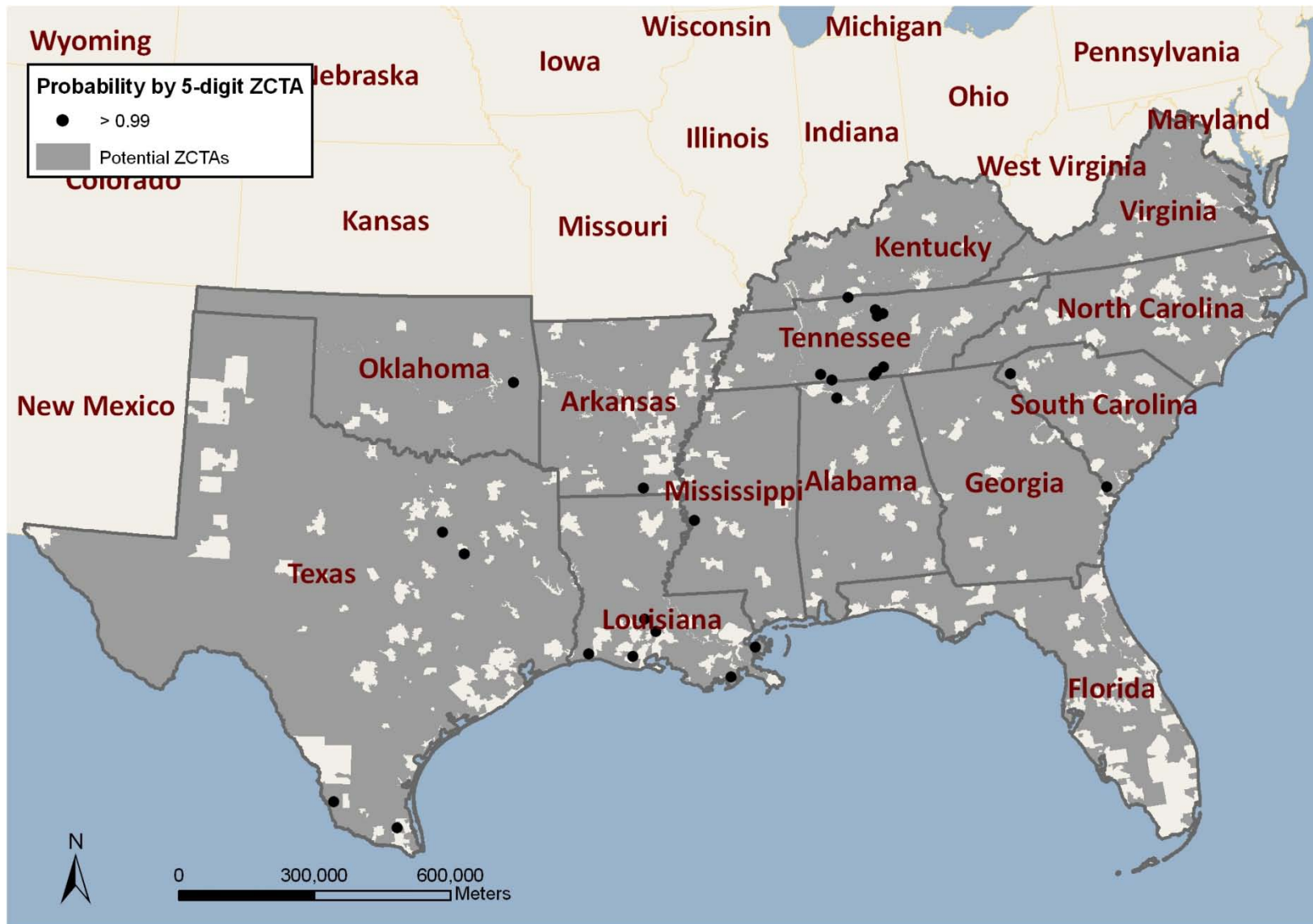


Figure 4-13 Top 25 optimal locations based upon the hybrid model for Group III biomass-using facilities in Southeastern 13 states

Chapter 5 Conclusions and Future Research

Recently, the United States, Europe, China, Indian, and Brazil, whose economies are heavily dependent on petroleum, turn their eyes toward new renewable, long-term sustainable energy sources (e.g. biomass, solar, wind power, hydropower, etc.). In addition, concerns over environmental health, possible exhaustion of oil reserves and national security accelerate the interests of scientists, industry leaders and politicians towards the development of renewable energy technologies. Biofuels derived from biomass resources have offered a promising solution for substitution away from petroleum-based energy, and it becomes one of the fastest growing renewable energy technologies.

Identifying economically favorable sites for biorefineries is an important issue for ensuring the viability of low-cost biofuel production. Since biofuel production is highly geographically dependent on feedstock sources, various factors, i.e., environmental impacts, economic influences, political incentives, and transportation availability, have strong influence on optimal low cost locations. This thesis applies Bayesian Logistic models, which assume four different prior probability distributions in the use of maximum likelihood estimates from classical logistic regression as prior knowledge, to evaluate the factors that mostly influence the location of biomass-using facilities, and to rank existing sites for potential future development of these facilities. Three specific groups of biomass-using facilities are studied. By comparing classical logistic regression with Bayesian logistic regression models under different priors, the model with highest predictive power is found for each of three groups. Population density,

which has negative influence on the relative appeal of a 5-digit ZCTA as a potential location, is the common statistically significant factor for all three groups.

The analysis of Group I biomass-using facilities, which combine all traditionally biomass-using mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants, indicate that eighteen explanatory variables⁵ have statistically significant impacts on the location of a Group I facility. Bayesian estimation assuming a Gaussian prior distribution provides the best predictive power, which indicates that a good location of a Group I facility should be in a 5-digit ZCTA with low median family income, low population density, low harvesting costs of wheat winter all straw, and low cumulative total quantity of certain agricultural residues, i.e., barley straw, corn stover, and sorghum straw, within a haul-driving distance more than 120 miles, but large population size, good railroad accessibility, and high cumulative total cost of wheat all straw residues within 40-mile and 160-mile haul distances.

The best sitting locations (5-digit ZCTAs) of Group II, which contain pulp and paper mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants, also have the same attributes of low population density and low cumulative total quantity of sorghum straw wheat winter all straw residues of a 160-mile haul distance. Bayesian inference using non-informative uniform prior, which keeps almost the same specificity but increases the sensitivity lot, has the highest misclassification rate. Note that primary wood processing mills and secondary wood processing mills are significant variables and have positive impacts on the locations of Group II biomass-using facilities. This observation reveals that existing primary wood processing mills and secondary wood processing mills are important residue providers for

⁵ Median_Family_Income, Sorghum_RES_Qty, WheatWinter_RES_HC, Population, Population_Density, Barley_120_TQty, Corn_80_TCost, Corn_120_TQty, Oat_160_MCost, Oat_160_ACost, Sorghum_160_TQty, WheatAll_40_TCost, WheatAll_120_TQty, WheatAll_160_TCost, WheatAll_160_MCost, WheatWinter_160_TQty, and RailroadAvailability

pulp and paper production and may have a synergistic relationship with Group II biomass-using facilities.

Results of Group III biomass-using facilities, which contain food processing mills, biorefineries using agricultural residues and wood-using bioenergy/biofuels plants, indicate that seven explanatory variables⁶ have statistically significant impacts on the siting locations (5-digit ZCTAs). Bayesian inference using Gaussian prior, Laplace prior or Cauchy prior has a better predictive power, by improving both the sensitivity and specificity rates in the misclassification table, than maximum likelihood estimation does. Estimates from Gaussian prior suggest that a good location of a Group III facility should be in a 5-digit ZCTA with larger water area, high residue quantity of sorghum straws but low population density and low harvesting costs and cumulative trucking total cost of wheat winter all straw residues within an 80-mile haul distance.

Twenty-five optimal locations (ZCTAs) are predicted and plotted in the 13 U.S. Southeastern states for each biomass-using facility group based on the best fitted Bayesian logistic regression model. Mississippi, Kentucky, Tennessee, and Arkansas are predicted to have possible locations of Group I and Group II biomass-using facilities, and Texas, Florida, Louisiana, and North Carolina attract more possible locations of Group III facilities.

In order to further improve the predictive performance for the Group II and Group III biomass-using facilities because of the low sensitivity, this research used a hybrid model proposed by Abu-Hanna and Keizer (2003) that integrates classification trees with local logistic regression, and applies Bayesian inference instead of using maximum likelihood methods for the estimates in local models. Results of Group II biomass-using facilities show that the hybrid

⁶ Population_Density, Sqmiwater, Sorghum_RES_Qty, WheatWinter_RES_HC, Oat_40_ACost, Oat_120_TCost, and WheatWinter_80_TCost

model effectively increases the sensitivity of Group II biomass-using facilities from 58.54% to 64.40%, and has a higher correct classification rate of 96.51% than the one developed best fitted Bayesian logistic regression model does for the whole set. Analysis of Group III biomass-using facilities indicate that both of the specificity and sensitivity are improved significantly by the hybrid model, which eventually increases the correct classification rate from 92.67% to 94.04%. Twenty-five locations with the highest ranking scores in the Southeastern states (5-digit ZCTAs), based upon the hybrid model, are then predicted and mapped out. North Carolina, Tennessee, Virginia, and South Carolina are predicted to have more possible locations for Group II biomass-using facilities, and Tennessee, Louisiana, Texas, and South Carolina attract more locations of Group III biomass-using facilities.

The database built for this study is another important contribution, which not only benefits future research on this topic but also supports data availability and accessibility through the public domain site www.BioSAT.net. The application of GIS to estimate agricultural residues significantly improves data quality and consistency for this database. The transportation cost variables derived from the BioSAT model construct the economic foundation on evaluating the location feasibility and quality of biomass-using facilities. Combining the above data with other socio-demographic data, such as population, employment, water and land areas for each 5-digit ZCTA, this database broadens the biomass characteristics at the level of 5-digit ZCTAs, and adds to the richness of the BioSAT model.

There are several interesting research directions that may improve the analysis effectiveness on optimal location problems. First, collecting more detailed data, such as natural environmental characteristics on the biomass, i.e., soil types, rainfall precipitation, elevation and climate changes; policy regulations; industry spatial competition and state subsidies, would

significantly improve building up our prior knowledge and therefore benefit the Bayesian models predictive ability. Designing a more accurate prior specification and learning algorithm of Bayesian methods can help to explore and obtain a greater understanding of the underlying association between influential factors and location decision. Other statistical methods such as decision tree and Bayesian neural networks could also be applied in this area and may provide better interpretation of the model. Moreover, this approach can be generalized to address similar problems on other spatial regions as long as maintaining a similar data structure.

References

- Abu-Hanna, A., and Keizer, N. (2003). Integrating classification trees with local logistic regression in Intensive Care prognosis. *Artificial Intelligence in Medicine*, 29, 5-23.
- Agresti, A. (1990). *Categorical data analysis*. John Wiley and Sons, New York.
- Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley and Sons, New Jersey.
- Amigun, B., Sigamoney, R., and Blottnitz, H. (2008). Commercialisation of biofuel industry in Africa: a review. *Renewable and Sustainable Energy Reviews*, 12(3), 690-711.
- 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>>.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London*, 53, 370-418.
- Berndes, G., Hoogwijk, M., and Van den Broek, R. (2003). The contribution of biomass in the future global energy supply: a review of 17 studies. *Biomass and Bioenergy*, 25(1), 1-28.
- Brechbill, S.C., and Tyner, W.E., (2008). *The economics of biomass collection, transportation, and supply to Indiana cellulosic and electric utility facilities*. Working Paper #08-03, Dept. of Agricultural Economics, Purdue University.
- Biomass Research and Development Board (2009). *Increasing feedstock production for biofuels: economic drivers, environmental implications, and the role of research* [Web document]. Retrieved from

- < http://www.brdisolutions.com/Site%20Docs/Increasing%20Feedstock_revised.pdf>.
- Box, G. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society*, 143(4), 383-430.
- Brechbill, S. and Tyner, W. (2008). *The economics of renewable energy: corn stover and switchgrass*. Retrieved from <<http://www.ces.purdue.edu/extmedia/ID/ID-404.pdf>>.
- Byrne, J., Shen, B., and Li, X. (1996). The Challenge of sustainability: balancing China's energy, economic and environmental goals. *Energy Policy*, 24(5), 455-462.
- Calle, F., Groot, P., Hemstock S., and Woods J. (2007). *The Biomass assessment handbook: bioenergy for a sustainable environment*. Earthscan, London.
- 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>.
- CSX Corporation, Inc. (2005). *CSX state fact sheets* [Data file]. Available 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>>.
- Demirbas, A. (2007). Progress and recent trends in biofuels. *Progress in Energy and Combustion Science*, 33(1), 1-18.

- Fischer, G. and Schrattenholzer, L. (2002). Global bioenergy potentials through 2050. *Biomass and bioenergy*, 20(3), 151-159.
- Fried, J., Christensen, G., Weyermann, D., ..., Pinjuv, G. (2003). Modeling opportunities and feasibility of siting wood-fired electrical generating facilities to facilitate landscape-scale fuel treatment with FIA BioSum. *Systems Analysis in Forest Resources: Proceedings of the 2003 Symposium*, 207-216.
- Freppaz, D., Minciardi, R., Robba, M., ..., Taramasso, A. (2004). Optimizing forest biomass exploitation for energy supply at a regional level. *Biomass and Bioenergy*, 26(1), 15-25.
- Fulton, L., Howes, T., and Hardy, J. (2004). *Biofuel for transport: an international perspective*. OECD, International Energy Agency, Paris.
- Galik, C., Abt, R., and 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.
- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall, London.
- Genkin, A., Lewis, D., and Madigan D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3), 291-304.
- Graham, R., English, B., and Noon, C. (2000). A geographic information system-based modeling system for evaluating the cost of delivered energy crop feedstock. *Biomass and Bioenergy*, 18(4), 309-329.

- Groscurth, H., Almeida, A., Bauen, A., ..., Widmann, B. (2000). Total costs and benefits of biomass in selected regions of the European Union. *Energy*, 25(11), 1081-1095.
- Hayes, M. (1998). *Agricultural residues: a promising alternative to virgin wood fiber* [Web document]. Retrieved from < <http://www.rca-info.org/alts/meghanhayes.html>>.
- Hilborn, R., Pikitch, E., and McAllister, M. (1994). A Bayesian estimation and decision analysis for an age-structured model using biomass survey data. *Fisheries Research*, 19(1-2), 17-30.
- Hill, J., Nelson, E., Tilman, D., ..., Tiffany, D. (2006). Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of National Academic Science USA*, 103, 11206-11210.
- Holm-Nielsen, J., Madsen, M., and Popiel, P. (2006). Predicted energy crop potentials for bioenergy, worldwide and for EU-25. *World Bioenergy 2006 Conference on Biomass for Energy*, Jonkoping, Sweden.
- Hoogwijk, M., Faaij, A., Van den Broek, R., ..., Turkenburg, W. (2002). Exploration of the ranges of the global potential of biomass for energy. *Biomass and Bioenergy*, 25(2), 119-133.
- Hoskinson, R., Karlen, D., Birrell, S., ..., Wilhelm, W. (2007). Engineering, nutrient removal, and feedstock conversion evaluations of four corn stover harvest scenarios. *Biomass and Bioenergy*, 31(2-3), 136-136.
- Hosmer, D. and Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Sons, New York.

- IEA Bioenergy Task 39 (2009). *Status of 2nd generation biofuels demonstration facilities* [Data file]. Retrieved from <<http://biofuels.abc-energy.at/demoplants>>.
- Jaakkola, T. and Jordan, M. (1996). A variational approach to Bayesian logistic regression models and their extensions. *Proceedings of the sixth international workshop on artificial intelligence and statistics*.
- Jaynes, E. (1957). Information theory and statistical mechanics. *The Physical Review*, 106(4), 620-630.
- Jeffery, H. (1939). *Theory of probability*. Oxford University Press, USA.
- Kamm, B., Kamm, M., and Soye, K. (1998). The green biorefinery, concept of technology. *First international symposium on green biorefinery*, Society of Ecological Technology and System Analysis, Berlin.
- Kamm, B. and Kamm, M. (2004). Applied microbiology and biotechnology. *Principles of biorefineries*, 64(2), 137-145.
- Kim, S. and Dale, B. (2004). Global potential bioethanol production from wasted crops and crop residues. *Biomass and Bioenergy*, 26(4), 361-375.
- Laplace, P. (1774). Memoir on the probability of causes of events (translation by S. Stigler in 1986). *Statistical Science*, 1, 364-378.
- Leduc, S., Schwab, D., Dotzauer, E., ..., Obersteiner, M. (2008). Optimal location of wood gasification plants for methanol production with heat recovery. *International Journal of Energy Research*, 32(12), 1080-1091.

- Liu, X. (2009). *A statistical analysis of key factors influencing the location of biomass-using facilities*. Master Thesis, University of Tennessee, Knoxville, TN.
- Manta (2009). *Food Processing Companies* [Data file]. Available from
<http://www.manta.com/mb_33_C3_000/food_processing>.
- Marshall, G., Shroyer, A., Grover, F., and Hammermeister, K. (1994). Bayesian-logit model for risk assessment in coronary artery bypass grafting. *The Annals of Thoracic Surgery*, 57, 1492-1499.
- Massachusetts Sustainable Forest Bioenergy Initiative (2008). *Woody biomass energy: local renewable fuel for commercial, institutional and industrial facilities* [Web document]. Retrieved from
<<http://www.mass.gov/dcr/stewardship/forestry/utilmark/Wood%20Energy%20Brochure.pdf>>.
- Merriam-Webster (2010). *Biomass* [Biomass definition]. Retrieved from < <http://www.merriam-webster.com/dictionary/biomass> >.
- Mckendry, P. (2002). Energy production from biomass (part 1): overview of biomass. *Bioresource Technology*, 83(1), 37-46.
- 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>>.

- Moon, E., Saveyn, B., Proost, S., and 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.
- Norfolk Southern System (2009). *Norfolk southern system map* [Graphic illustration of Norfolk southern railroad system]. Available from

<<http://www.nscorp.com/nscorp/pdf/systemmap2008.pdf>>.
- Perez-Verdin, G., Grebner, D., ..., Matney, T. (2009). Woody biomass availability for bioethanol conversion in Mississippi. *Biomass and Bioenergy*, 33(3), 492-503.
- Parhizkar, O. and Smith, R. (2008). Application of GIS to estimate the availability of Virginia's biomass residues for bioenergy production. *Forest Products Journal*, 58(2), 71-76.
- Perlack, R., Wright, L., Turhollow, A., ..., Erbach, D. (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.
- Pimentel, D., Moran, M., Fast, S., ..., Young, M. (1981). Biomass energy from crop and forest residues. *Science*, 212(4499), 1110-1115.
- Ptasinski, K., Prins, M., and Pierik, A. (2007). Exergetic evaluation of biomass gasification. *Energy*, 32(4), 568-574.
- Puhan, S., Vedaraman, N., Rambrahamam, B., and Nagarajan, G. (2005). Mahua (*madhuca indica*) seed oil: A source of renewable energy in India. *Journal of Scientific and Industrial Research*, 64(11), 890-896.

- Reijnders, L. (2006). Conditions for the sustainability of biomass based fuel use. *Energy Policy*, 34(7), 863-876.
- Renewable Fuels Association (2009). *Biorefinery locations* [Data file]. Retrieved from <
<http://www.ethanolrfa.org/bio-refinery-locations/> >.
- Savage, L. (1954). *The foundations of statistics*, Dover Publications, New York.
- Shell International Petroleum Company, Inc. (1995). *The Shell global scenarios to 2025* [Web document]. Retrieved from
http://www-static.shell.com/static/aboutshell/downloads/our_strategy/shell_global_scenarios/exsum_23052005.pdf.
- Singh, S., Ekanem, E., Wakefield, T., and Comer, S. (2003). Emerging importance of bio-based products and bio-energy in the U.S. economy: information dissemination and training of students. *World Food and Agribusiness Symposium and Forum*, Cancun, Mexico.
- Smeets, E., Faaij, A., Lewandowski, I., and Turkenburg, W. (2007). A bottom-up assessment and review of global bio-energy potentials to 2050. *Progress in Energy and Combustion Science*, 33(1), 56-106.
- Smith, A. and Spiegelhalter, D. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society*, 42(2), 213-220.
- Smith, M. (2010). Spatial search and fishing location choice: methodological challenges of empirical modeling. *American Journal of Agricultural Economics*, 82(5), 1198-1206.

- Soccol, C., Vandenberghe, L., Costa, B., ..., Medeiros, A. (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.
- Thompson, W., Meyer, S., and Green, T. (2010). The U.S. biodiesel use mandate and biodiesel feedstock markets. *Biomass and Bioenergy*, 34(6), 883-889.
- Union Pacific (2009). *UPRR system map* [Graphic illustration of the Union Pacific railroad system]. Available from <<http://www.uprr.com/aboutup/maps/sysmap.shtml>>.
- 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 and U.S. Department of Energy (2005). *Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply* [Web document]. Retrieved from <http://feedstockreview.ornl.gov/pdf/billion_ton_vision.pdf>.

U.S. Department of Agriculture National Agricultural Statistics Service (2008). *Agricultural Residues* [Data File]. Available from

<http://www.nass.usda.gov/Data_and_Statistics/index.asp>

U.S. Department of Agriculture and Energy (2009). *DOE and USDA select projects for more than \$24 million in biomass research and development grants* [Web document].

Retrieved from

<<http://www.usda.gov/wps/portal/usda/usdahome?navid=ENERGY&navtype=MS>>.

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>.

U.S. Department of Energy (2010). *Bioenergy* [Bioenergy definition]. Retrieved from

< <http://www.energy.gov/energysources/bioenergy.htm>>.

U.S. Department of Energy (2010). *National algal biofuels technology roadmap* [Web document]. Retrieved from

< http://www1.eere.energy.gov/biomass/pdfs/algal_biofuels_roadmap.pdf>.

U.S. National Land Cover Dataset (2000). *Multi-resolution land characteristics consortium* [Data File]. Available from < <http://gisdata.usgs.gov/website/mrlc/viewer.htm> >.

Utah Department of Natural Resources (2010). *Biomass* [Web document]. Retrieved from < http://geology.utah.gov/sep/renewable_energy/biomass.htm >.

- Vries, B., Vuuren, D., and Hoogwijk, M. (2007). Renewable energy sources: their global potential for the first-half of the 21st century at a global level: an integrated approach. *Energy Policy*, 35(4), 2590-2610.
- 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.
- West, M., Harrison, P., and Migon, H. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389), 73-83.
- White, D., and Sifneos, J. (2002). Regression tree cartography. *Journal of Computational and Graphical Statistics*, 11(3), 600-614.
- Wong, A. (1996). *The agri-pulp newsprint alternative*. Presented at the Newspaper Association of America Super Conference, Miami, FL.
- Wright, L. (2006). Worldwide commercial development of bioenergy with a focus on energy crop-based projects. *Biomass and Bioenergy*, 30(8-9), 706-714.
- Xu, J., M.ASCE, M., Fischbeck, P., and M.ASCE, J. (2010). Integrating location models with Bayesian analysis to inform decision making. *Journal of Water Resources Planning and Management*, 136(2), 209-216.
- Xu, Z. and Akella, R. (2008). A Bayesian logistic regression model for active relevance feedback. *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*. 227-234.

- Young, T., Ostermeier, D., Thomas, J., and Brooks, R. (1991). The economic availability of woody biomass for the Southeastern United States. *Bioresource Technology*, 37(1), 7-15.
- Young, T., Hartsell, A., Perdue, J., ..., Rials, T. (2009). A real-time web-based optimal Biomass Site Assessment Tool (BioSAT) – An economic assessment of biomass for the 33 eastern United States. *Proceedings of international conference on woody biomass utilization*, Mississippi State University, Starkville, MS. *In Press*.

Appendices

A-1 ArcGIS codes for Biomass Estimation

```
Private Sub GetRasterPixels_Click()

    Dim pDoc As IMxDocument
    Dim pMap As IMap
    Dim pFLayer As IFeatureLayer
    Dim pFc As IFeatureClass
    Dim pINFeatureClassName As IFeatureClassName
    Dim pDataSet As IDataset
    Dim pInDsName As IDatasetName
    Dim pFSel As IFeatureSelection
    Dim pSelSet As ISelectionSet
    Dim pFeatureClassName As IFeatureClassName
    Dim pOutDatasetName As IDatasetName
    Dim pWorkspaceName As IWorkspaceName
    Dim pExportOp As IExportOperation

    Set pDoc = ThisDocument
    Set pMap = pDoc.FocusMap
    Set pFLayer = pMap.Layer(0)
    Set pFc = pFLayer.FeatureClass

    'Get the FcName from the featureclass
    Set pDataSet = pFc
    Set pINFeatureClassName = pDataSet.FullName
    Set pInDsName = pINFeatureClassName

    'Get the selection set
    Set pFSel = pFLayer
    Set pSelSet = pFSel.SelectionSet
    Dim FeatureCount As Long
    FeatureCount = pFc.FeatureCount(Nothing)

    If pSelSet.Count > 0 Then
        pFSel.Clear
    End If

    Dim i As Long
    For i = 1 To FeatureCount
        Set pRasWS = Nothing
        Set pWSF = Nothing
        Set pWorkspaceFactory = Nothing
        Set pFeatureWorkspace = Nothing

        Dim pQFilter As IQueryFilter
        Set pQFilter = New QueryFilter

        pQFilter.WhereClause = "O_ID2=" & i
        Dim NameValue As String
```

```
Set pSelSet = pFc.Select(pQFilter, esriSelectionTypeHybrid, esriSelectionOptionNormal, Nothing)
```

```
'Define the output feature class name
```

```
Set pFeatureClassName = New FeatureClassName
```

```
Set pOutDatasetName = pFeatureClassName
```

```
pOutDatasetName.Name = i
```

```
Set pWorkspaceName = New WorkspaceName
```

```
pWorkspaceName.PathName = "d:\temp"
```

```
pWorkspaceName.WorkspaceFactoryProgID = "esriCore.shapefileworkspacefactory.1"
```

```
Set pOutDatasetName.WorkspaceName = pWorkspaceName
```

```
pFeatureClassName.FeatureType = esriFTSimple
```

```
pFeatureClassName.ShapeType = esriGeometryAny
```

```
pFeatureClassName.ShapeFieldName = "Shape"
```

```
'Export
```

```
Set pExportOp = New ExportOperation
```

```
pExportOp.ExportFeatureClass pInDsName, Nothing, pSelSet, Nothing, pOutDatasetName, 0
```

```
'Create the RasterExtractionOp object
```

```
Dim pExtractionOp As IExtractionOp
```

```
Set pExtractionOp = New RasterExtractionOp
```

```
'Declare the input objects
```

```
Dim pInputRaster As IRaster
```

```
Dim pMaskfeature As IFeatureClass
```

```
Dim Filename As String
```

```
Filename = i
```

```
'Call function to open the raster datasets
```

```
'Set pInputRasterDataSet = OpenRasterDataset("D:\area_11_landcover", "landcover11_3k_022007")
```

```
'Get raster input from layer
```

```
Dim pRLayer As IRasterLayer
```

```
Dim pRLayerNo As Long
```

```
pRLayerNo = pMap.LayerCount
```

```
Dim k As Integer
```

```
For k = 0 To pRLayerNo - 1
```

```
    If TypeOf pMap.Layer(k) Is IRasterLayer Then
```

```
        Set pRLayer = pMap.Layer(k)
```

```
        Set pInputRaster = pRLayer.Raster
```

```
    End If
```

```
Next k
```

```
Dim Rastername As String
```

```
Rastername = pRLayer.Name
```

```
Set pMaskfeature = OpenFeatureClass("D:\temp", Filename)
```

```

'Get the O_ID Name'
Dim pFeature As IFeature
Dim pFCursor As IFeatureCursor
Set pFCursor = pMaskfeature.Search(Nothing, False)
FieldName = "O_ID"
FieldIndex = pMaskfeature.FindField(FieldName)
Dim NoFeatureCount As Long
NoFeatureCount = pMaskfeature.FeatureCount(Nothing)

For k = 1 To NoFeatureCount
    Set pFeature = pFCursor.NextFeature
    NameValue = pFeature.Value(FieldIndex)
Next k

'Declare the output dataset
Dim pOutputRaster As IRaster

'Call the method
Set pOutputRaster = pExtractionOp.Raster(pInputRaster, pMaskfeature)

Dim pRaster As IRaster
Set pRaster = pOutputRaster

'Get Another Table
Dim pStandTab As IStandaloneTable
Dim pOutTable As ITable
Dim pOutRow As IRow
Dim pOutCursor As ICursor
Dim pOutFieldIndex As Integer
Dim pOutFieldName As String
Dim pStandTabColl As IStandaloneTableCollection
Set pStandTabColl = pMap
Set pStandTab = pStandTabColl.StandaloneTable(0)
Set pOutTable = pStandTab.Table

'Get the Raster Attribute Table
Dim ptable As ITable
Dim pBand As IRasterBand
Dim pBandCol As IRasterBandCollection
Set pBandCol = pRaster
Set pBand = pBandCol.Item(0)
Dim TableExist As Boolean
Set ptable = pBand.AttributeTable
Dim NumOfValue As Integer
NumOfValue = ptable.RowCount(Nothing)

Dim Fieldindex1, Fieldindex2 As Integer
Dim Fieldname1, Fieldname2 As String

Fieldname1 = "VALUE"
Fieldname2 = "COUNT"

```

```

Fieldindex1 = ptable.FindField(Fieldname1)
Fieldindex2 = ptable.FindField(Fieldname2)
Set pDataSet = ptable

'Do the loop to get each attribute value from the Raster Table and Write to the Table
Dim pCursor As ICursor
Dim pRow As IRow
Set pCursor = ptable.Search(Nothing, False)
Set pRow = pCursor.NextRow
Dim pOutRowBuff As IRowBuffer
Set pOutRowBuff = pOutTable.CreateRowBuffer

Dim RowValue As Integer
Dim CountValue As Long

For j = 1 To NumOfValue
    RowValue = 0
    RowValue = pRow.Value(Fieldindex1)
    CountValue = pRow.Value(Fieldindex2)
    Set pRow = pCursor.NextRow

    pOutFieldName = "V_" & RowValue

    If Not pOutFieldName = "" Then
        pOutFieldIndex = pOutTable.FindField(pOutFieldName)

        If pOutFieldIndex >= 0 Then
            pOutRowBuff.Value(pOutFieldIndex) = CountValue
        End If
    End If
Next j

pOutRowBuff.Value(1) = NameValue
Set pOutCursor = pOutTable.Insert(True)
pOutCursor.InsertRow pOutRowBuff
Set pOutRaster = Nothing
Set pInputRaster = Nothing
Set pMaskfeature = Nothing

Next i

MsgBox "Done!"

End Sub

```

```

Private Sub AttachValue_Click()

    Dim pDoc As IMxDocument
    Dim pMap As IMap
    Dim pFLayer As IFeatureLayer
    Dim pFc As IFeatureClass
    Dim pDataSet As IDataset
    Dim pFSel As IFeatureSelection
    Dim pSelSet As ISelectionSet

    Set pDoc = ThisDocument
    Set pMap = pDoc.FocusMap
    Set pFLayer = pMap.Layer(0)
    Set pFc = pFLayer.FeatureClass

    Dim FeatureCount As Long
    FeatureCount = pFc.FeatureCount(Nothing)

    Dim i As Integer
    Dim fieldName As String
    Dim FieldIndex As Integer
    Dim pCursor As IFeatureCursor
    Dim pFeature As IFeature

    Set pCursor = pFc.Search(Nothing, False)
    fieldName = "O_ID2"
    FieldIndex = pFc.FindField(fieldName)

    For i = 1 To FeatureCount

        Set pFeature = pCursor.NextFeature
        pFeature.Value(FieldIndex) = i
        pFeature.Store

    Next i
    MsgBox "Done!"

End Sub

```

A-2 Maps of Agricultural Residue Availability in 13 Southeastern U.S. states

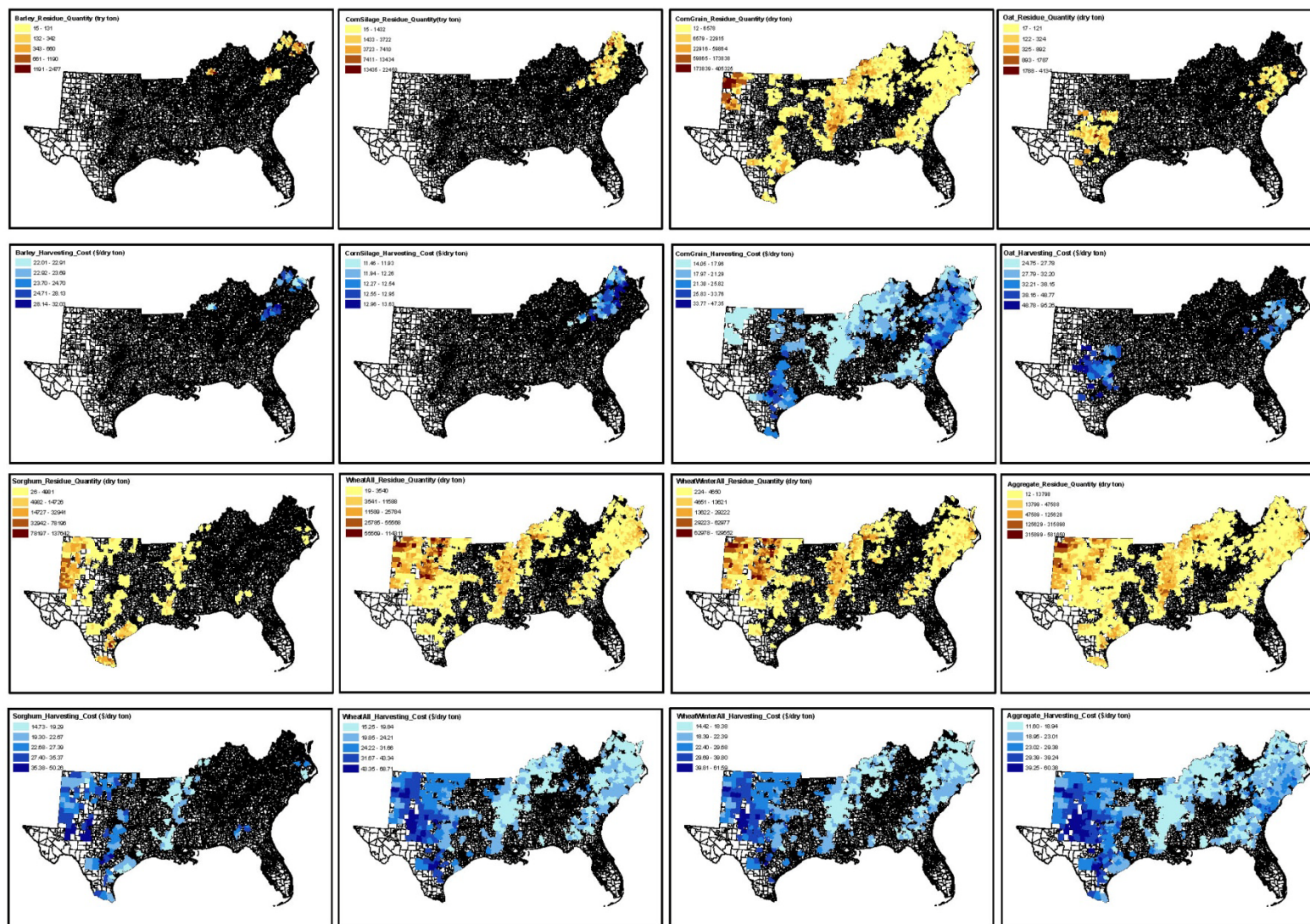


Figure A-2.1 Agricultural residue quantity and harvest costs at the level of 5-digit ZCTA in 13 Southeastern states

A-3 A SAS macro for converting datasets from SAS to R and WinBUGS

/* _LEXPORT Documentation

Original Author: Rodney Sparapani

Export a BUGS or R data file of vectors from a SAS Dataset.

REQUIRED Parameters

FILE= file to create

VAR= list of variables to be included

Specific OPTIONAL Parameters

APPEND= append to data file, instead of creating

CENTER= list of variables to center, i.e. their new mean is zero

CHAIN=1 If CHAIN=1, then generate data file and init file,
if any. Otherwise, generate init file only.
default number of chains to sample
MEAN/PROB summaries are unaltered by default
however, if CHAIN>1, they are calculated
analogously to the chain initializations
of the SAS Bayesian Procedures, i.e.
 $MEAN + ((-1)^{(CHAIN-1)} * (2 + FLOOR(CHAIN/2))) * SE$

CLOSE=1 defaults to closing data file on exit

FORMAT=best12. default format for variables

INIT= default for data file creation rather
than an init file, set to a file name to
be created to over-ride

INITAPPEND= append to init file, instead of creating

INITCLOSE=1 defaults to closing init file on exit

INITINSERT= values to insert into INIT, see INSERT below

INSERT= values to insert into FILE, for TYPE=JAGS
/ force out linefeeds

DATA=_LAST_ default SAS dataset used


```

initclose=1, initinsert=, insert=, linesize=80, logbase=e, logit=, logmean=, ls=&linesize,
mean=, meanlog=, n=N, out=&data, prec=,
prob=, standard=, type=BUGS, log=);

%_require(&file &var);

%if %length(&log) %then %_printto(log=&log);

%local nobs h i j k arg args var0 create miss scratch meanonly preconly savelast
stats closelast /*centerdata*/ logitonly logmeanonly meantype temp;

%let file=%scan(&file, 1, "");
%let type=%upcase(&type);
%let data=%upcase(&data);
%let savelast=&syslast;

%if "&type"="BUGS" %then %do;
  %if %index(%lowcase(&file),.dump) %then %let type=JAGS;
  %else %if %index(%lowcase(&file),.dmp) %then %let type=JAGS;
  %else %if %index(%lowcase(&file),.r) %then %let type=R;
  %else %if %index(%lowcase(&file),.s) %then %let type=S;
%end;

%if "&type"="R" %then %let n=;

%let nobs=%_nobs(data=&data);
%let var=%_blist(&var, data=&data);
%let prob=%_blist(&prob, data=&data, nofmt=1);
%let prec=%_blist(&prec, data=&data, nofmt=1);
%let mean=%_blist(&mean, data=&data, nofmt=1);
%let meanlog=%_blist(&meanlog, data=&data, nofmt=1);
%let logmean=%_blist(&logmean, data=&data, nofmt=1);
%let logit=%_blist(&logit, data=&data, nofmt=1);
%let args=%_count(&var);

%let var0=0;

%do i=1 %to &args;
  %let arg=%scan(&var, &i, %str( ));

  %if %index(&arg, .) %then %let format&var0=&arg;
  %else %do;
    %let var0=%eval(&var0+1);
    %local var&var0 format&var0;
    %let var&var0=&arg;
    %let format&var0=&format;
  %end;
%end;

%if &var0=0 %then %do;
  %put ERROR: no variables found, DATA=&data, VAR=&var;

```

```

    %_abend;
%end;
%else %if &nobs=0 %then %do;
    %put ERROR: no observations found, DATA=&data, NOBS=&nobs;
    %_abend;
%end;

%if "&append"="REQUIRED" %then %let create=1;
%else %let create=0;

%let mean=&mean &meanlog &logmean &logit;
%let stats=%_count(&mean &prec &prob);
%let closelast=0;
%let h=0;

%if %length(&center) %then %do;
    %let scratch=%_scratch;

    proc standard mean=0 &standard data=&data out=&scratch;
        var &center;
    run;

    %let data=&scratch;
%end;

%if %length(&meanlog) %then %do;
    %let logbase=%upcase(&logbase);
    %let scratch=%_scratch;

    data &scratch;
        set &data;

        %do i=1 %to %_count(&meanlog);
            %let temp=%scan(&meanlog, &i, %str( ));

            %if "&logbase"="E" %then &temp=log(&temp);
            %else %if "&logbase"="2" | "&logbase"="10" %then &temp=log&logbase(&temp);
            %else %if "%datatype(&logbase)"="NUMERIC" %then &temp=log(&temp)/log(&logbase);
        %end;
    run;

    %let data=&scratch;
%end;

%do i=1 %to &var0;

    %let preconly=%sysfunc(indexw(%upcase(&prec), %upcase(&&var&i)));
    %let meanonly=%sysfunc(indexw(%upcase(&mean), %upcase(&&var&i)));
    %let logitonly=%sysfunc(indexw(%upcase(&logit), %upcase(&&var&i)));

    %if &logitonly %then %let logmeanonly=0;

```

```

%else %let logmeanonly=%sysfunc(indexw(%upcase(&logmean), %upcase(&&var&i)));

%if &preonly | &meanonly %then %do;
  %if &preonly & &meanonly %then %do;
    %let preonly=0;
    %let meanonly=0;
    %let stats=%eval(&stats-1);
  %end;

  %let h=%eval(&h+1);
  %let scratch=%_scratch;

proc univariate normal plot data=&data;
  var &&var&i;
  output out=&scratch mean=mean_&&var&i var=var_&&var&i n=n_&&var&i;
run;

data &scratch;
  set &scratch;
  prec_&&var&i=1/var_&&var&i;

  %if &chain>1 %then %do;
    %if &logitonly %then mean_&&var&i=mean_&&var&i+
      ((-1)**(&chain-1))*(2+floor(&chain/2))*sqrt(mean_&&var&i*(1-
mean_&&var&i)/n_&&var&i);
    %else %if &logmeanonly %then mean_&&var&i=mean_&&var&i+
      ((-1)**(&chain-1))*(2+floor(&chain/2))*sqrt(mean_&&var&i/n_&&var&i);
    %else mean_&&var&i=mean_&&var&i+
      ((-1)**(&chain-1))*(2+floor(&chain/2))*sqrt(var_&&var&i/n_&&var&i);
  %end;

  drop n_&&var&i var_&&var&i;
  %if &logmeanonly %then %do;
    mean_&&var&i;
    log_mean_&&var&i=log(mean_&&var&i)
    %let meantype=log_mean;
  %end;
  %else %if &logitonly %then %do;
    mean_&&var&i;
    logit_&&var&i=log(mean_&&var&i/(1-mean_&&var&i))
    %let meantype=logit;
  %end;
  %else %let meantype=mean;;
run;

%if &meanonly %then %let var=&meantype._&&var&i;
%else %if &preonly %then %let var=prec_&&var&i;
%else %let var=&meantype._&&var&i prec_&&var&i;

```

```

%if %length(&init) %then %do;
  %if &h=&stats %then %let closelast=&initclose;

  %if &h=1 & %length(&initappend)=0 %then
    %_lexport(insert=&initinsert, file=&init, data=&scratch, var=&var, n=,
      close=&closelast, type=&type);
  %else %_lexport(append=&init, data=&scratch, var=&var, n=,

    close=&closelast, type=&type);
%end;
%else %do;
  %let append=&file;

  %if &h=1 & &create=1 %then %do;
    %_lexport(insert=&insert, file=&file, data=&scratch, var=&var,
      n=%_ifelse(%length(&n), then=&n=&nobs), close=0, type=&type);
    %let insert=;
  %end;
  %else %_lexport(append=&file, data=&scratch, var=&var, n=,
    close=0, type=&type);
%end;
%end;

%if %sysfunc(indexw(%upcase(&prob), %upcase(&&var&i))) %then %do;
  %let h=%eval(&h+1);
  %let scratch=%_scratch;

  proc freq data=&data;
    tables &&var&i / out=&scratch;
  run;

  %local n_&&var&i;

  data &scratch;
    set &scratch end=eof;
    where n(&&var&i);
    retain n_&&var&i 0;

    percent=percent/100;

    n_&&var&i=n_&&var&i+count;

    if eof then call symput("n_&&var&i", trim(left(n_&&var&i)));
  run;

  %let miss=%_nobs(data=&scratch);

  data &scratch;
    drop total percent;
    total=0;

```

```

%do j=1 %to &miss-1;
  j=&j;
  set &scratch point=j;

  %if &chain>1 %then
    percent=min(1, max(0, percent+
      ((-1)**(&chain-1))*(2+floor(&chain/2))*sqrt(percent*(1-percent)/&&&n_&&var&i)));;

    percent=input(put(percent, &format), &format);
    total=total+percent;
    prob_&&var&i..&j=percent;
%end;

%if &chain>1 %then %do;
  if total>1 then do;
    put "ERROR: CHAIN=&chain results in probabilities that sum to >1.";
    put "    Add 1 to CHAIN and try again.";
  end;
%end;

prob_&&var&i..&miss=1-total;
output;
stop;
run;

%end;

%if &chain=1 %then %do;
  %do i=1 %to &var0;
    proc format;
      value _&i._
        ., .A-.Z, .='NA'
        other=[&&format&i]
      ;
    run;
  %end;

  data _null_;
  %if "&append"="REQUIRED" %then %do;
    file "&file" linesize=&ls;

    %if %length(&n) & &var0>1 %then %do;
      put %do i=1 %to &var0-1; "#&&var&i, " %end; "#&&var&var0";
    %end;

    put
      %if "&type"="R" | "&type"="S" %then "%lowercase(%trim(&out)) <- list(";
      %else %if "&type"="BUGS" %then "list(";

    %if %length(&n) %then %do;
      %if "&type"="BUGS" %then %do;

```

```

        %if %index(&n, =) %then "%upcase(&n), ";
        %else "&n=&nobs, ";
    %end;
    %else "`&n` <- &nobs";
%end;

%end;
%else file "&file" linesize=&ls mod;;

%if %length(&insert) %then %do;
    put %if "&type"="JAGS" %then %do;
        %let k=%_count(&insert, split=);

        %do j=1 %to &k-1;
            "%left(%scan(&insert, &j, /))" /

            %end; "%left(%scan(&insert, &k, /))"
        %end;
        %else "&insert,";;
    %end;

%do i=1 %to &var0;
    put %if "&type"="JAGS" %then "%_tr(&&var&i, from=_, to=.)` <- " /;
        %else "%_tr(&&var&i, from=_, to=.) = ";
        @;

    %if &nobs>1 %then put "c(";

/*
    %if %sysfunc(indexw(%upcase(&center), %upcase(&&var&i))) %then %do;
        if _n_=1 then set %scan(&centerdata, &i, %str( )) point=_n_;
    %end;
*/

do i=1 to &nobs;
    set &data(keep=&&var&i) point=i;

    %*if %sysfunc(indexw(%upcase(&center), %upcase(&&var&i))) %then
        if n(&&var&i) then &&var&i=&&var&i-_mean_&&var&i;;

    if i=&nobs then put &&var&i _&i._-r
        %if "&type"="JAGS" %then %do;
            %if &nobs>1 %then ');
            %else ' ';
        %end;
        %else %if &i<&var0 %then %do;
            %if &nobs>1 %then '), ' ;
            %else ' , ' ;
        %end;
        %else %do;
            %if &nobs>1 %then %do;

```

```

        %if &close=1 %then ');
        %else ',';
    %end;
    %else %if &close=1 %then ');
    %else ',';
    %end;
;
    else put &&var&i _&i._-r ',' @;
end;
%end;

stop;
run;
%end;

%let syslast=&savelast;
%if %length(&log) %then %_printto;

%mend _lexport;

%*VALIDATION TEST STREAM;
/* un-comment to re-validate
%_limport(infile=/usr/local/doc/jags/examples/vol1/rats/ratsmiss-data.R, file=ratmis.sas, out=ratmis);

proc print data=_last_;
run;

%_lexport(file=_lexport.txt, data=ratmis, format=3., var=x y, prob=x, mean=x y, prec=y);
*/

```

A-4 WinBUGS and R codes for Bayesian inferences

Uniform prior for Group I biomass-using facilities:

```
model
{
  for( i in 1 : N ) {
    total.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * median_family_income[i] +
beta[3] * railroadavailability1[i] + beta[4] * railroadavailability2[i]+ beta[5] *
Sorghum_RES_Qty[i] + beta[6] * WheatWinter_RES_HC[i] + beta[7] * Population[i] + beta[8]
*Population_Density[i] + beta[9] * Barley_120_TQty[i] +beta[10]*Corn_80_TCost[i]
+beta[11]*Corn_120_TQty[i] +beta[12]*Corn_160_TQty[i]
+beta[13]*Oat_160_MCost[i]+beta[14]*Oat_160_ACost[i]+ beta[15]*Sorghum_160_TQty[i]
+beta[16]*WheatAll_40_TCost [i] +beta[17]*WheatAll_120_TQty [i]
+beta[18]*WheatAll_160_TCost[i] +beta[19]*WheatAll_160_MCost[i]
+beta[20]*WheatWinter_160_TQty[i]
  }
  beta[1] ~ dunif (1,3)
  beta[2] ~ dunif (-0.45,-0.25)
  beta[3] ~ dunif (0.3,0.5)
  beta[4] ~ dunif (-0.06,-0.01)
  beta[5] ~ dunif (-1.3,-1.1)
  beta[6] ~ dunif (-0.085,-0.065)
  beta[7] ~ dunif (0.85,1.11)
  beta[8] ~ dunif (-2.4,-2.2)
  beta[9] ~ dunif (-1.12,-0.95)
  beta[10] ~ dunif (-0.11,0.08)
  beta[11] ~ dunif (-3.6,-3.3)
  beta[12] ~ dunif (5.4,5.8)
  beta[13] ~ dunif (-1.2,-0.98)
  beta[14] ~ dunif (1.2,1.4)
  beta[15] ~ dunif (-1.95,-1.8)
  beta[16] ~ dunif (0.12,0.15)
  beta[17] ~ dunif (1.9,2.2)
  beta[18] ~ dunif (0.72,0.85)
  beta[19] ~ dunif (-1.4,-1.2)
  beta[20] ~ dunif (-5.8,-5.2)
}
```

Gaussian prior for Group I biomass-using facilities:

```

model
{
  for( i in 1 : N ) {
    total.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * median_family_income[i] +
beta[3] * railroadavailability1[i] + beta[4] * railroadavailability2[i]+ beta[5] *
Sorghum_RES_Qty[i] + beta[6] * WheatWinter_RES_HC[i] + beta[7] * Population[i] + beta[8]
*Population_Density[i] + beta[9] * Barley_120_TQty[i] +beta[10]*Corn_80_TCost[i]
+beta[11]*Corn_120_TQty[i] +beta[12]*Corn_160_TQty[i]
+beta[13]*Oat_160_MCost[i]+beta[14]*Oat_160_ACost[i]+ beta[15]*Sorghum_160_TQty[i]
+beta[16]*WheatAll_40_TCost [i] +beta[17]*WheatAll_120_TQty [i]
+beta[18]*WheatAll_160_TCost[i] +beta[19]*WheatAll_160_MCost[i]
+beta[20]*WheatWinter_160_TQty[i]
  }
  beta[1] ~ dnorm(1.5708,1)
  beta[2] ~ dnorm(-0.3815,1)
  beta[3] ~ dnorm(0.3905,1)
  beta[4] ~ dnorm (-0.0392,1)
  beta[5] ~ dnorm(-1.1464,1)
  beta[6] ~ dnorm(-0.0791,1)
  beta[7] ~ dnorm (0.9892,1)
  beta[8] ~ dnorm(-2.3042,1)
  beta[9] ~ dnorm (-1.0940,1)
  beta[10] ~ dnorm(-0.0913,1)
  beta[11] ~ dnorm (-3.4211,1)
  beta[12] ~ dnorm(5.6637,1)
  beta[13] ~ dnorm(-1.0973,1)
  beta[14] ~ dnorm(1.3059,1)
  beta[15] ~ dnorm(-1.8725,1)
  beta[16] ~ dnorm(0.1364,1)
  beta[17] ~ dnorm (2.0452,1)
  beta[18] ~ dnorm (0.7814,1)
  beta[19] ~ dnorm (-1.2930,1)
  beta[20] ~ dnorm(-5.4681,1)
}

```

Laplace prior for Group I biomass-using facilities:

```

model
{
  for( i in 1 : N ) {
    total.biorefscore[i] ~ dbern(p[i])
    ID[i] <- zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * median_family_income[i] +
beta[3] * railroadavailability1[i] + beta[4] * railroadavailability2[i] + beta[5] *
Sorghum_RES_Qty[i] + beta[6] * WheatWinter_RES_HC[i] + beta[7] * Population[i] + beta[8]
*Population_Density[i] + beta[9] * Barley_120_TQty[i] + beta[10]*Corn_80_TCost[i]
+beta[11]*Corn_120_TQty[i] +beta[12]*Corn_160_TQty[i]
+beta[13]*Oat_160_MCost[i]+beta[14]*Oat_160_ACost[i]+ beta[15]*Sorghum_160_TQty[i]
+beta[16]*WheatAll_40_TCost[i] +beta[17]*WheatAll_120_TQty[i]
+beta[18]*WheatAll_160_TCost[i] +beta[19]*WheatAll_160_MCost[i]
+beta[20]*WheatWinter_160_TQty[i]
  }
  beta[1] ~ ddexp(1.5708,1)
  beta[2] ~ ddexp(0.3815,1)
  beta[3] ~ ddexp(0.3905,1)
  beta[4] ~ ddexp(0.0392,1)
  beta[5] ~ ddexp(1.1464,1)
  beta[6] ~ ddexp(0.0791,1)
  beta[7] ~ ddexp(0.9892,1)
  beta[8] ~ ddexp(2.3042,1)
  beta[9] ~ ddexp(1.0940,1)
  beta[10] ~ ddexp(0.0913,1)
  beta[11] ~ ddexp(3.4211,1)
  beta[12] ~ ddexp(5.6637,1)
  beta[13] ~ ddexp(1.0973,1)
  beta[14] ~ ddexp(1.3059,1)
  beta[15] ~ ddexp(1.8725,1)
  beta[16] ~ ddexp(0.1364,1)
  beta[17] ~ ddexp(2.0452,1)
  beta[18] ~ ddexp(0.7814,1)
  beta[19] ~ ddexp(1.2930,1)
  beta[20] ~ ddexp(5.4681,1)
}

```

Cauchy prior for Group I biomass-using facilities:

```
#Read the Data from a txt file;
data<-read.csv(file="Group1.csv", header=TRUE, sep=",");
list(names(data));
n<-dim(data)[1];
n.parms<-dim(data)[2]-2;

#Apply Cauchy prior model;
library(arm);
x1<- as.numeric (t(data[3]));
x2<- as.numeric (t(data[4]));
x3<- as.numeric (t(data[5]));
x4<- as.numeric (t(data[6]));
x5<- as.numeric (t(data[7]));
x6<- as.numeric (t(data[8]));
x7<- as.numeric (t(data[9]));
x8<- as.numeric (t(data[10]));
x9<- as.numeric (t(data[11]));
x10<- as.numeric (t(data[12]));
x11<- as.numeric (t(data[13]));
x12<- as.numeric (t(data[14]));
x13<- as.numeric (t(data[15]));
x14<- as.numeric (t(data[16]));
x15<- as.numeric (t(data[17]));
x16<- as.numeric (t(data[18]));
x17<- as.numeric (t(data[19]));
x18<- as.numeric (t(data[20]));
x19<- as.numeric (t(data[21]));

y<-as.numeric(t(data[2]));

M1<-bayesglm(y ~
x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19,family=binomial(link="logit"), prior.scale=Inf, prior.df=Inf, n.iter = 10000);

display(M1)
y_predict<-predict(M1)

# Get Posterior Distribution
M1.sim<-sim(M1);
```

```

#Get Quantiles
probs<-c(0.025,0.5,0.975)
for (i in 1:20)
{
    out<-capture.output(quantile(M1.sim$coef[,i],probs))
    cat(out,file="Group1_quantile.txt",sep="\n",append=TRUE)
}

#Get Multiple Posterior Probability Plots;
par(mfrow=c(4,5));
plot(density(x0.posterior, bw=0.1),col="red", main=" ", xlab="beta[0]");
plot(density(x1.posterior, bw=0.1),col="red", main=" ", xlab="beta[1]");
plot(density(x2.posterior, bw=0.1),col="red", main=" ", xlab="beta[2]");
plot(density(x3.posterior, bw=0.1),col="red", main=" ", xlab="beta[3]");
plot(density(x4.posterior, bw=0.1),col="red", main=" ", xlab="beta[4]");
plot(density(x5.posterior, bw=0.1),col="red", main=" ", xlab="beta[5]");
plot(density(x6.posterior, bw=0.1),col="red", main=" ", xlab="beta[6]");
plot(density(x7.posterior, bw=0.1),col="red", main=" ", xlab="beta[7]");
plot(density(x8.posterior, bw=0.1),col="red", main=" ", xlab="beta[8]");
plot(density(x9.posterior, bw=0.1),col="red", main=" ", xlab="beta[9]");
plot(density(x10.posterior, bw=0.1),col="red", main=" ", xlab="beta[10]");
plot(density(x11.posterior, bw=0.1),col="red", main=" ", xlab="beta[11]");
plot(density(x12.posterior, bw=0.1),col="red", main=" ", xlab="beta[12]");
plot(density(x13.posterior, bw=0.1),col="red", main=" ", xlab="beta[13]");
plot(density(x14.posterior, bw=0.1),col="red", main=" ", xlab="beta[14]");
plot(density(x15.posterior, bw=0.1),col="red", main=" ", xlab="beta[15]");
plot(density(x16.posterior, bw=0.1),col="red", main=" ", xlab="beta[16]");
plot(density(x17.posterior, bw=0.1),col="red", main=" ", xlab="beta[17]");
plot(density(x18.posterior, bw=0.1),col="red", main=" ", xlab="beta[18]");
plot(density(x19.posterior, bw=0.1),col="red", main=" ", xlab="beta[19]");

```

Uniform prior for Group II biomass-using facilities:

```
model
{
  for( i in 1 : N ) {
    pulppaper.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * Corn_RES_HC[i] +
beta[3]*Population[i] + beta[4] * Population_Density[i] + beta[5] * Primary_Mill_No[i]+ beta[6]
* Barley_120_ACost[i] + beta[7] * Barley_160_TCost[i] + beta[8] * Corn_160_TCost[i] +
beta[9] *Oat_160_TCost[i] + beta[10] * Sorghum_80_TQty[i] +beta[11]*Sorghum_160_TQty[i]
+beta[12]*WheatAll_120_MCost[i] +beta[13]*WheatAll_160_TQty[i]
+beta[14]*WheatWinter_160_TQty[i]+beta[15]*Secondary_Mill_NO[i]
  }
  beta[1] ~ dunif (-5.9,-5.3)
  beta[2] ~ dunif (-0.0002,0.1)
  beta[3] ~ dunif (-0.65,-0.45)
  beta[4] ~ dunif (-1.7,-1.5)
  beta[5] ~ dunif (1.2,1.6)
  beta[6] ~ dunif (-0.002,-0.001)
  beta[7] ~ dunif (0,0.005)
  beta[8] ~ dunif (1,1.3)
  beta[9] ~ dunif (0,0.0008)
  beta[10] ~ dunif (-3.9,-3.3)
  beta[11] ~ dunif (-2.2,-1.5)
  beta[12] ~ dunif (-0.001,0)
  beta[13] ~ dunif (7.3,8)
  beta[14] ~ dunif (-10,-8.5)
  beta[15] ~ dunif (0.2,0.34)
}
```

Gaussian prior for Group II biomass-using facilities:

```

model
{
  for( i in 1 : N ) {
    pulppaper.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * Corn_RES_HC[i] +
beta[3]*Population[i] + beta[4] * Population_Density[i] + beta[5] * Primary_Mill_No[i]+ beta[6]
* Barley_120_ACost[i] + beta[7] * Barley_160_TCost[i] + beta[8] * Corn_160_TCost[i] +
beta[9] *Oat_160_TCost[i] + beta[10] * Sorghum_80_TQty[i] +beta[11]*Sorghum_160_TQty[i]
+beta[12]*WheatAll_120_MCost[i] +beta[13]*WheatAll_160_TQty[i]
+beta[14]*WheatWinter_160_TQty[i]+beta[15]*Secondary_Mill_NO[i]
  }
  beta[1] ~ dnorm (-5.6,1)
  beta[2] ~ dnorm (0,1)
  beta[3] ~ dnorm (0.55,1)
  beta[4] ~ dnorm (-1.6,1)
  beta[5] ~ dnorm (1.42,1)
  beta[6] ~ dnorm (0,1)
  beta[7] ~ dnorm (0,1)
  beta[8] ~ dnorm (1.17,1)
  beta[9] ~ dnorm (0,1)
  beta[10] ~ dnorm (-3.54,1)
  beta[11] ~ dnorm (-1.81,1)
  beta[12] ~ dnorm (0,1)
  beta[13] ~ dnorm (7.8,1)
  beta[14] ~ dnorm (-9.19,1)
  beta[15] ~ dnorm (0.3,1)
}

```

Laplace prior for Group II biomass-using facilities:

```
model
{
  for( i in 1 : N ) {
    pulppaper.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * Corn_RES_HC[i] +
beta[3]*Population[i] + beta[4] * Population_Density[i] + beta[5] * Primary_Mill_No[i]+ beta[6]
* Barley_120_ACost[i] + beta[7] * Barley_160_TCost[i] + beta[8] * Corn_160_TCost[i] +
beta[9] *Oat_160_TCost[i] + beta[10] * Sorghum_80_TQty[i] +beta[11]*Sorghum_160_TQty[i]
+beta[12]*WheatAll_120_MCost[i] +beta[13]*WheatAll_160_TQty[i]
+beta[14]*WheatWinter_160_TQty[i]+beta[15]*Secondary_Mill_NO[i]
  }
  beta[1] ~ ddexp(-5.6247,1)
  beta[2] ~ ddexp(-0.00015,1)
  beta[3] ~ ddexp(-0.5539,1)
  beta[4] ~ ddexp (-1.6069,1)
  beta[5] ~ ddexp(1.4167,1)
  beta[6] ~ ddexp(-0.00134,1)
  beta[7] ~ ddexp(0.00161,1)
  beta[8] ~ ddexp(1.1713,1)
  beta[9] ~ ddexp(0.00016,1)
  beta[10] ~ ddexp(-3.5397,1)
  beta[11] ~ ddexp(-1.8137,1)
  beta[12] ~ ddexp(-0.00021,1)
  beta[13] ~ ddexp(7.8371,1)
  beta[14] ~ ddexp(-9.193,1)
  beta[15] ~ ddexp(0.2969,1)
}
```

Cauchy prior for Group II biomass-using facilities:

```
#Read the Data from a txt file;
data<-read.csv(file="Group2.csv", header=TRUE, sep=",");
list(names(data));
n<-dim(data)[1];
n.parms<-dim(data)[2]-2;

#Apply Cauchy prior model;
library(arm);
x1<- as.numeric (t(data[3]));
x2<- as.numeric (t(data[4]));
x3<- as.numeric (t(data[5]));
x4<- as.numeric (t(data[6]));
x5<- as.numeric (t(data[7]));
x6<- as.numeric (t(data[8]));
x7<- as.numeric (t(data[9]));
x8<- as.numeric (t(data[10]));
x9<- as.numeric (t(data[11]));
x10<- as.numeric (t(data[12]));
x11<- as.numeric (t(data[13]));
x12<- as.numeric (t(data[14]));
x13<- as.numeric (t(data[15]));
x14<- as.numeric (t(data[16]));

y<-as.numeric(t(data[2]));

M1<-bayesglm(y ~
x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14,family=binomial(link="logit"),
prior.scale=Inf, prior.df=Inf, n.iter = 10000);

display(M1)
y_predict<-predict(M1)

# Get Posterior Distribution
M1.sim<-sim(M1);

#Get Quantiles
probs<-c(0.025,0.5,0.975)
for (i in 1:15)
{
out<-capture.output(quantile(M1.sim$coef[,i],probs))
cat(out,file="Group2_quantile.txt",sep="\n",append=TRUE)
}
```

```

#Get Multiple Posterior Probability Plots;
par(mfrow=c(4,5));
plot(density(x0.posterior, bw=0.1),col="red", main=" ", xlab="beta[0]");
plot(density(x1.posterior, bw=0.1),col="red", main=" ", xlab="beta[1]");
plot(density(x2.posterior, bw=0.1),col="red", main=" ", xlab="beta[2]");
plot(density(x3.posterior, bw=0.1),col="red", main=" ", xlab="beta[3]");
plot(density(x4.posterior, bw=0.1),col="red", main=" ", xlab="beta[4]");
plot(density(x5.posterior, bw=0.1),col="red", main=" ", xlab="beta[5]");
plot(density(x6.posterior, bw=0.1),col="red", main=" ", xlab="beta[6]");
plot(density(x7.posterior, bw=0.1),col="red", main=" ", xlab="beta[7]");
plot(density(x8.posterior, bw=0.1),col="red", main=" ", xlab="beta[8]");
plot(density(x9.posterior, bw=0.1),col="red", main=" ", xlab="beta[9]");
plot(density(x10.posterior, bw=0.1),col="red", main=" ", xlab="beta[10]");
plot(density(x11.posterior, bw=0.1),col="red", main=" ", xlab="beta[11]");
plot(density(x12.posterior, bw=0.1),col="red", main=" ", xlab="beta[12]");
plot(density(x13.posterior, bw=0.1),col="red", main=" ", xlab="beta[13]");
plot(density(x14.posterior, bw=0.1),col="red", main=" ", xlab="beta[14]");

```

Uniform prior for Group III biomass-using facilities:

```

model
{
  for( i in 1 : N ) {
    foodprocess.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * Population_Density[i] +
beta[3]*WheatWinter_80_TCost[i] + beta[4] * Oat_120_TCost[i] + beta[5] * Oat_40_ACost[i]+
beta[6] * Sorghum_RES_Qty[i] + beta[7] * Sqmiwater[i] + beta[8] *WheatWinter_RES_HC[i]
  }
  beta[1] ~ dunif (-1.95,-1.55)
  beta[2] ~ dunif (-1.75,-1.25)
  beta[3] ~ dunif (-0.25,-0.13)
  beta[4] ~ dunif (-0.18,-0.12)
  beta[5] ~ dunif (0,0.8)
  beta[6] ~ dunif (1.0,2.5)
  beta[7] ~ dunif (2.0,4.0)
  beta[8] ~ dunif (-0.2,-0.1)
}

```

Gaussian prior for Group III biomass-using facilities:

```

model
{
  for( i in 1 : N ) {
    foodprocess.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * Population_Density[i] +
beta[3]*WheatWinter_80_TCost[i] + beta[4] * Oat_120_TCost[i] + beta[5] * Oat_40_ACost[i]+
beta[6] * Sorghum_RES_Qty[i] + beta[7] * Sqmiwater[i] + beta[8] *WheatWinter_RES_HC[i]
  }
  beta[1] ~ dnorm (-1.7149,1)
  beta[2] ~ dnorm (-1.4671,1)
  beta[3] ~ dnorm(-0.1805,1)
  beta[4] ~ dnorm(-0.1532,1)
  beta[5] ~ dnorm(0.3019,1)
  beta[6] ~ dnorm(1.7449,1)
  beta[7] ~ dnorm(2.9637,1)
  beta[8] ~ dnorm(-0.1606,1)
}

```

Laplace prior for Group III biomass-using facilities:

```
model
{
  for( i in 1 : N ) {
    foodprocess.biorefscore[i] ~ dbern(p[i])
    ID[i]<-zcta[i]
    logit(p[i]) <- beta[1] + beta[2] * Population_Density[i] +
beta[3]*WheatWinter_80_TCost[i] + beta[4] * Oat_120_TCost[i] + beta[5] * Oat_40_ACost[i]+
beta[6] * Sorghum_RES_Qty[i] + beta[7] * Sqmiwater[i] + beta[8] *WheatWinter_RES_HC[i]
  }
  beta[1] ~ ddexp (-1.7149,1)
  beta[2] ~ ddexp (-1.4671,1)
  beta[3] ~ ddexp (-0.1805,1)
  beta[4] ~ ddexp (-0.1532,1)
  beta[5] ~ ddexp (0.3019,1)
  beta[6] ~ ddexp (1.7449,1)
  beta[7] ~ ddexp (2.9637,1)
  beta[8] ~ ddexp (-0.1606,1)
}
```

Cauchy prior for Group III biomass-using facilities:

```
#Read the Data from a txt file;
data<-read.csv(file="Group3_WithoutPSO.csv", header=TRUE, sep=",");
list(names(data));
n<-dim(data)[1];
n.parms<-dim(data)[2]-2;

#Apply Cauchy prior model;
library(arm);
x1<- as.numeric (t(data[3]));
x2<- as.numeric (t(data[4]));
x3<- as.numeric (t(data[5]));
x4<- as.numeric (t(data[6]));
x5<- as.numeric (t(data[7]));
x6<- as.numeric (t(data[8]));
x7<- as.numeric (t(data[9]));

y<-as.numeric(t(data[2]));
M1<-bayesglm(y ~ x1+x2+x3+x4+x5+x6+x7,family=binomial(link="logit"), prior.scale=Inf,
prior.df=Inf, n.iter = 10000);

display(M1)
```

```

y_predict<-predict(M1)

# Get Posterior Distribution
M1.sim<-sim(M1);

#Get Quantiles
probs<-c(0.025,0.5,0.975)
for (i in 1:8)
{
out<-capture.output(quantile(M1.sim$coef[,i],probs))
cat(out,file="Group3_WithoutPSO_quantile.txt",sep="\n",append=TRUE)
}

#Get Multiple Posterior Probability Plots;
par(mfrow=c(4,4));
plot(density(x0.posterior, bw=0.1),col="red", main=" ", xlab="beta[0]");
plot(density(x1.posterior, bw=0.1),col="red", main=" ", xlab="beta[1]");
plot(density(x2.posterior, bw=0.1),col="red", main=" ", xlab="beta[2]");
plot(density(x3.posterior, bw=0.1),col="red", main=" ", xlab="beta[3]");
plot(density(x4.posterior, bw=0.1),col="red", main=" ", xlab="beta[4]");
plot(density(x5.posterior, bw=0.1),col="red", main=" ", xlab="beta[5]");
plot(density(x6.posterior, bw=0.1),col="red", main=" ", xlab="beta[6]");
plot(density(x7.posterior, bw=0.1),col="red", main=" ", xlab="beta[7]");

```

A-5 Estimations of Bayesian inferences

Table A-5.1 Parameter estimates under Uniform prior for Group I biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	1.44	0.1265	1.196	1.44	1.691
Median_Family_Income	-0.4137	0.0272	-0.4487	-0.419	-0.3485
RailRoadAvailability 1	0.4067	0.05586	0.3069	0.4093	0.4953
RailRoadAvailability 2	-0.02882	0.01356	-0.05682	-0.0263	-0.01067
Sorghum_RES_Qty	-1.171	0.05191	-1.283	-1.161	-1.103
WheatWinter_RES_HC	-0.07365	0.005491	-0.08412	-0.07308	-0.06544
Population	1.017	0.06194	0.8794	1.026	1.106
Population_Density	-2.299	0.05491	-2.393	-2.298	-2.206
Barley_120_TQty	-1.023	0.04647	-1.112	-1.018	-0.9532
Corn_80_TCost	-0.08029	0.01536	-0.1065	-0.08114	-0.04868
Corn_120_TQty	-3.406	0.0753	-3.573	-3.392	-3.304
Corn_160_TQty	5.671	0.09195	5.461	5.689	5.795
Oat_160_MCost	-1.083	0.05192	-1.186	-1.076	-1.009
Oat_160_ACost	1.275	0.05381	1.203	1.266	1.382
Sorghum_160_TQty	-1.863	0.04087	-1.942	-1.858	-1.803
WheatAll_40_TCost	0.1327	0.007867	0.1207	0.132	0.1481
WheatAll_120_TQty	2.059	0.0834	1.911	2.064	2.193
WheatAll_160_TCost	0.7857	0.03625	0.7244	0.7853	0.8459
WheatAll_160_MCost	-1.284	0.04407	-1.365	-1.283	-1.208
WheatWinter_160_TQty	-5.506	0.1363	-5.759	-5.505	-5.245

Table A-5.2 Parameter estimates under Laplace prior for Group I biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	1.538	0.2338	1.091	1.561	1.931
Median_Family_Income	-0.4384	0.06545	-0.5414	-0.435	-0.3345
RailRoadAvailability 1	0.4952	0.1708	0.1731	0.4943	0.8136
RailRoadAvailability 2	0.2499	0.2222	0.06275	0.2662	0.4671
Sorghum_RES_Qty	-0.9106	0.2333	-1.37	-0.8944	-0.4898
WheatWinter_RES_HC	-0.04409	0.0224	-0.06705	-0.04476	-0.02019
Population	1.196	0.1512	0.9641	1.186	1.515
Population_Density	-2.416	0.1927	-2.725	-2.41	-2.156
Barley_120_TQty	-0.6578	0.223	-0.9804	-0.6505	-0.3555
Corn_80_TCost	-0.07383	0.04739	-0.1131	-0.07542	-0.04272
Corn_120_TQty	-2.447	0.8971	-3.347	-2.683	-0.4248
Corn_160_TQty	5.552	0.6185	3.865	5.677	6.368
Oat_160_MCost	1.701	0.1167	1.538	1.666	1.915
Oat_160_ACost	-1.7	0.1304	-1.945	-1.662	-1.518
Sorghum_160_TQty	-2.553	0.3929	-2.943	-2.518	-2.232
WheatAll_40_TCost	0.127	0.0201	0.0965	0.1258	0.1578
WheatAll_120_TQty	1.163	0.9334	-1.441	1.474	2.238
WheatAll_160_TCost	0.04371	0.2348	-0.5031	0.1546	0.274
WheatAll_160_MCost	-0.4213	0.2627	-0.681	-0.5507	0.1927
WheatWinter_160_TQty	-3.634	1.124	-4.966	-4.026	-0.8357

Table A-5.3 Parameter estimates under Cauchy prior for Group I biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	0.11	0.13	-0.18011	0.101677	0.394132
Median_Family_Income	-0.26	0.04	-0.34568	-0.25561	-0.19161
RailRoadAvailability 1	0.43	0.12	0.215877	0.425573	0.700413
RailRoadAvailability 2	0.23	0.09	0.092131	0.21999	0.431203
Sorghum_RES_Qty	-0.59	0.14	-0.84587	-0.60316	-0.33518
WheatWinter_RES_HC	-0.02	0.01	-0.04815	-0.02384	-0.00086
Population	0.61	0.10	0.41956	0.606557	0.78078
Population_Density	-1.40	0.10	-1.56566	-1.39439	-1.19745
Barley_120_TQty	-0.51	0.12	-0.73342	-0.50798	-0.28539
Corn_80_TCost	-0.05	0.01	-0.08096	-0.05115	-0.02588
Corn_120_TQty	-0.05	0.25	-0.56076	-0.11904	0.397878
Corn_160_TQty	2.28	0.25	1.918289	2.355351	2.76611
Oat_160_MCost	0.00	0.03	-0.04541	-0.00249	0.045228
Oat_160_ACost	0.10	0.03	0.03858	0.101846	0.143986
Sorghum_160_TQty	-1.78	0.10	-1.97384	-1.77607	-1.64715
WheatAll_40_TCost	0.08	0.01	0.061165	0.083236	0.105998
WheatAll_120_TQty	-0.70	0.24	-1.17034	-0.6751	-0.29378
WheatAll_160_TCost	-0.20	0.05	-0.32134	-0.19939	-0.10685
WheatAll_160_MCost	-0.09	0.06	-0.18725	-0.08467	0.047066
WheatWinter_160_TQty	-0.83	0.24	-1.19428	-0.82867	-0.35255

Table A-5.4 Parameter estimates under Gaussian prior for Group II biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-5.460	0.1521	-5.845	-5.461	-5.309
Corn_RES_HC	-0.00012	0.00015	-0.00018	-0.00015	-0.0002
Population	-0.5113	0.0504	-0.632	-0.5107	-0.4526
Population_Density	-1.595	0.0557	-1.696	-1.608	-1.51
Primary_Mill_No	1.362	0.1021	1.139	1.37	1.583
Barley_120_ACost	-0.00102	0.00011	-0.0011	-0.00104	-0.001
Barley_160_TCost	0.0011	0.0001	0.0012	0.00128	0.00153
Corn_160_TCost	1.179	0.0749	1.026	1.1990	1.298
Oat_160_TCost	0.0002	0.00002	0.00014	0.00021	0.00029
Sorghum_80_TQty	-3.325	0.04641	-3.457	-3.333	-3.303
Sorghum_160_TQty	-1.543	0.05733	-1.705	-1.54	-1.501
WheatAll_120_MCost	-0.00013	0.00005	-0.00033	-0.00017	-0.00006
WheatAll_160_TQty	7.66	0.196	7.346	7.67	7.988
WheatWinter_160_TQty	-8.93	0.243	-9.434	-8.941	-8.549
Secondary_Mill_No	0.3062	0.0287	0.231	0.3132	0.342

Table A-5.5 Parameter estimates under Laplace prior for Group II biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-5.639	1.41	-8.76	-5.625	-2.624
Corn_RES_HC	-0.0002822	1.415	-3.157	-0.0177	2.902
Population	-0.5446	1.462	-3.629	-0.5436	2.474
Population_Density	-1.642	1.403	-4.699	-1.658	1.268
Primary_Mill_No	1.457	1.356	-1.427	1.429	4.377
Barley_120_ACost	-0.001462	1.451	-3.138	-0.01513	3.103
Barley_160_TCost	0.001221	1.479	-3.026	0.05868	3.152
Corn_160_TCost	1.093	1.411	-1.849	1.111	4.031
Oat_160_TCost	0.0002436	1.354	-3.066	0.001652	2.761
Sorghum_80_TQty	-3.58	1.384	-6.566	-3.565	-0.6699
Sorghum_160_TQty	-1.812	1.441	-4.783	-1.812	1.087
WheatAll_120_MCost	-0.0001142	1.382	-2.683	0.002104	3.048
WheatAll_160_TQty	7.796	1.36	4.835	7.792	10.52
WheatWinter_160_TQty	-9.148	1.435	-11.99	-9.175	-5.969
Secondary_Mill_No	0.3112	1.338	-2.559	0.3317	3.094

Table A-5.6 Parameter estimates under Cauchy prior for Group II biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-3.78	0.46	-4.8374	-3.77665	-2.8918
Corn_RES_HC	0	0	-1.71E-04	-1.28E-04	-9.61E-05
Population	-0.15	0.12	-0.41929	-0.14151	0.069555
Population_Density	-1.61	0.28	-2.15163	-1.64958	-1.08193
Primary_Mill_No	1.29	0.14	1.047941	1.296951	1.5787
Barley_120_ACost	0	0	-0.0003	-0.0001	7.48E-05
Barley_160_TCost	0	0	-3.76E-05	1.16E-04	3.21E-04
Corn_160_TCost	0	0	-1.10E-04	-3.10E-05	7.12E-05
Oat_160_TCost	0	0	0.000215	0.000274	0.000334
Sorghum_80_TQty	-0.04	0.28	-0.64324	-0.04735	0.578372
Sorghum_160_TQty	-0.71	0.28	-1.34609	-0.7635	-0.18357
WheatAll_120_MCost	0	0	-0.00022	-0.00014	-6.2E-05
WheatAll_160_TQty	1.36	1.81	-1.30273	1.665798	5.88296
WheatWinter_160_TQty	-2.24	1.82	-6.85788	-2.57049	0.542865
Secondary_Mill_No	0.39	0.08	0.270391	0.400214	0.526993

Table A-5.7 Parameter estimates under Uniform prior for Group III biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-1.71	0.1085	-1.929	-1.694	-1.557
Population_Density	-1.519	0.1369	-1.739	-1.527	-1.271
WheatWinter_80_TCost	-0.1606	0.02324	-0.2164	-0.1555	-0.1312
Oat_120_TCost	-0.1368	0.01327	-0.1688	-0.1336	-0.1206
Oat_40_ACost	0.2967	0.02516	0.2484	0.2967	0.3466
Sorghum_RES_Qty	1.529	0.3103	1.044	1.495	2.208
Sqmiwater	2.608	0.3571	2.045	2.572	3.41
WheatWinter_RES_HC	-0.1756	0.01755	-0.199	-0.1787	-0.1344

Table A-5.8 Parameter estimates under Laplace prior for Group III biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-1.466	0.545	-2.693	-1.421	-0.6653
Population_Density	-1.536	0.3108	-2.106	-1.52	-0.9866
WheatWinter_80_TCost	-0.1379	0.05149	-0.2243	-0.1372	-0.05963
Oat_120_TCost	-0.09876	0.03681	-0.1617	-0.1023	-0.01984
Oat_40_ACost	0.2483	0.05736	0.1373	0.2505	0.3531
Sorghum_RES_Qty	1.479	0.341	0.8286	1.487	2.113
Sqmiwater	2.47	0.4318	1.62	2.476	3.286
WheatWinter_RES_HC	-0.1906	0.04865	-0.2502	-0.1945	-0.1326

Table A-5.9 Parameter estimates under Cauchy prior for Group III biomass-using facilities

Parameter	Mean	Standard Deviation	2.5%	Median	97.5%
Intercept	-1.22	0.33	-1.77333	-1.25387	-0.55864
Population_Density	-1.59	0.31	-2.16389	-1.59939	-0.98528
WheatWinter_80_TCost	-0.13	0.04	-0.19621	-0.12402	-0.05981
Oat_120_TCost	-0.10	0.03	-0.17272	-0.10066	-0.03857
Oat_40_ACost	0.23	0.05	0.145262	0.230403	0.332664
Sorghum_RES_Qty	1.33	0.36	0.826919	1.379602	1.90479
Sqmiwater	2.23	0.44	1.474683	2.256987	3.136598
WheatWinter_RES_HC	-0.20	0.03	-0.25568	-0.1969	-0.11372

Vita

Xia Huang is a Graduate Research Assistant in the Center for Renewable Carbon at the University of Tennessee, Knoxville. She plans to graduate from the University of Tennessee with a Master of Science degree in Statistics in December 2010. She also has a M.S. Degree (concentration Geographic Information System (GIS) -Transportation) in Geography from the University of Tennessee and B.S. (Geography) from Beijing Normal University. She worked as a Graduate Research Assistant during the period of 2006-2007 participating in a National Science Foundation Project (# BCS-0616724), and had two years as a graduate teaching assistant tutoring undergraduates and graduates in their introductory statistics and GIS courses. In 2006, she earned the excellent thesis award for Bachelor Degree from Beijing Normal University. In 2010, she presented “Enhancing the BioSAT Model: Assessing Site Location for Biorefineries using Bayesian Logistic Regression” in the United State Forest Service FIA Symposium. She also has been an invited speaker in EXPO 2005 International Forum in Aichi Japan and an environmental ambassador of Beijing studying in Denmark, 2001. She is a member of the American Statistical Association, and Association of American Geographers (AAG). She was session chair in the AAG - San Francisco, 2007 and AAG – Boston, 2008.