A hybrid model integrating Classification Tree with local Bayesian Logistic Regression to Assess Preselected Site Locations for Biorefineries

Xia Huang¹, James H Perdue², Timothy M. Young¹, Frank M. Guess³

¹ Center for Renewable Carbon, The University of Tennessee, Knoxville
² U.S. Forest Service, Southern Research Station
³ Department of Statistics, Operations, and Management Science, The University of Tennessee, Knoxville
Biomass Site Assessment Tool

“Identify possible site locations from a statistical perspective”
# Research Goal & Objectives

## Research Goal

- Quantify influential variables on the location of wood-using biorefineries.
- Identify possible site locations for potential wood-using biorefineries.

## Objectives

- Develop an expanded database from BIOSAT, including social economic variables, woody residue availability variables, transportation-related variables, land cover variables, etc.
- Focus on data quality and consistency in the use of geographic spatial analysis.
- Identify variables influencing locations for wood-using biorefineries.
- Predict possible locations for potential wood-using biorefineries.
Research Scope and Resolution

Introduction

Regional Level
(All 13 Southeastern States together)
Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, Oklahoma, North Carolina, South Carolina, Tennessee, Texas, Virginia.

Resolution
U.S. Census Bureau ZIP Code Tabulation Areas
(5-digit ZCTA)
Total 5-digit ZCTAs: 10,016
**Definition**

- A biorefinery is a facility that integrates biomass conversion processes and equipment to produce fuels, power, or value-added chemicals from **biomass**.

**Type of Biomass**

- Wood
- Crops
- Garbage
- Landfill Gas
- Alcohol Fuels
Biomass-using Facilities

New-emerging biomass-using facilities

- Biorefineries that use woody/agricultural residues

Traditional biomass-using facilities

- Primary mills
- Secondary mills
  - Pulp and paper mills
“Not Suitable” Location for Wood-using Biorefineries

Federal Lands

Population Density > 150 people/mile
(Wear et al. 1999)

Slope > 40% (21.8°)
(Kimsey et al. 2011)

Not Suitable Ecoregions

Data
“Not Suitable” Location for Wood-using Biorefineries (cont.)

Total 5-digit ZCTAs excluded: 5,019
(5,019/10,016=50.1%)
**Response Variable**
- Y=1 for location of existing biorefinery using woody residues
- Y=0 for location of primary mills with small capacity (<779 dry tons)

**Explanatory Variable**

- **Social/Economic Variables**
  - Population density
  - Median family income
  - Farm net income
  - Road density

- **Woody Residue Availability**
  - Residue quantity
  - Residue harvesting cost
  - Growth / removal ratio

- **Transportation-Related Variables**
  - Transportation cost
  - Railroad availability
  - # of Water ports

- **Land Cover Variables**
  - Forest area ratio
  - Urban land ratio
  - Water area ratio
  - Slope
Logistic Regression

\[
\text{logit}[\theta_i] = \log_e \frac{\theta_i}{1 - \theta_i} \text{ where } \theta_i = \text{probability of occurrence of the event}
\]

\[
\log_e \frac{\theta_i}{1 - \theta_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n
\]

\[
\theta_i = \frac{1}{1 + e^{-[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n]}}
\]

- No assumption on linearity and normality

Bayes’ Theorem

\[
p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta) p(y \mid \theta)}{p(y)}
\]

- Prior and posterior probability distributions
- Extend logistic regression model in a Bayesian framework (Xu and Akella 2008)
- Use Bayesian Inference Methods for coefficient estimates ($\beta$)
In Mathematics

\[ p(y = 1 \mid x, M, D) = \int_{\beta} p(y = 1, \beta \mid x, M, D) d\beta \]

\[ = \int_{\beta} p(y = 1 \mid x, \beta, D) p(\beta \mid M, D) d\beta \]

where

\[ p(y = 1 \mid x, \beta, D) = \{1 + \exp[-\beta^T x]\}^{-1} \]

Priors

- **Non-informative prior:**
  - Prior 1: Uniform prior distribution
  \[ p(\beta) \propto \text{constant} \]

- **Informative prior:**
  - Prior 2: Gaussian prior distribution
  \[ p(\beta \mid \mu, \sigma^2) \propto \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(\beta - \mu)^2}{2\sigma^2} \right) \]
Logic Analysis Methods

Classical Logistic Regression

- Social/Economic Variables
- Biomass Availability
- Transportation-related Variables
- Land Cover Variables

Model 1: Simple Regression
Model 2: Variable Selection → Regression
Model 3: Variable Transformation → Regression
Model 4: Variable Transformation → Variable Selection → Regression

Model Selection by BIC

Bayesian Logistic Regression

Build Priors for Parameters
Derive Posteriors for Parameters
 Obtain Estimates

Model Assessment

<table>
<thead>
<tr>
<th>Predicted Value</th>
<th>ZCTAs of Primary Mills with small capacity (0)</th>
<th>ZCTAs of existing biorefineries (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Value</td>
<td>True Prediction</td>
<td>False Prediction</td>
</tr>
<tr>
<td>ZCTAs of Primary Mills with small capacity (0)</td>
<td>True Prediction</td>
<td>False Prediction</td>
</tr>
<tr>
<td>ZCTAs of existing biorefineries (1)</td>
<td>False Prediction</td>
<td>True Prediction</td>
</tr>
</tbody>
</table>
Classical Logistic Regression

Logistic Regression Models

<table>
<thead>
<tr>
<th>Model Name</th>
<th>BIC Value</th>
<th>Misclassification Rate (Training - 80%)</th>
<th>Misclassification Rate (Test - 20%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 Simple Regression</td>
<td>189.94</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>Model 2 Variable Selection → Regression</td>
<td>189.94</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>Model 3 Variable Transformation → Regression</td>
<td>178.22</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>Model 4 Variable Transformation → Variable Selection → Regression</td>
<td>184.83</td>
<td>0.18</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Model Assessment

<table>
<thead>
<tr>
<th>Predicted Value</th>
<th>Actual Value</th>
<th>0 (Training + Test)</th>
<th>1 (Training + Test)</th>
<th>Total (Training + Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>143 + 32</td>
<td>3 + 4</td>
<td>146 + 36 = 182</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>28 + 7</td>
<td>19 + 5</td>
<td>47 + 12 = 59</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
<td>171 + 39</td>
<td>22 + 9</td>
<td>241</td>
</tr>
</tbody>
</table>

Best

Low
Classification Trees with Local Bayesian Logistic Regression

**Objective**

- To improve the predictive performance

**Model**

- Integrate classification trees (CTs) with local logistic regression
- Apply Bayesian inference instead of using maximum likelihood methods for the estimates in local logistic regressions

**Strategy**

- First to use CTs partitioning the whole data set in several manageable subsets
- Then to fit a local logistic regression in each subset to get a probability estimate of siting biomass-using facilities
- The number of subsets is no more than 10
- The number of observations in each subset greater than 15
Classification Trees

**Results – The Hybrid Model**

**Biomass-Using Facilities**
- ZCTAs of Existing Mills: 59
- ZCTAs of Primary Mills with Small Capacity: 182

**Subset 1:**
- **Road Density >= 3.6 km/km²**
  - ZCTAs of Existing Mills: 16
  - ZCTAs of Primary Mills with Small Capacity: 2

**Subset 2:**
- **Median Family Income >= 40,845 dollars**
  - ZCTAs of Existing Mills: 35
  - ZCTAs of Primary Mills with Small Capacity: 53

**Subset 3:**
- **Median Family Income < 40,845 dollars**
  - ZCTAs of Existing Mills: 3
  - ZCTAs of Primary Mills with Small Capacity: 31

**Subset 4:**
- **Total Cost_80mile >= 21 mil dollars**
  - ZCTAs of Existing Mills: 38
  - ZCTAs of Primary Mills with Small Capacity: 84

**Road Density < 3.6 km/km²**
- ZCTAs of Existing Mills: 43
- ZCTAs of Primary Mills with Small Capacity: 180

**Subset 4:**
- **Total Cost_80mile < 21 mil dollars**
  - ZCTAs of Existing Mills: 5
  - ZCTAs of Primary Mills with Small Capacity: 96
### Results – The Hybrid Model

#### Significant Variables

<table>
<thead>
<tr>
<th>Significant Variables</th>
<th>Whole Set (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subset 1</td>
</tr>
<tr>
<td>Road Density</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Total Cost _ 80 mile</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Median Family Income</td>
<td>0.0009</td>
</tr>
<tr>
<td>Average Cost _ 80 mile</td>
<td>0.0425</td>
</tr>
<tr>
<td>Total Quantity _80 mile</td>
<td>0.0165</td>
</tr>
<tr>
<td>Number of Water Ports</td>
<td>0.0660</td>
</tr>
</tbody>
</table>

#### Inference Methods

Correct Classification Rate = \( \frac{"0" \cdot (\text{True Prediction}) + "1" \cdot (\text{True Prediction})}{\text{Total}} \)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Likelihood Estimation</td>
<td>2 + 14/18 = 99%</td>
<td>48 + 14/88 = 70%</td>
<td>30 + 2/34 = 94%</td>
<td>96 + 0/101 = 95%</td>
<td>206/241 = 85%</td>
</tr>
<tr>
<td>Bayesian Inference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform Prior</td>
<td>2 + 15/18 = 94%</td>
<td>49 + 16/88 = 74%</td>
<td>30 + 2/34 = 94%</td>
<td>96 + 0/101 = 95%</td>
<td>210/241 = 87%</td>
</tr>
<tr>
<td>Gaussian Prior</td>
<td>2 + 14/18 = 94%</td>
<td>48 + 14/88 = 70%</td>
<td>30 + 2/34 = 94%</td>
<td>96 + 0/101 = 95%</td>
<td>206/241 = 85%</td>
</tr>
</tbody>
</table>

**Predicted Value**

<table>
<thead>
<tr>
<th>Actual Value</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Value</td>
<td>175</td>
<td>177</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>26</td>
<td>33</td>
<td>59</td>
</tr>
</tbody>
</table>

**Total**

<table>
<thead>
<tr>
<th></th>
<th>203</th>
<th>38</th>
<th>241</th>
</tr>
</thead>
</table>
Possible Locations

Results – Prediction

Possible Biorefineries Locations

“Classification Tree”
### Possible Locations

**Result of the hybrid model**

- **Existing Biorefineries (59)**
- **Possible Biorefineries Locations (Prob > 0.8)**
- **Possible ZCTAs (4,756)**

**Result of landscape suitability Index**

- **Forest Suitability - High (987)**
- **Forest Suitability - Possibly Suitable**
Conclusions and Future Research

- The hybrid model integrating CTs with local Bayesian logistic regression improved the predictive power.
- The hybrid model quantified the significant factors that mostly influence the location of biomass-using facilities.
- The application of GIS improved data quality and consistency.

- Assess the interrelationship of the landscape’s physical features and human geography paired with economically viable and ecologically sustainable locations.
- Develop spatially-explicit models for the suitability of biorefineries locations.
- Validate the developed models.


Acknowledgements

• USDA Forest Service Southern Research Station – Asheville, NC

• USDA Southeastern Sun Grant Center, Knoxville, TN

• University of Tennessee, Agricultural Experiment Station, Knoxville, TN

• University of Tennessee, College of Business, Knoxville, TN
THANK YOU