Estimating Control Total Acres for Desired Geographies Using Cropland Data Layer (CDL)

Mingyue Hu, DaeGyu Jang, Cindy Yu, Zhu Zhengyuan

Center for Survey and Statistical Methodology

Iowa State University

August 5, 2024

Background

- The goal of the Conservation Effects Assessment Project (CEAP): examine the environmental impact of the USDA's conservation program.
- The goal of this sub-project: obtain an accurate estimate of the total area of rangeland and pastureland for any arbitrary domain (such as Ecoregions, counties) using two data sources: NRI photo interpretation data and CDL data.

National Resources Inventory (NRI)

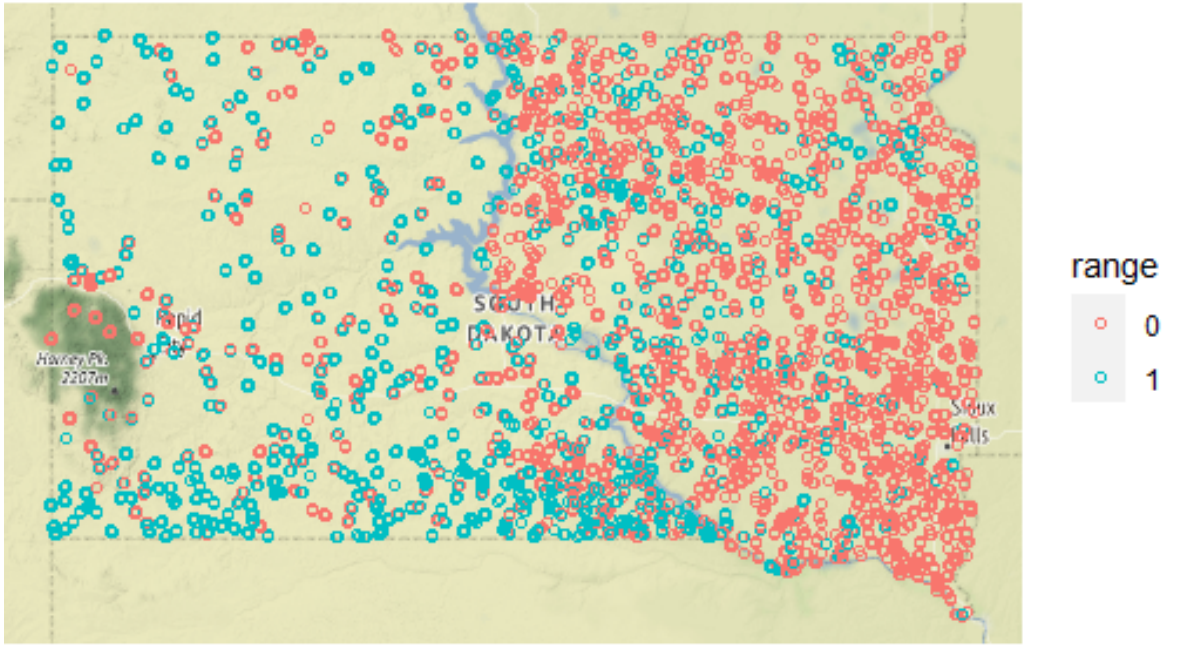


Figure: NRI sample points in the state of South Dakota

Design-based NRI Estimator for Each State

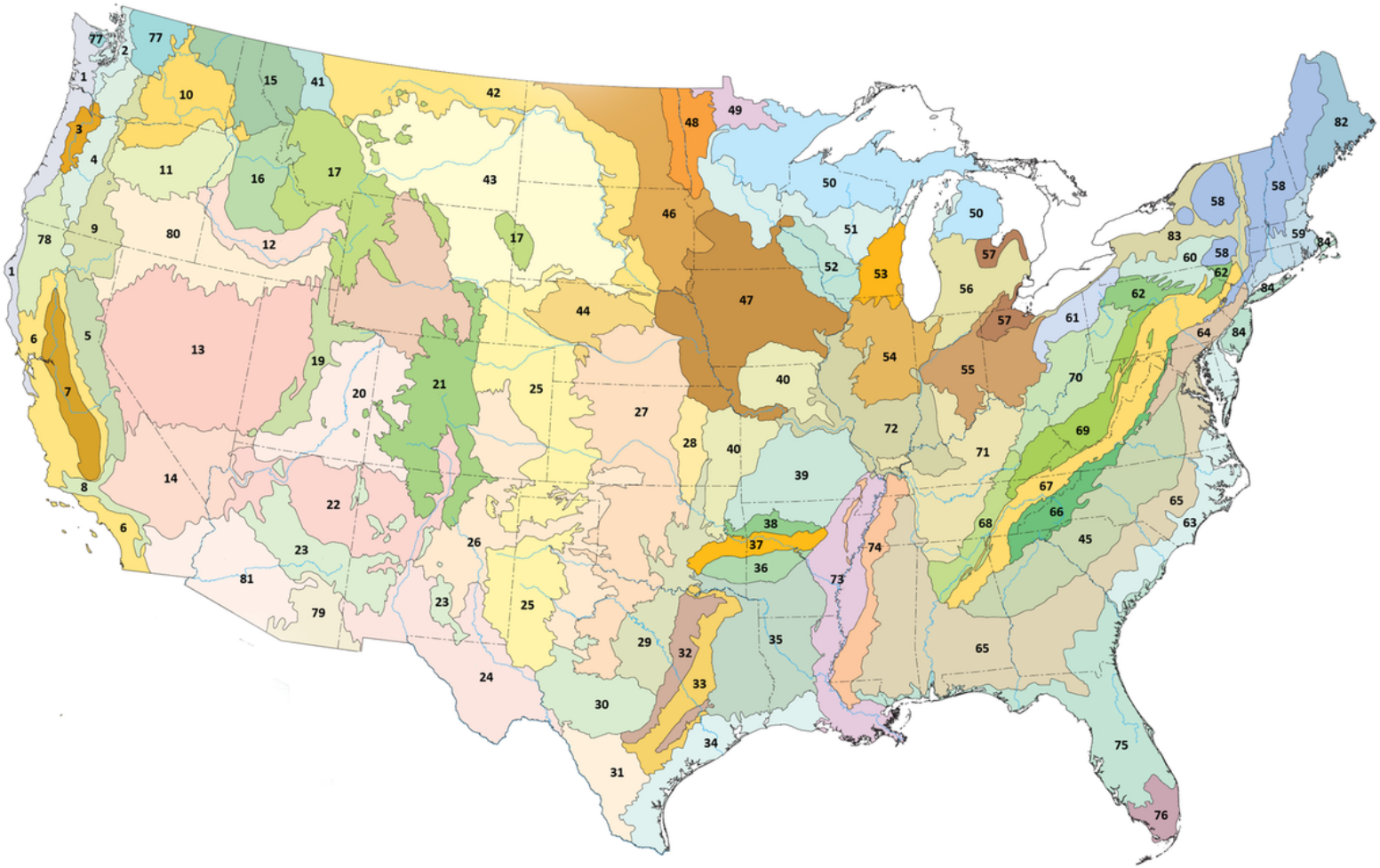
- $\hat{T}_s = \sum_i \tilde{W}_{i,s} Y_{i,s}$: The NRI estimator of total acres of (nonfederal) rangeland/pastureland in state s, where

$$Y_{i,s} = \begin{cases} 1, & \text{if } i\text{th NRI sample point in state } s \text{ is range/pasture,} \\ 0, & \text{otherwise,} \end{cases}$$

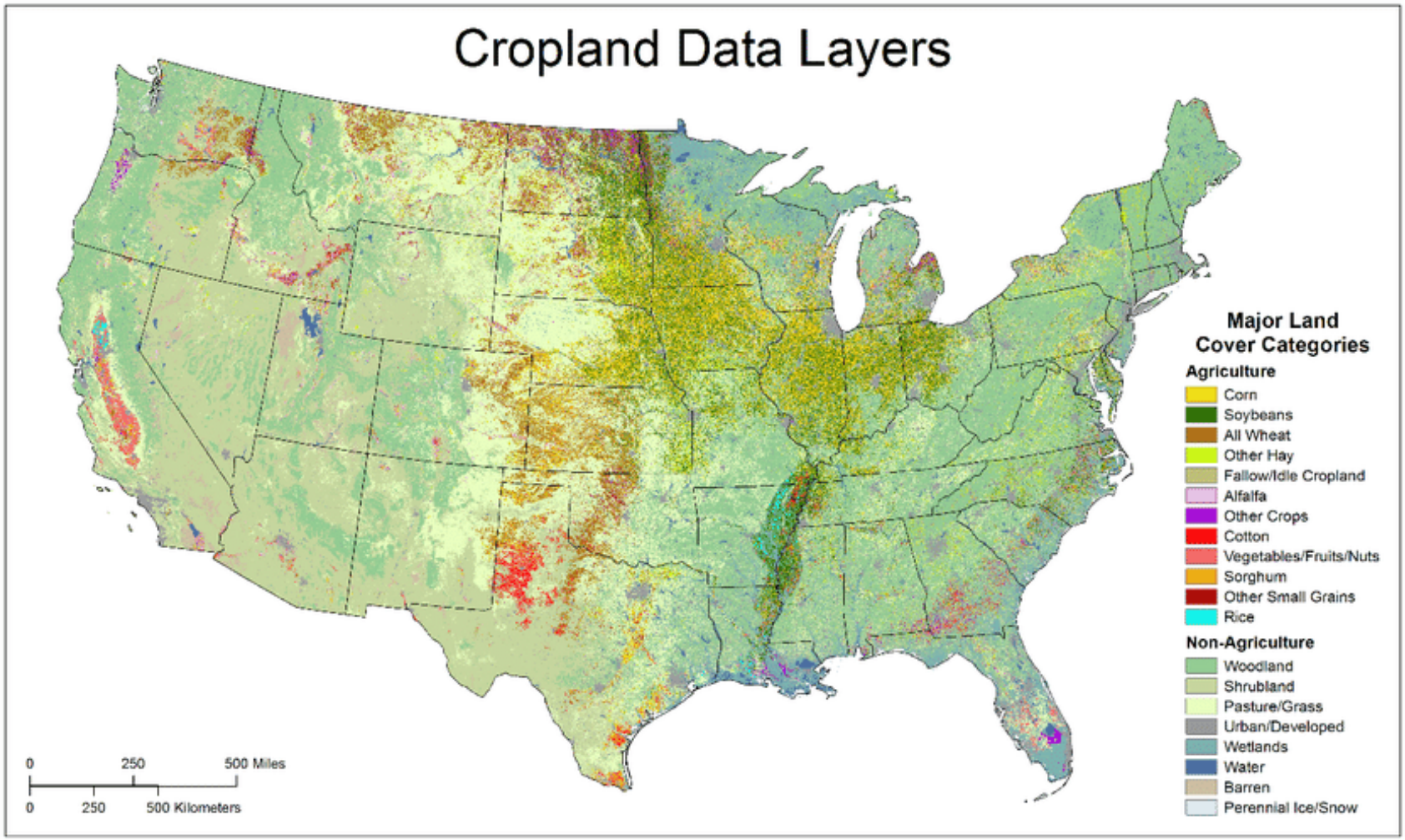
$\tilde{W}_{i,s} =$ the weight for ith NRI sample point in state s.

- The weights $\tilde{W}_{i,s}$ of NRI data are developed to represent non-federal land of states and counties only.

New Desired Regions: Ecoregion LV3



Cropland Data Layer (CDL)



CDL data together with NRI samples

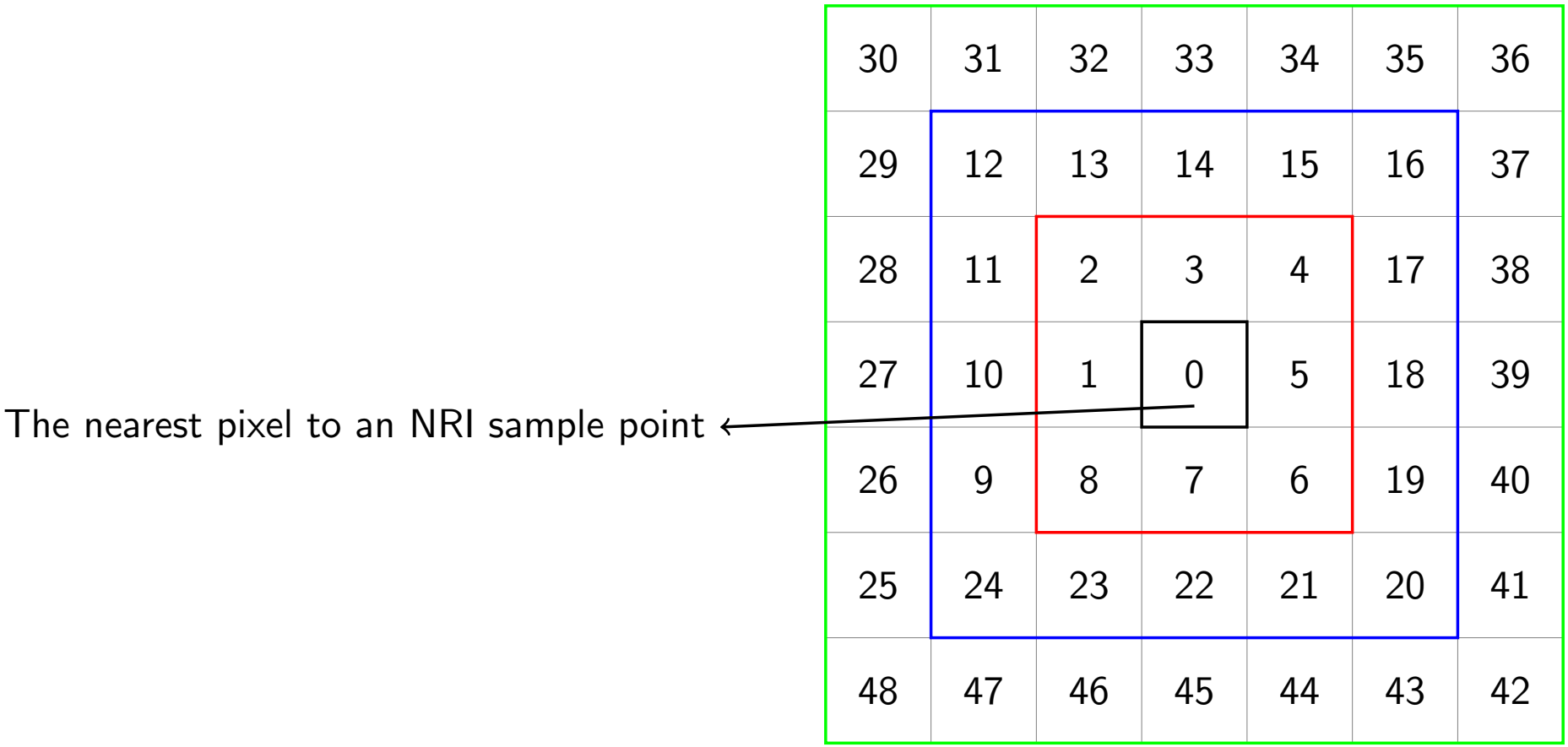


Figure: The sequence numbers of observations from 7×7 pixels centered on the nearest pixel to an NRI sample point.

Variables

$$
\begin{aligned}
Y_i &= \begin{cases} 1, & \text{if ith NRI sample point is rangeland/pastureland,} \\ 0, & \text{otherwise} \end{cases} \\
X_{ij} &= \frac{(\text{the number of pixels of category } j \text{ in the square area.})}{(\text{the number of pixels in the square area.})}, \\
&\quad \text{for } j = 1, \ldots, p,
\end{aligned}
$$

Statistical Model

- Logistic Regression Model:

$$P(Y_i = 1|\boldsymbol{X}_i;\boldsymbol{\beta}) = \pi(\boldsymbol{X}_i;\boldsymbol{\beta}) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij}\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij}\right)}, \quad (1)$$

  where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$.

- The log-likelihood function of $\boldsymbol{\beta}$:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\{y_i \log(P(Y_i = 1|\boldsymbol{X}_i;\boldsymbol{\beta})) + (1-y_i)\log(P(Y_i = 0|\boldsymbol{X}_i;\boldsymbol{\beta}))\} \\ &= \sum_{i=1}^{n}\left\{y_i \boldsymbol{X}_i^T\boldsymbol{\beta} - \log(1+\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))\right\}. \end{aligned} \quad (2)$$

LASSO

- The LASSO method computes the estimator of $\boldsymbol{\beta}$ by minimizing the following objective function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} [-l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1] \tag{3}$$

  where $l(\boldsymbol{\beta})$ is the log-likelihood function of $\boldsymbol{\beta}$ in (2), $\|\cdot\|_1$ is l_1 norm, and $\lambda > 0$ is the regularization parameter.
- The smaller λ is, the smoother the function is.

Model-Based Imputation Estimator

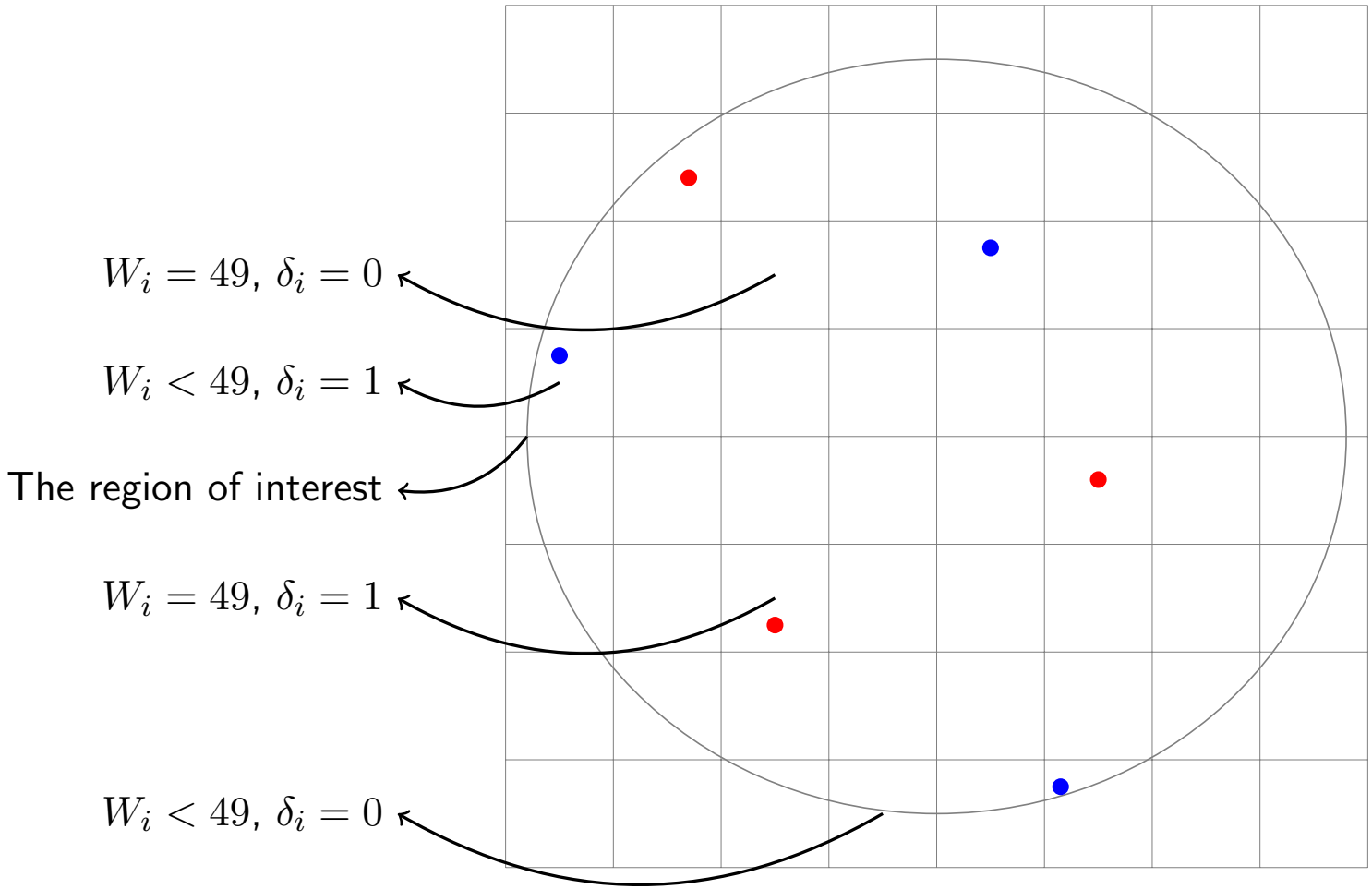


Figure: The revisited example of a grid. The ellipse represents the region of interest.

Parameter of Interest

$$T = \sum_{i=1}^{N} mW_i(1 - \eta_i)Y_i, \tag{4}$$

where $m = 0.2223946$ is the multiplier used to convert the number of pixels into acres, W_i is the size of the pixels for the overlap of the ith section and the target region, and $\eta_i = 1$ if the federal government owns the ith section, 0 otherwise.

$$Y_i = \begin{cases} 1, & \text{if ith NRI sample point is rangeland/pastureland,} \\ 0, & \text{otherwise.} \end{cases}$$

Model-based Estimator

- $\hat{T}_m$, the model-based estimator of T is as follows:

$$\begin{aligned}\hat{T}_m &= \sum_{i=1}^{N} mW_i(1-\eta_i)\left\{\delta_i Y_i + (1-\delta_i)E\left\{Y_i|\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}\right\}\right\},\\ &= \sum_{i=1}^{N} mW_i(1-\eta_i)\left\{\delta_i Y_i + (1-\delta_i)P\left(Y_i=1|\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}\right)\right\}, \qquad (5)\end{aligned}$$

where

$$\delta_i = \begin{cases} 1 & \text{if section } i \text{ contains NRI point,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\eta_i = \begin{cases} 1 & \text{if section } i \text{ is federal land} \\ 0 & \text{otherwise.} \end{cases}$$

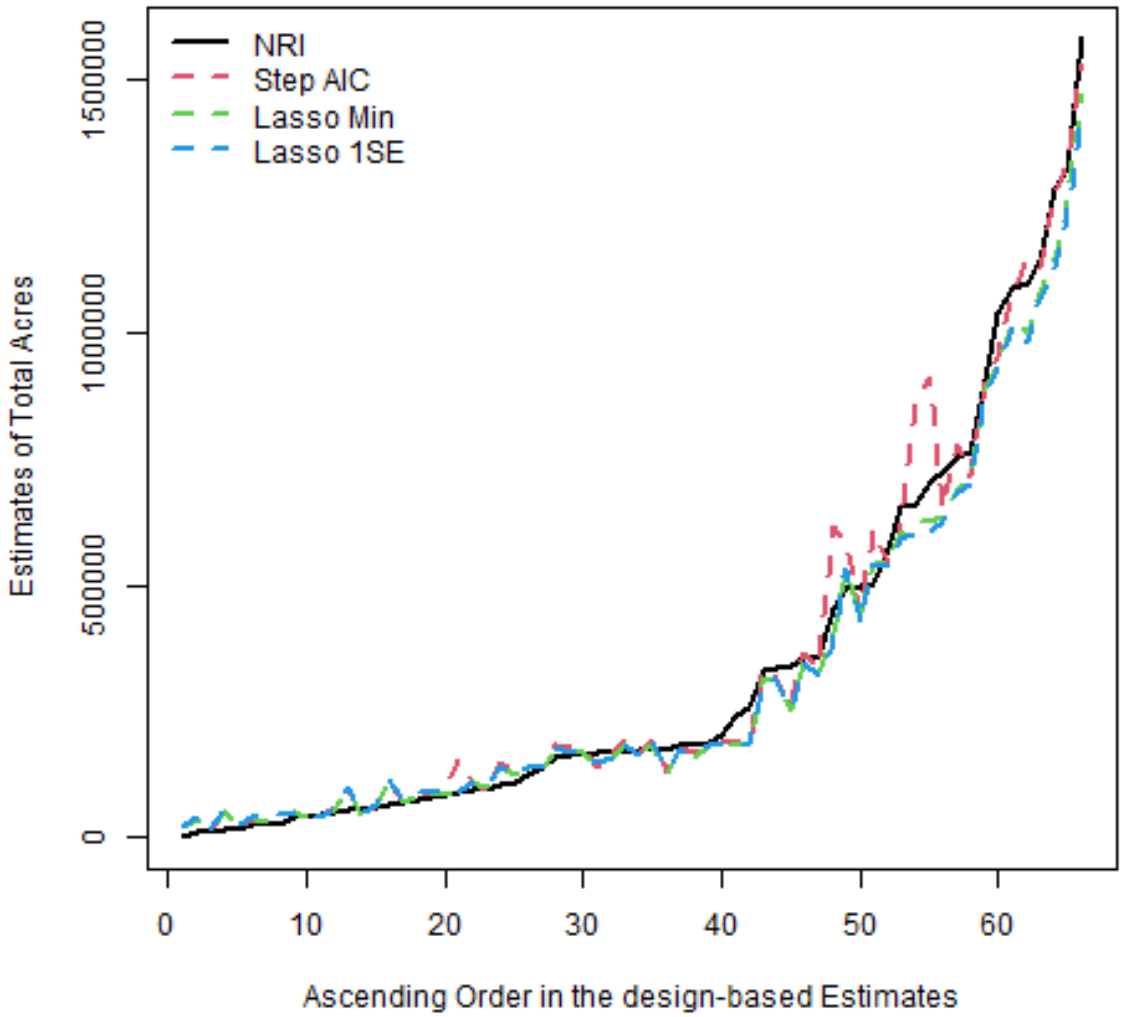
Mean AUCs - Rangeland

Table: The mean AUCs of 16 cases

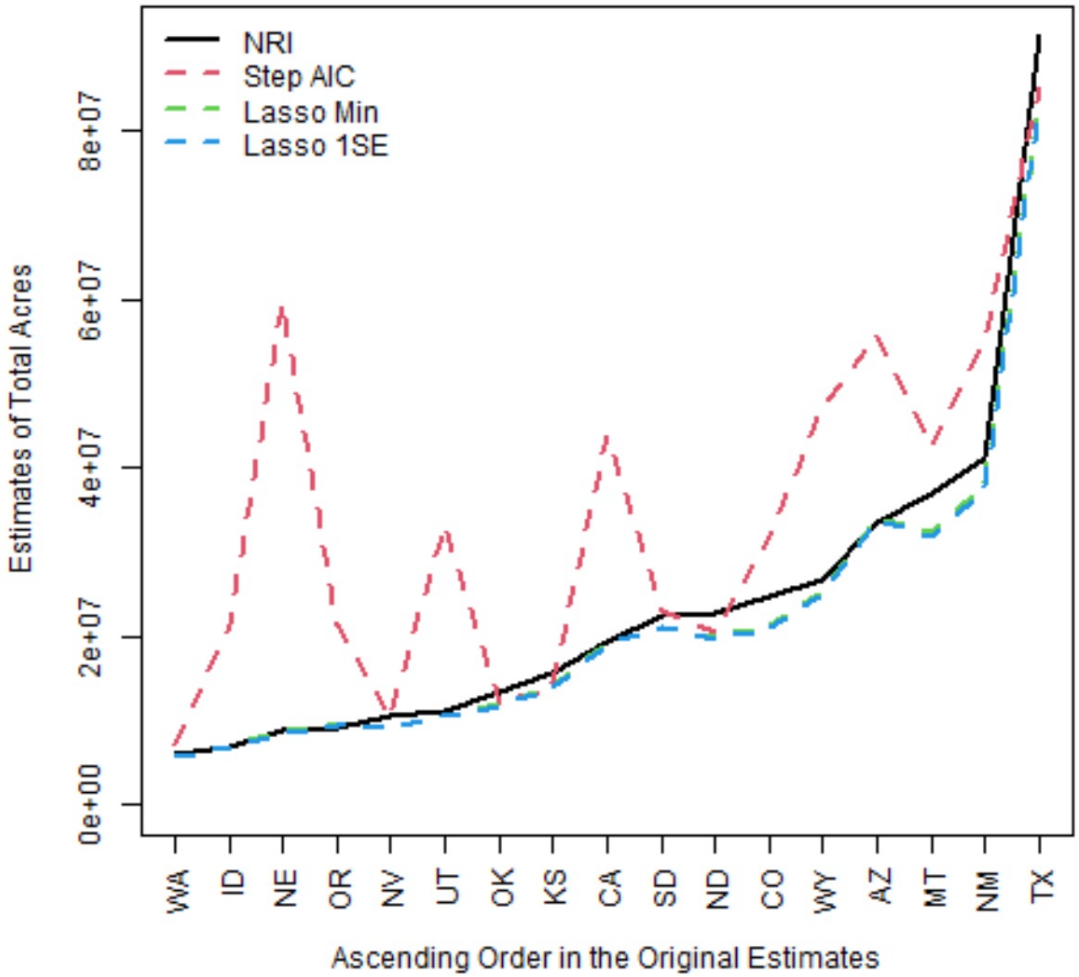
Size	Full Model	step AIC	LASSO 1SE	LASSO min
1	0.9071	0.9071	0.9076	0.9088
3	0.8023	0.9404	0.9409	0.9416
5	0.8477	0.9441	0.9439	0.9453
7	0.7812	0.9457	0.9470	0.9478

Note: The 1st column Size presents the size of squares, from which the data is generated. Each value in the 2nd to 4th columns presents the mean AUC. Full Model, the logistic regression model with all main effect terms; step AIC, the best model in terms of AIC; LASSO 1SE, the most parsimonious model with-in 1SE of the best LASSO Model; LASSO min, the best LASSO model.

County-level Estimators for Counties in South Dakota



State-level Estimators for 17 States

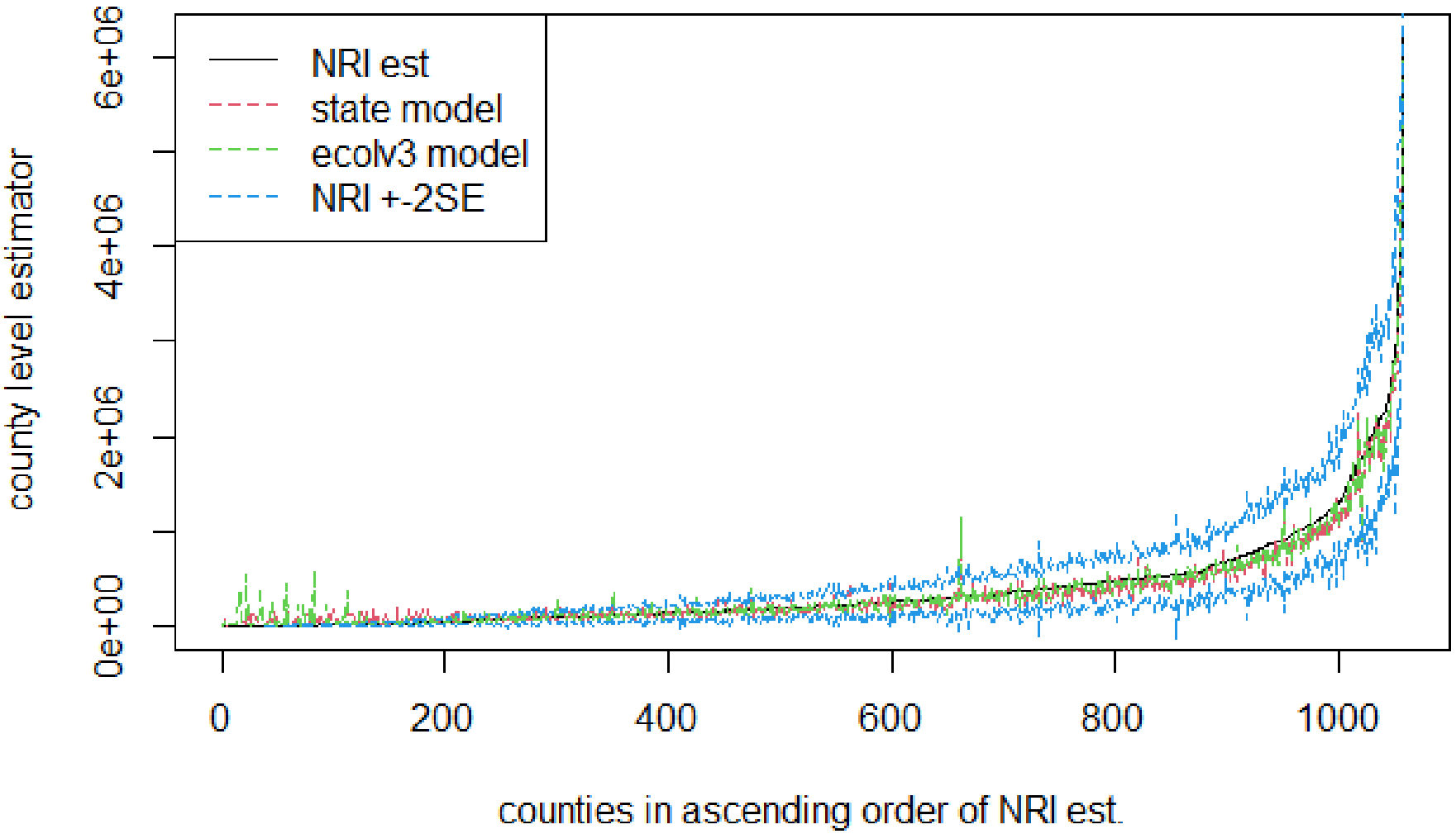


Combined Ecoregions - Rangeland

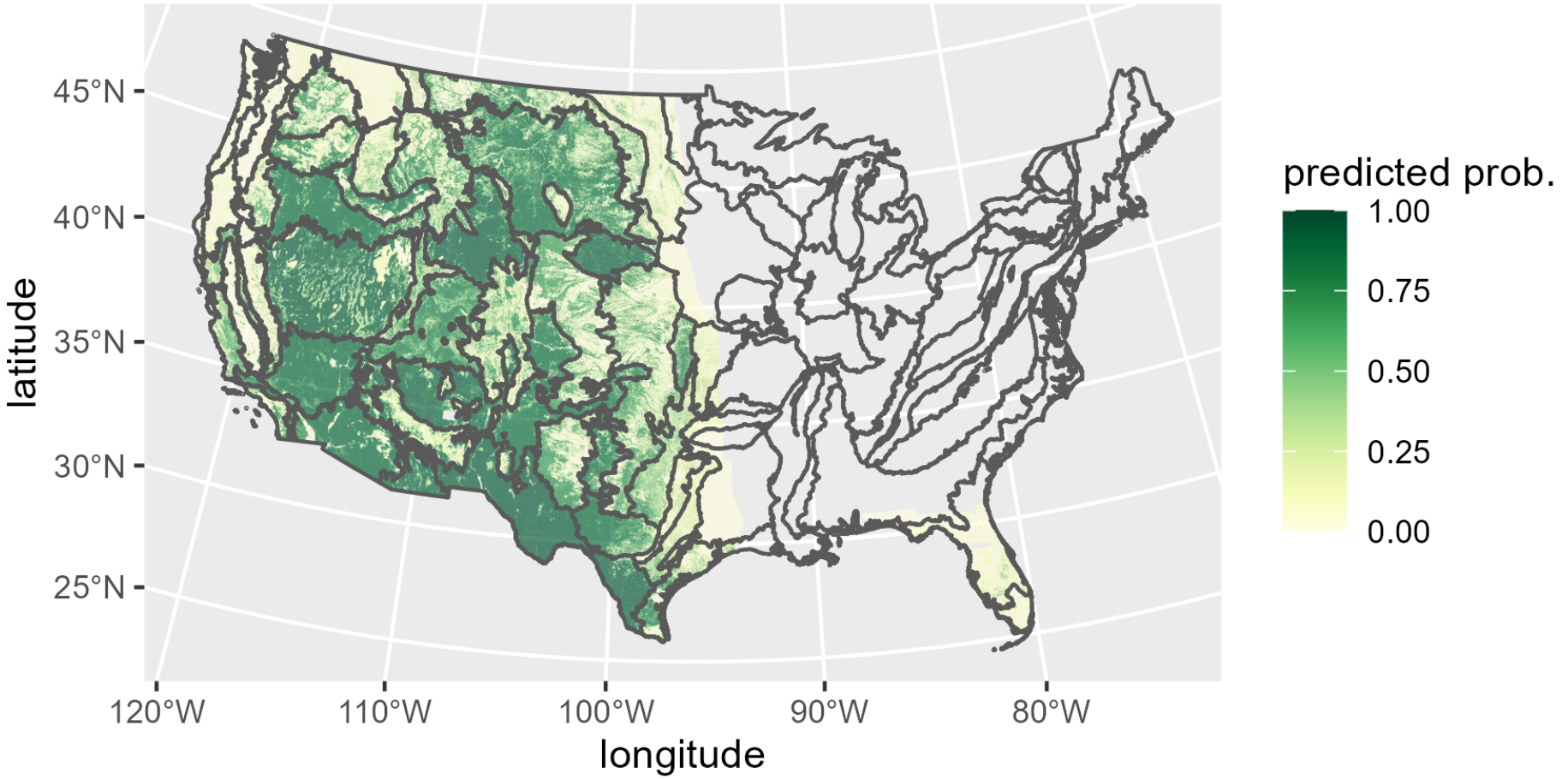
- Some ecoregions have too small numbers of NRI points to fit the model.
- Those ecoregions are combined with one of the adjacent ecoregions that produce state-level estimators closest to the NRI estimators.

Trouble Eco lv3	n_range	n_nonrange	Combined Eco lv3	n_range	n_nonrange
2	1	778	1	54	1169
38	0	156	37	37	572
39	10	2286	40	218	4013

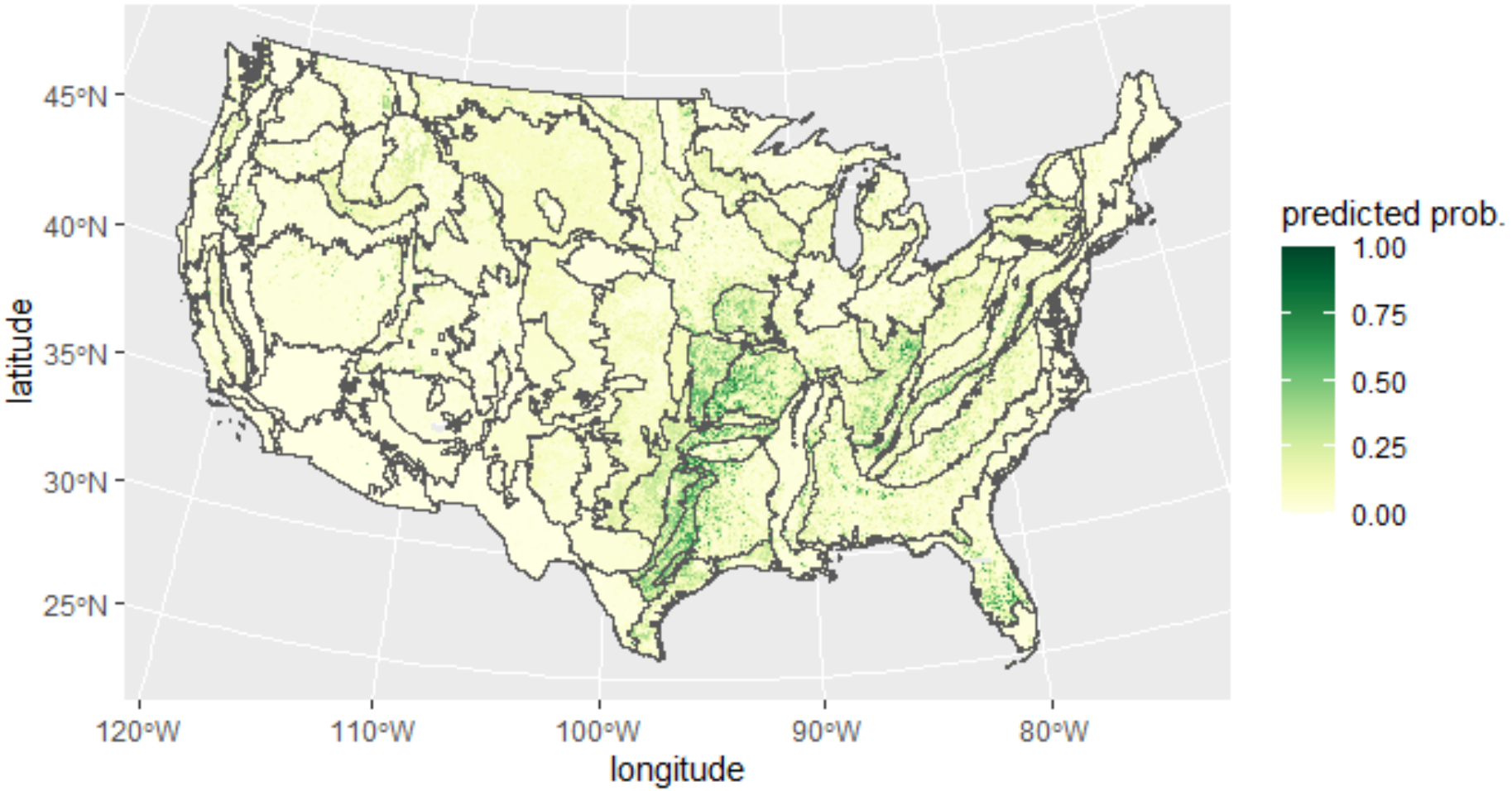
County-level Estimators for Ecoregion LV3



Ecoregion boundaries and predicted results - Rangeland



Ecoregion boundaries and predicted results - Pastureland



Contact Information

- Mingyue Hu: yue9615@iastate.edu