How Do We Assess the Performance of Our Small Area Estimators?

Jerzy Wieczorek[1], Kelly McConville[2], Grayson White[3], Tracey Frescino[4], Gretchen Moisen[4]

[1] Colby College, [2] Harvard University,
[3] Michigan State University, [4] USDA Forest Service, RMRS

Forest Inventory and Analysis Science Stakeholder Meeting **Nov 17, 2022**

When estimating the mean of some forest attribute of interest, we have so many choices!

When estimating the mean of some forest attribute of interest, we have so many choices!

- Design based?

$$\hat{\mu} = \frac{1}{n} \sum_{i \in s} y_i$$

When estimating the mean of some forest attribute of interest, we have so many choices!

- Design based?
- Model-assisted?

$$\hat{\mu} = \frac{1}{n} \sum_{i \in s} y_i$$

$$\hat{\mu} = \frac{1}{N} \sum_{i \in U} \hat{y}_i + \frac{1}{n} \sum_{i \in s} (y_i - \hat{y}_i)$$

When estimating the mean of some forest attribute of interest, we have so many choices!

- Design based?
- Model-assisted?
- Model-based?

$$\hat{\mu} = \frac{1}{n} \sum_{i \in s} y_i$$

$$\hat{\mu} = \frac{1}{N} \sum_{i \in U} \hat{y}_i + \frac{1}{n} \sum_{i \in s} (y_i - \hat{y}_i)$$

$$\hat{\mu} = X' \beta + \nu + \epsilon$$

When estimating the mean of some forest attribute of interest, we have so many choices!

- Design based?
- Model-assisted?
- Model-based?
  - EBLUP?
  - Hierarchical Bayes?
  - Spatial?

$$\hat{\mu} = \frac{1}{n} \sum_{i \in s} y_i$$

$$\hat{\mu} = \frac{1}{N} \sum_{i \in U} \hat{y}_i + \frac{1}{n} \sum_{i \in s} (y_i - \hat{y}_i)$$

Model form?

$$\hat{\mu} = X' \beta + \nu + \epsilon$$

Priors?

Fitting method?

When estimating the mean of some forest attribute of interest, we have so many choices!

- Design based?
- Model-assisted?
- Model-based?
  - EBLUP?
  - Hierarchical Bayes?
  - Spatial?

$$\hat{\mu} = \frac{1}{n} \sum_{i \in s} y_i$$

$$\hat{\mu} = \frac{1}{N} \sum_{i \in U} \hat{y}_i + \frac{1}{n} \sum_{i \in s} (y_i - \hat{y}_i)$$

$$\hat{\mu} = X' \beta + \nu + \epsilon$$

Model form?

Priors?

Fitting method?

How do we choose?

In order to address these choices, we have designed a unique simulation study using k nearest neighbors (kNN) and hot deck imputation.

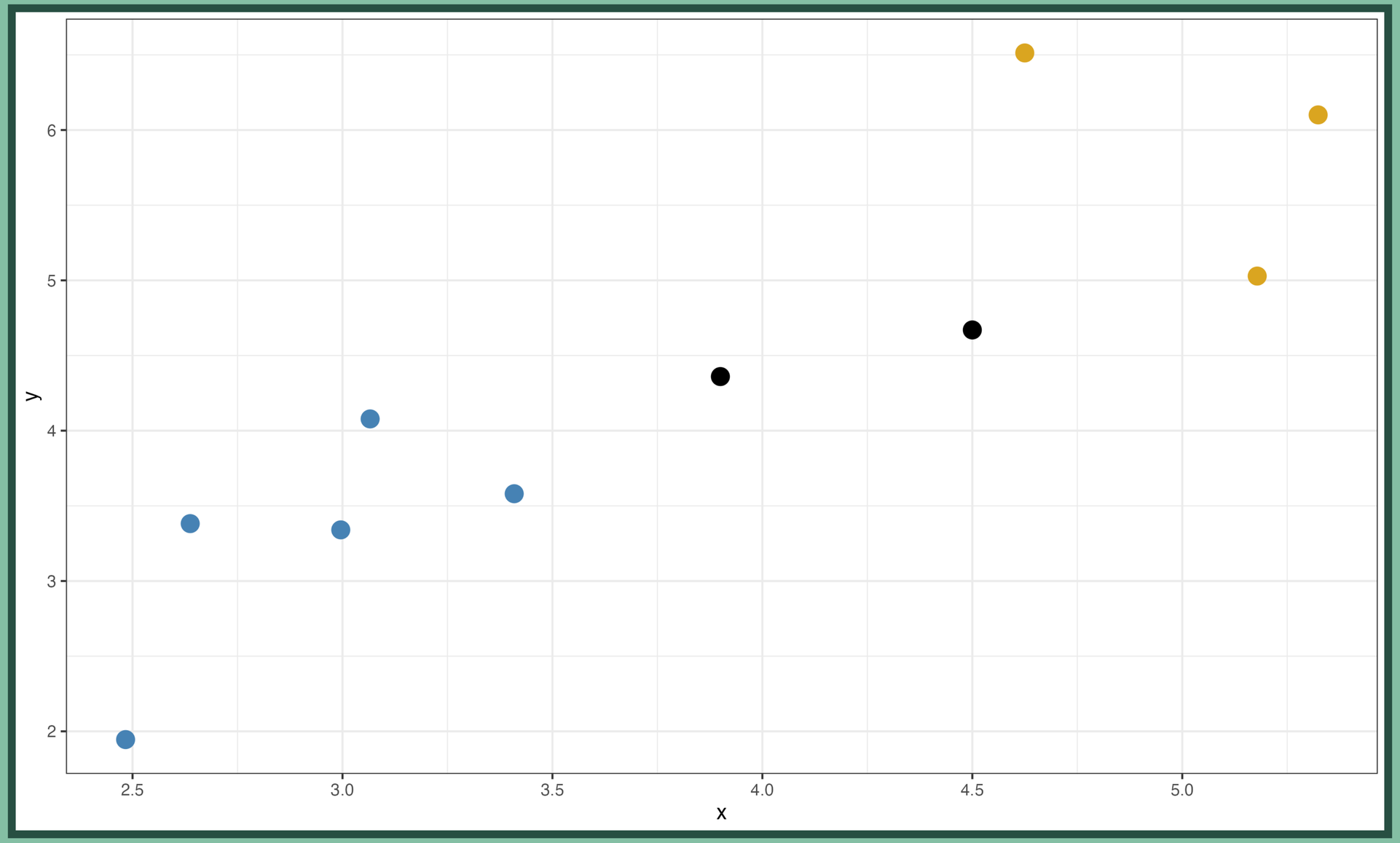
kNN

A non-parametric algorithm for finding the k "closest" observations to a given observation.

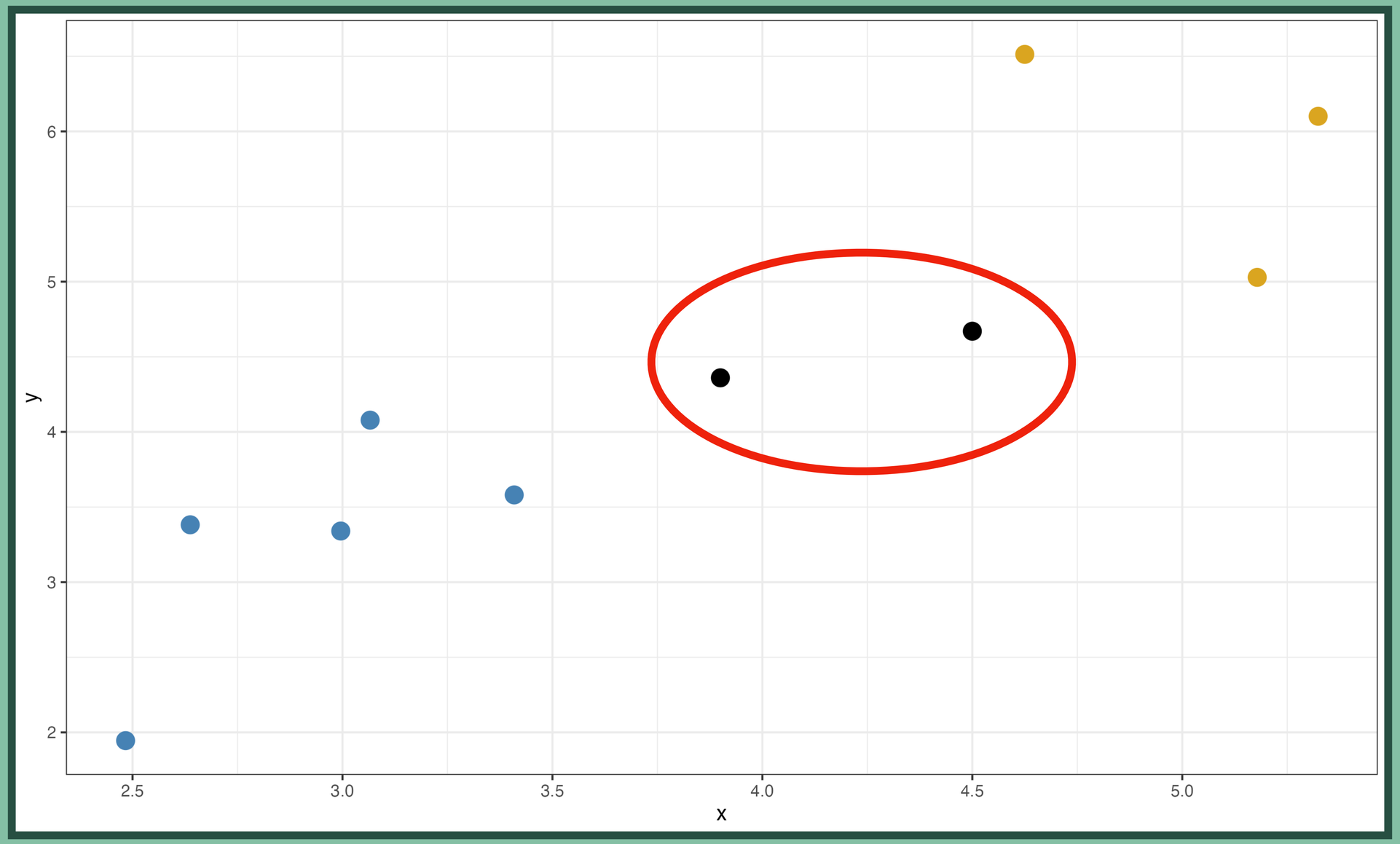
Hot deck imputation

A method for filling in missing data, where each new value is picked at random from a set of possible new values.

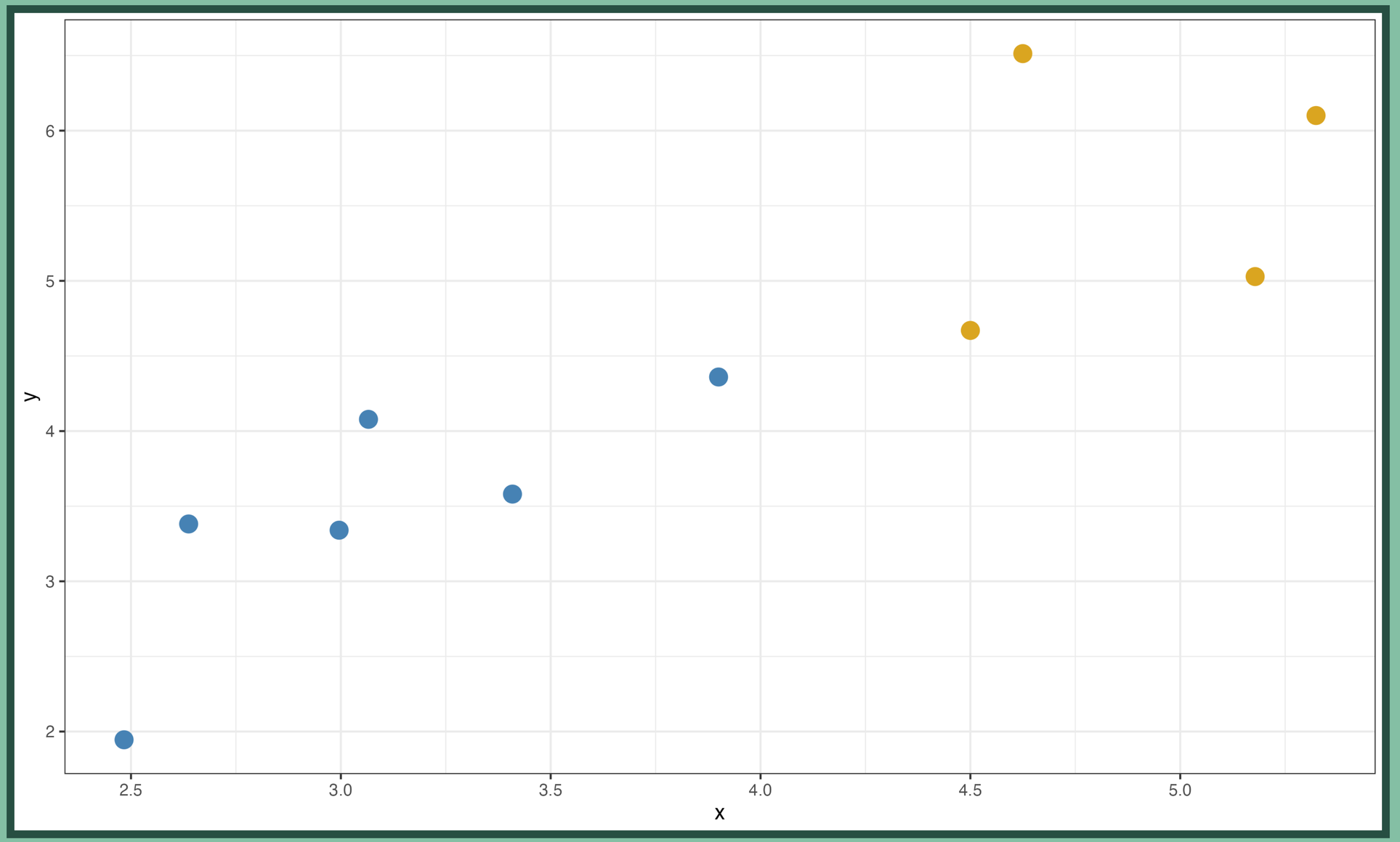
A brief overview of kNN



A brief overview of kNN



A brief overview of kNN

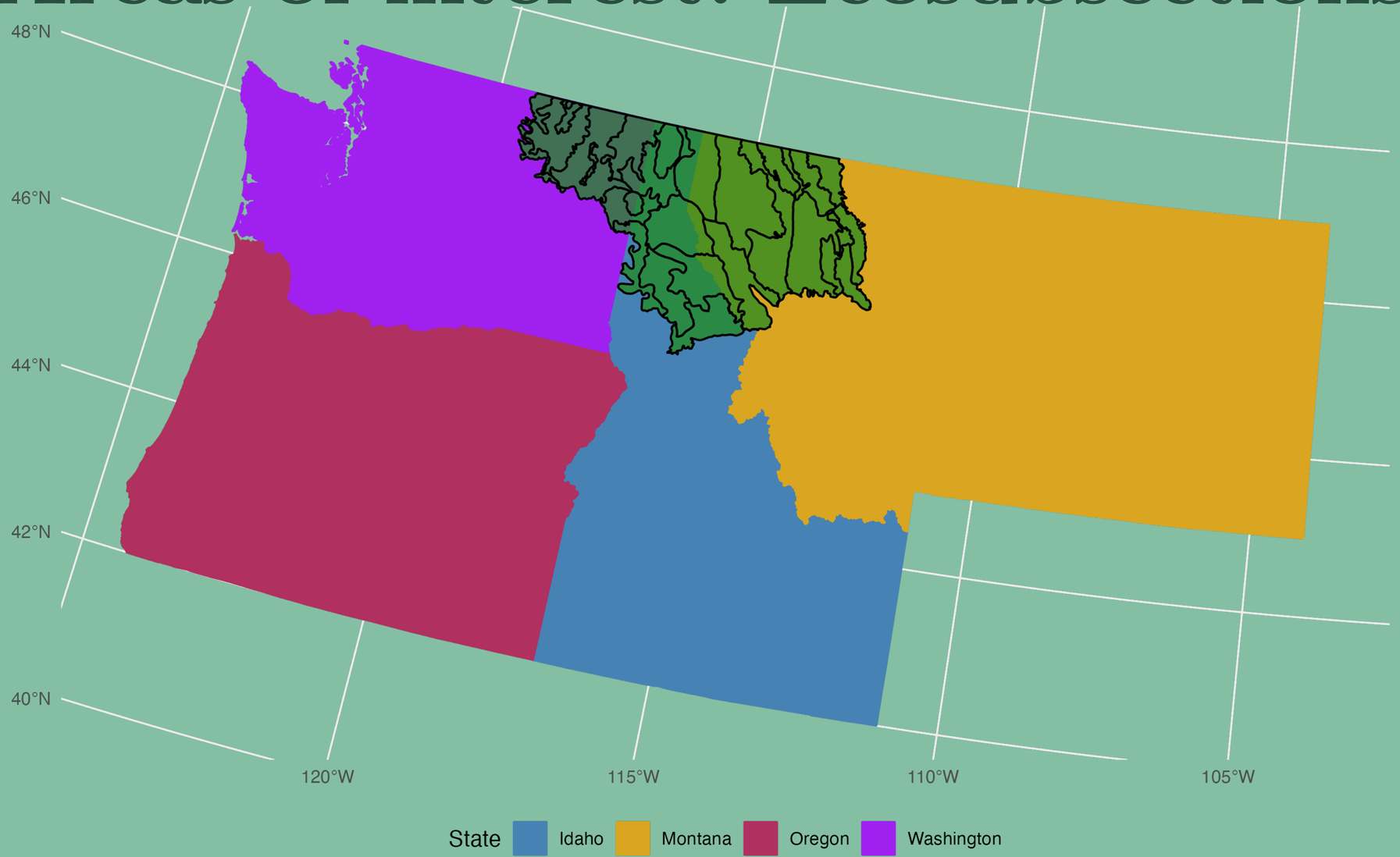


Simulation Design

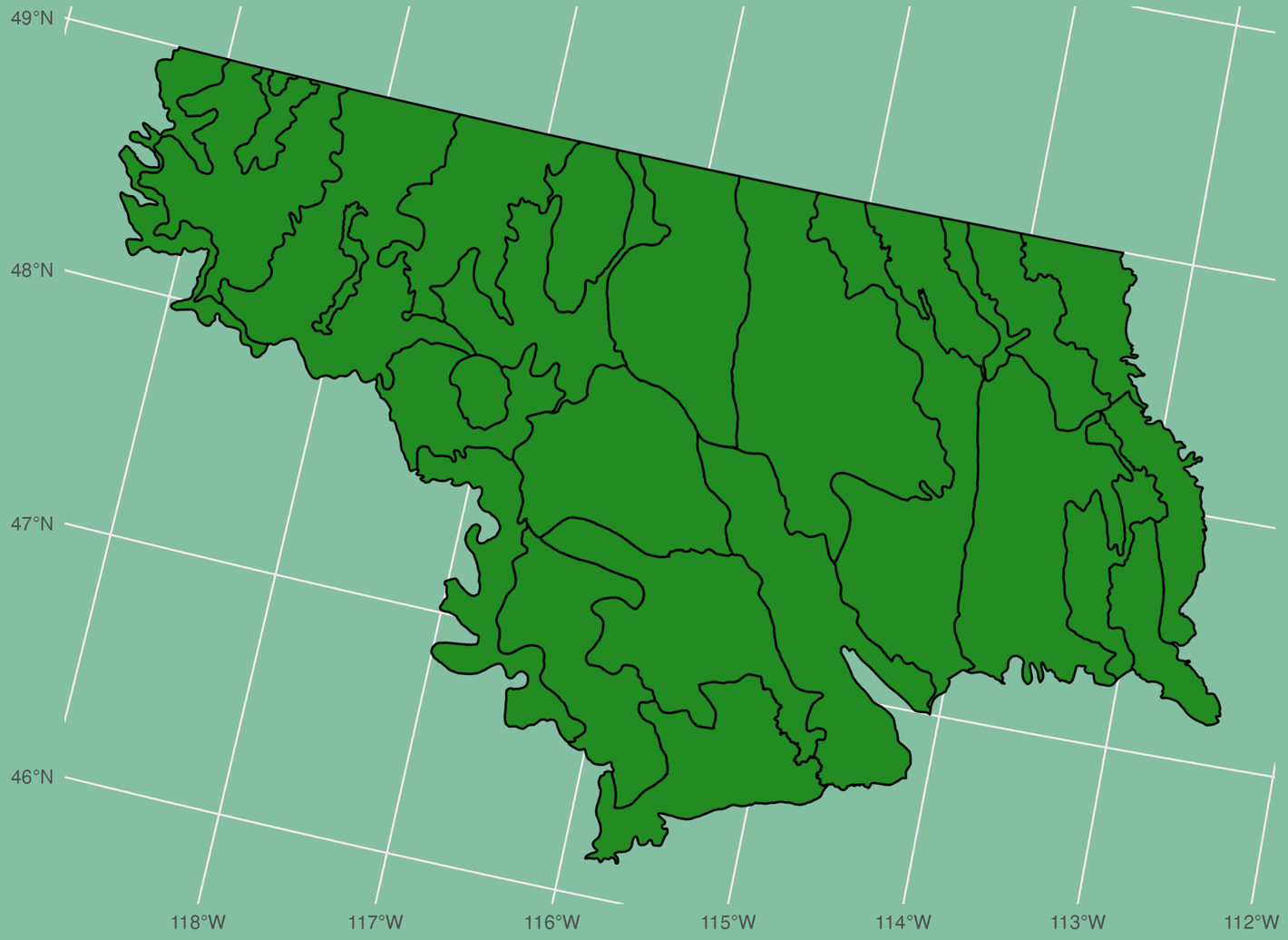
Orcas Island, Washington

Lansing, Michigan

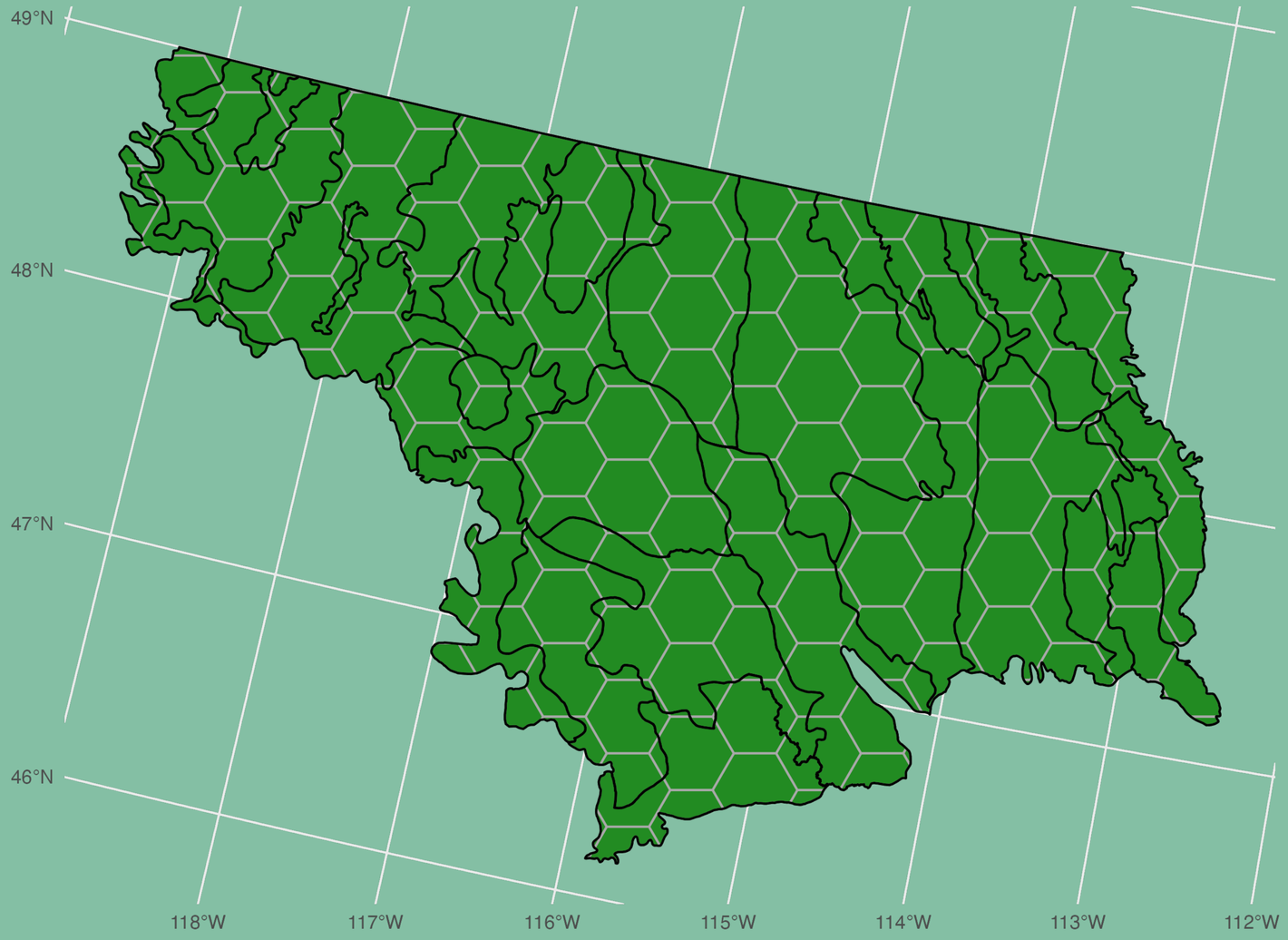
Study region: Ecoprovince M333, Areas of interest: Ecosubsections



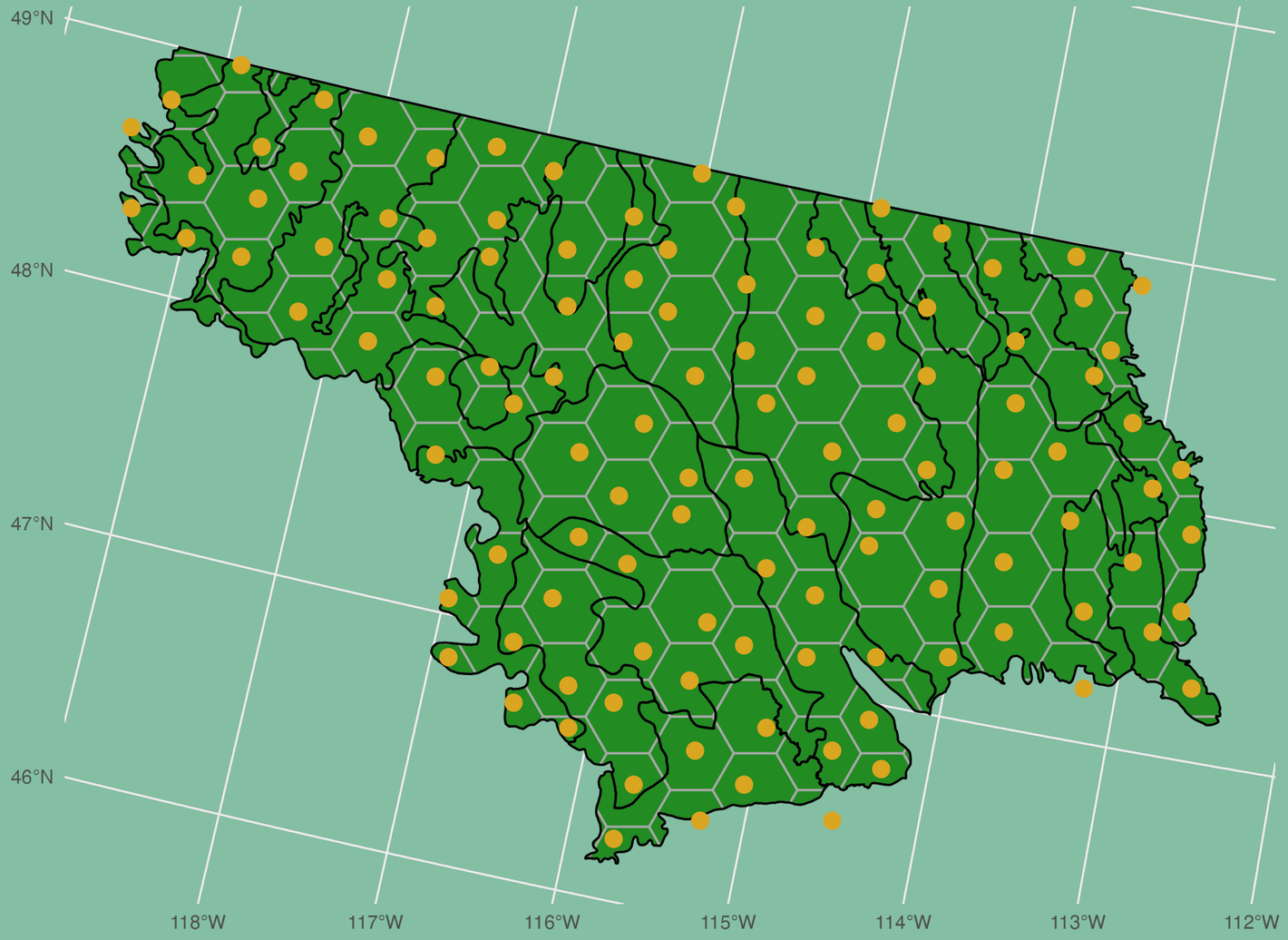
Study region: Ecoprovince M333, Areas of interest: Ecosubsections



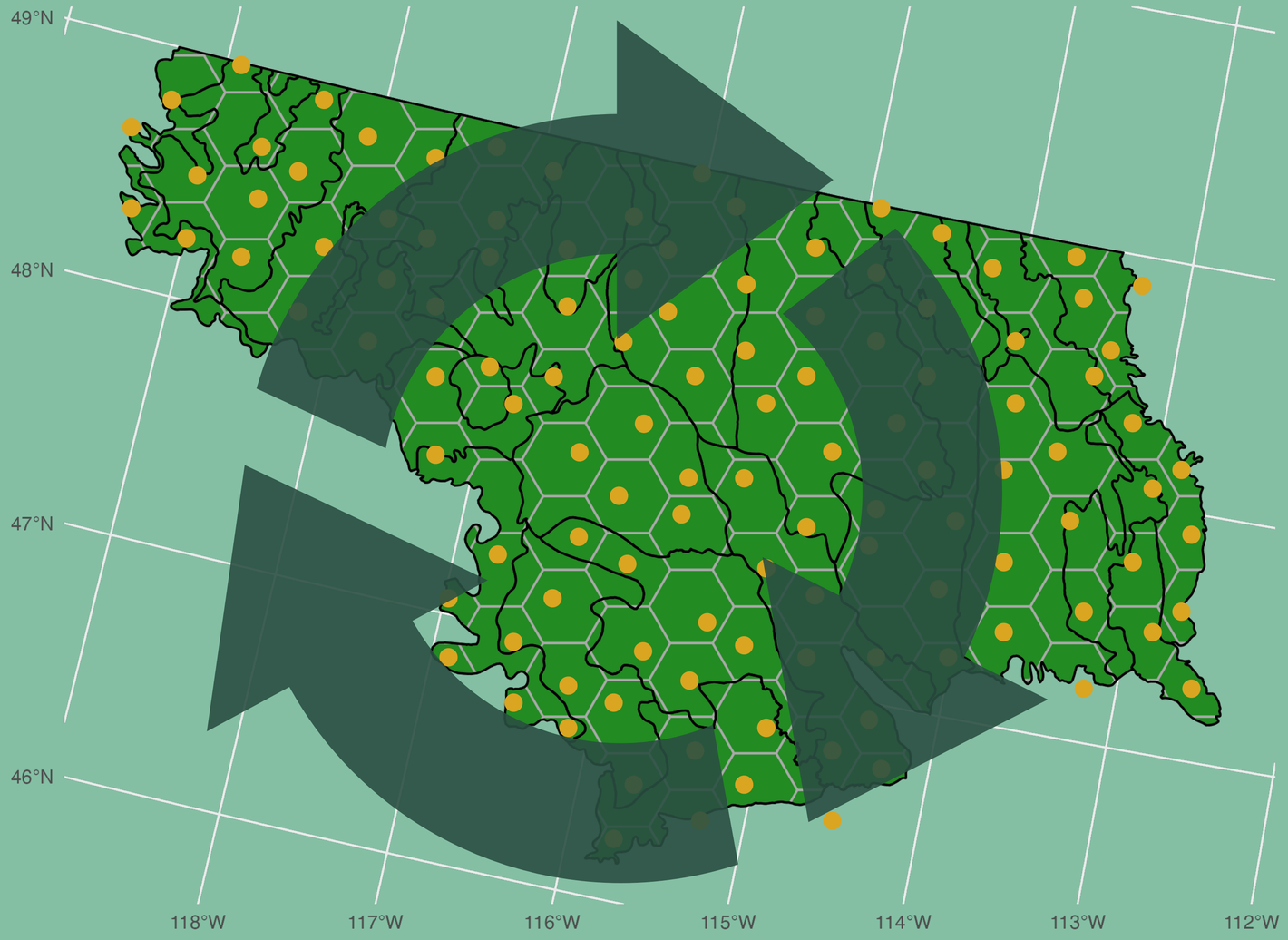
Creating Artificial Samples: Overlay FIA's Hexagonal Grid



Creating Artificial Samples: Randomly Select 90m Pixel Within Each Hex



Creating Artificial Samples: Repeat Pixel Selection Many Times



kNN Hot Deck Imputation: Step-by-step

Recipient row

tcc	tri	def	tnt
63	18.4	198	1

Step 1: select a recipient row of data

kNN Hot Deck Imputation: Step-by-step

Recipient row

tcc	tri	def	tnt
63	18.4	198	1

Potential donors

tcc	tri	def	tnt	BA
62	15.4	147	1	185
59	19.8	242	1	148
66	21.6	221	1	136

Step 1: select a recipient row of data

Step 2: find k nearest neighbors ("best matches") in donor dataset

kNN Hot Deck Imputation: Step-by-step

Recipient row

tcc	tri	def	tnt
63	18.4	198	1

Potential donors

tcc	tri	def	tnt	BA
62	15.4	147	1	185
59	19.8	242	1	148
66	21.6	221	1	136

Step 1: select a recipient row of data

Step 2: find k nearest neighbors ("best matches") in donor dataset

Step 3: choose one of the k best matches at random

kNN Hot Deck Imputation: Step-by-step

Recipient row

tcc	tri	def	tnt	BA
63	18.4	198	1	148

Potential donors

tcc	tri	def	tnt	BA
62	15.4	147	1	185
59	19.8	242	1	148
66	21.6	221	1	136

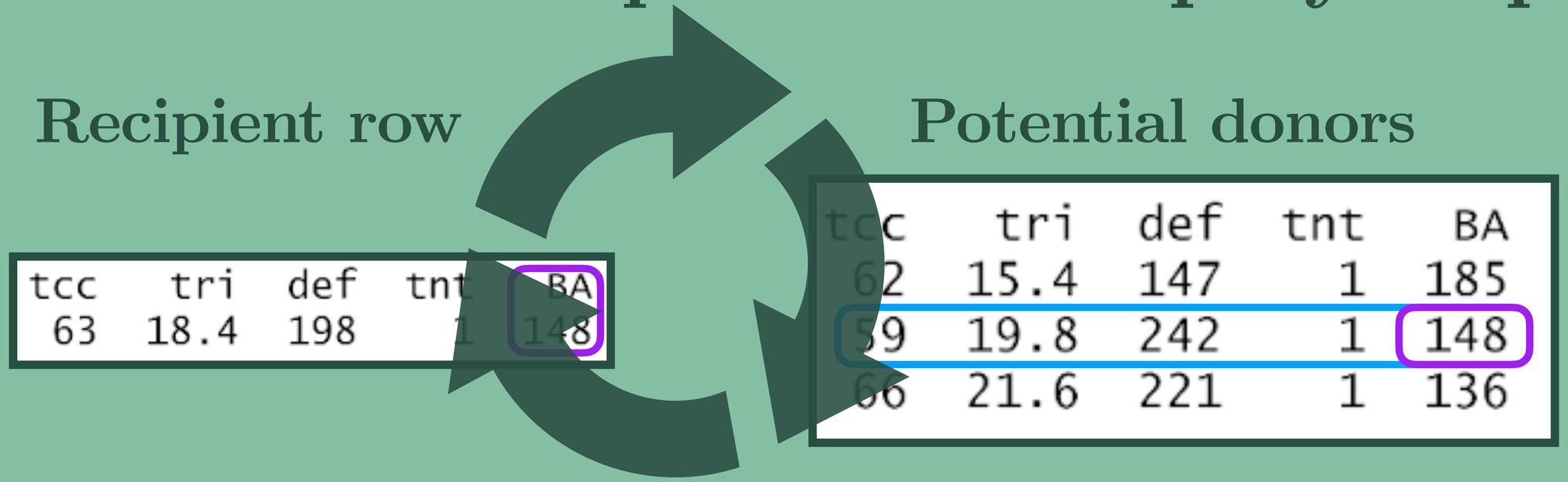
Step 1: select a recipient row of data

Step 2: find k nearest neighbors ("best matches") in donor dataset

Step 3: choose one of the k best matches at random

Step 4: Impute the donor's response value to the recipient row

kNN Hot Deck Imputation: Step-by-step



Step 1: select a recipient row of data

Step 2: find k nearest neighbors ("best matches") in donor dataset

Step 3: choose one of the k best matches at random

Step 4: Impute the donor's response value to the recipient row

Step 5: Repeat many times to generate simulation datasets

Why not a simpler alternative to kNN Hot Deck Imputation?

How else could we have generated our response variable values?

(1) Through fitting a regression model, or

(2) By treating an auxiliary variable as the response

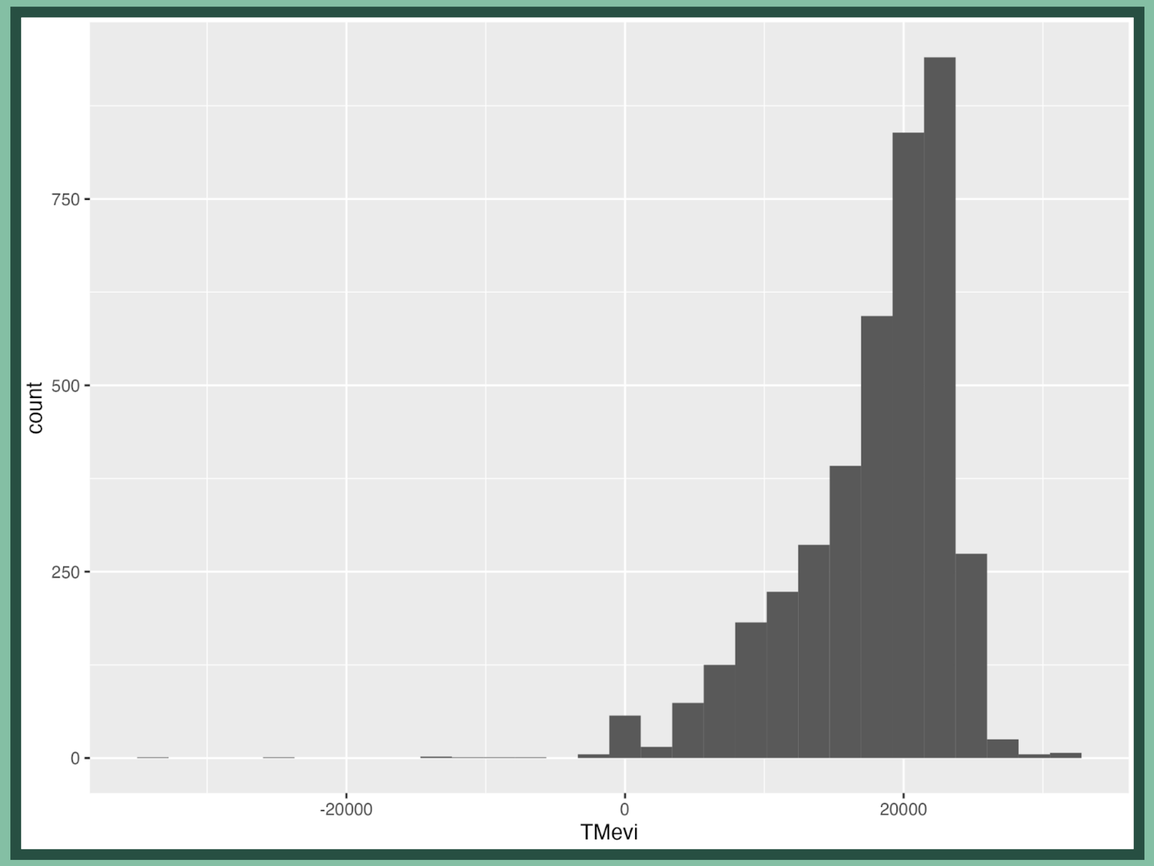
Both (1) and (2) have issues!

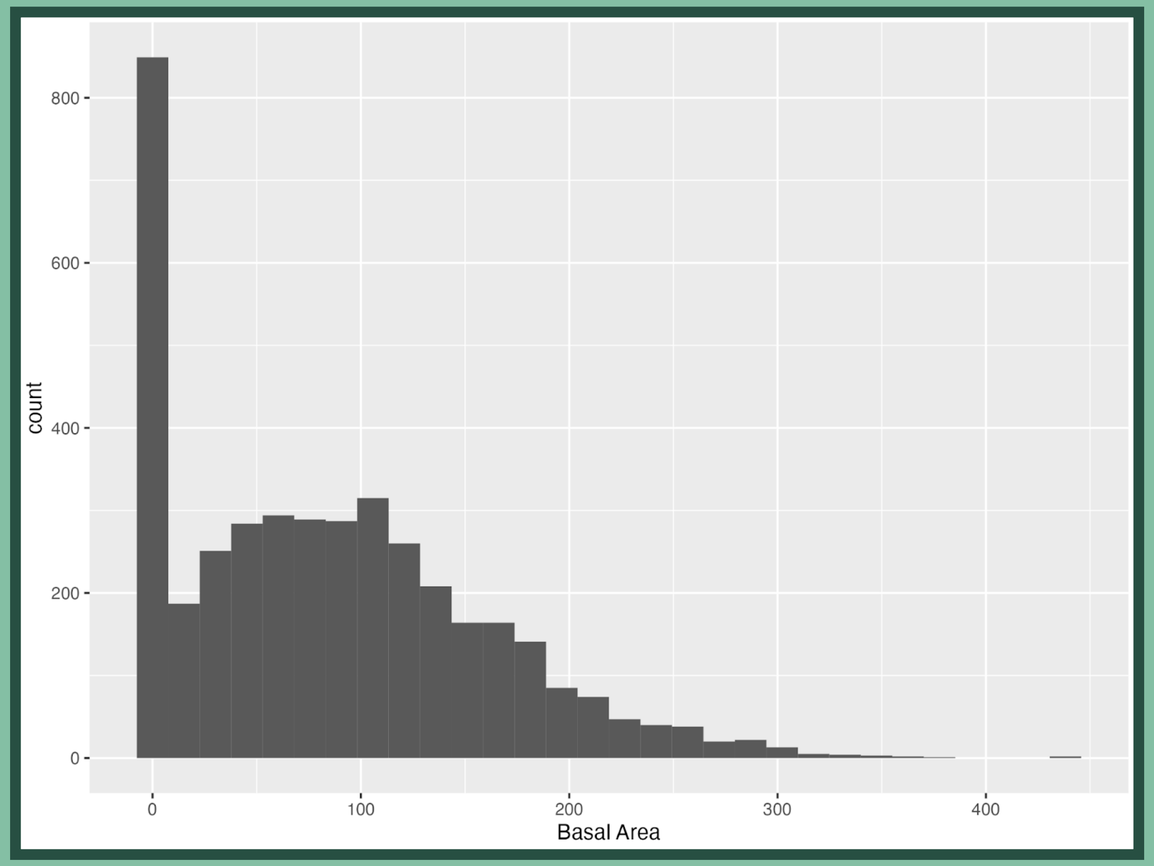
(1) Regression as Response:

Issue: If we fit a regression model and generated response values from it, our simulation study would be over-optimistic when evaluating estimators built on similar regression models.

(2) Auxiliary as Response

Issue: Our auxiliary variables are not "shaped" liked our response variables.





When estimating the mean of some forest attribute of interest, we have so many choices!

$$\hat{\mu} = \frac{1}{n}\sum_{i \in s} y_i$$

$$\hat{\mu} = \frac{1}{N}\sum_{i \in U} y_i + \frac{1}{n}\sum (y_i - \hat{y}_i)$$

Model form?

$$\hat{\mu} = X'\beta + \nu + \epsilon$$

Priors?

Fitting method?

How do we choose?

Thank you!

Get in touch:

Email me: whitegra@msu.edu

Website: www.graysonwhite.com

Slides/other work available: www.github.com/graysonwhite