Assessing small area estimates via bootstrap-weighted k-Nearest-Neighbor artificial populations

Grayson W. White[1], Jerzy A. Wieczorek[2], Zachariah W. Cody[2], Emily X. Tan[2], Jacqueline O. Chistolini[2], Kelly S. McConville[3], Tracey S. Frescino[4], Gretchen G. Moisen[4]

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

Joint Statistical Meetings 2024 Portland, Oregon, USA August 7, 2024

Outline

1. Context & motivation

2. Methodology

3. Data application & results

4. Future work

Motivating application: Small area estimation in forest inventory

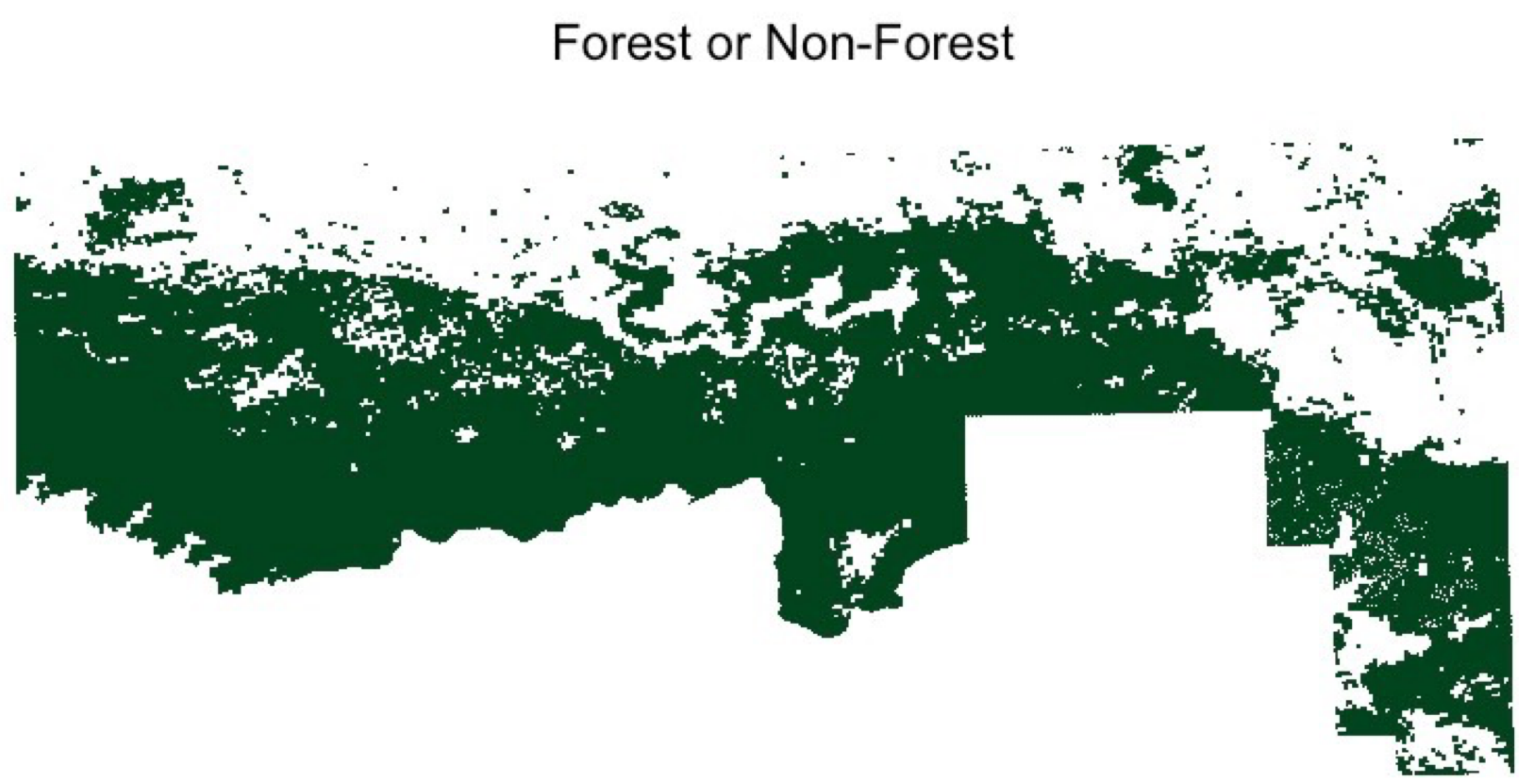
Motivating application: Small area estimation in forest inventory

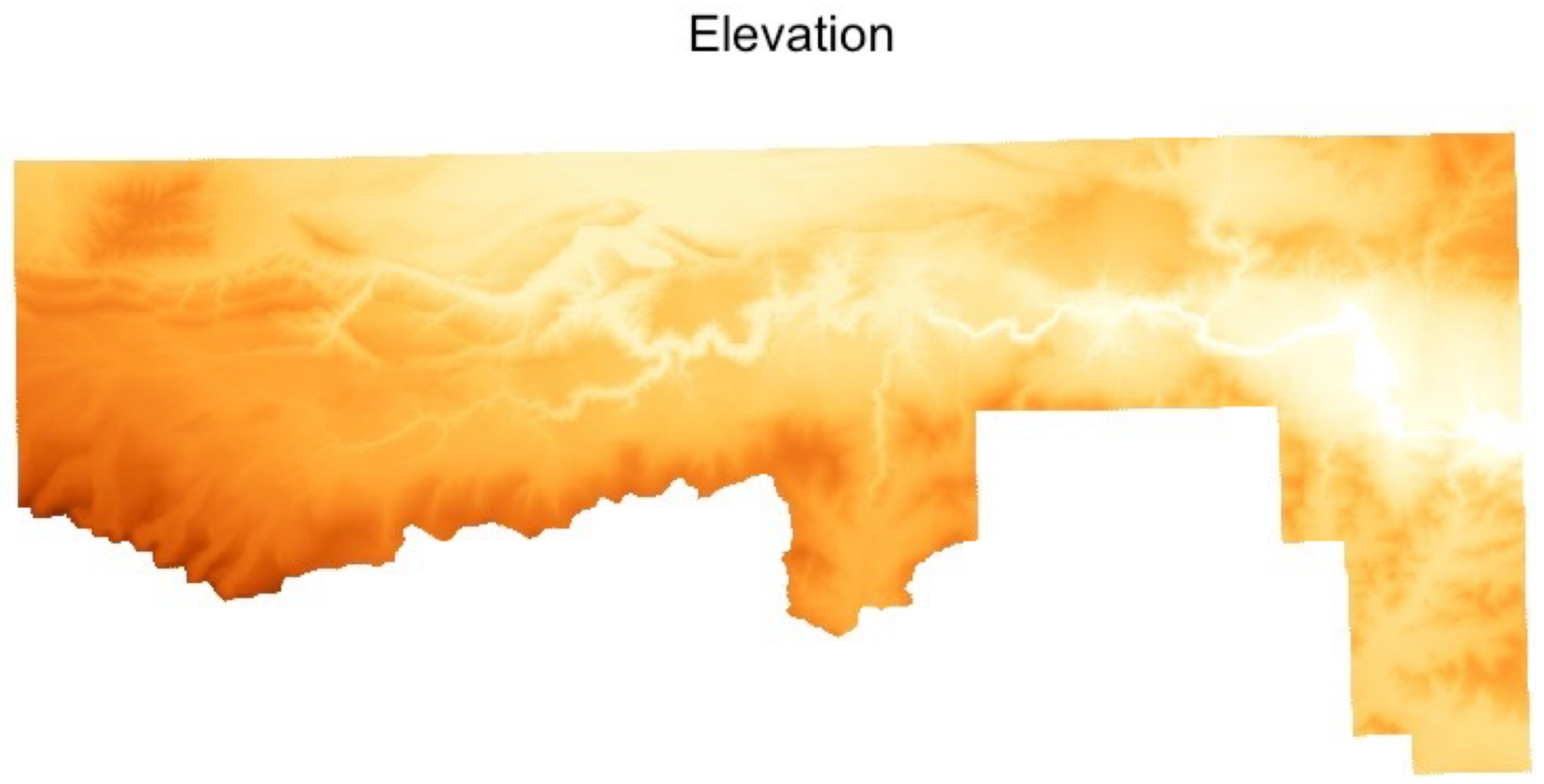
- The US Forest Service, Forest Inventory & Analysis (FIA) Program collects survey data on the forests across the US: ground crews visit selected locations to take measurements of timber supply, forest health, etc.

Motivating application: Small area estimation in forest inventory

- The US Forest Service, Forest Inventory & Analysis (FIA) Program collects <u>survey data</u> on the forests across the US: ground crews visit selected locations to take measurements of timber supply, forest health, etc.

- FIA supplements its survey data with a wide variety of remotely-sensed <u>auxiliary data</u>, with “wall-to-wall” coverage.





Motivating application: Small area estimation in forest inventory

- The FIA Program has funded a variety of projects aimed to operationalize model-based small area estimation of forest attributes at national scale.

- Despite a wide variety of funded projects, there has been little formal evaluation of small area estimators across projects.

When picking an approach to estimate some parameter μ of interest, we have so many choices!

When picking an approach to estimate some parameter μ of interest, we have so many choices!

- Direct?

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

When picking an approach to estimate some parameter μ of interest, we have so many choices!

- Direct?
- 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 picking an approach to estimate some parameter μ of interest, we have so many choices!

- Direct?
- 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 picking an approach to estimate some parameter μ of interest, we have so many choices!

- Direct?
- 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 picking an approach to estimate some parameter μ of interest, we have so many choices!

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

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

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

How do we choose?

Model form?

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

Priors?

Fitting method?

Artificial populations are a great tool to assess small area estimators*

Artificial populations are a great tool to assess small area estimators*

* When the artificial population created is a sensible depiction of reality, and

* the population generation process is largely different than the models driving the small area estimators of interest (i.e. it doesn't favor one approach over another).

KBAABB creates artificial populations based on kNN in auxiliary data space with selection weights derived from ABB

KBAABB creates artificial populations based on kNN in auxiliary data space with selection weights derived from ABB

- These weights correspond to probability of inclusion in a bootstrap sample.
- KBAABB is computationally efficient and simple to implement.

KBAABB Imputation

Recipient row

canopy_cover	roughness	water_def
0.517	1.748	-0.763

Step 1: select a recipient row of data

KBAABB Imputation

Recipient row

canopy_cover	roughness	water_def
0.517	1.748	-0.763

Step 1: select a recipient row of data

Step 2: find k nearest neighbors

Potential donors

biomass	canopy_cover	roughness	water_def
14	0.615	2.259	-0.329
33	0.500	2.083	-1.864
28	0.547	1.877	-0.748
22	0.505	2.452	-0.495
7	0.537	1.802	-0.742
17	0.495	1.907	-0.203
21	0.474	1.955	-0.289
10	0.391	1.801	-0.994
22	0.555	1.546	-1.297
40	0.522	1.295	-0.475

KBAABB Imputation

Potential donors

biomass	canopy_cover	roughness	water_def	Rank
14	0.615	2.259	-0.329	4
33	0.500	2.083	-1.864	8
28	0.547	1.877	-0.748	7
22	0.505	2.452	-0.495	2
7	0.537	1.802	-0.742	1
17	0.495	1.907	-0.203	6
21	0.474	1.955	-0.289	9
10	0.391	1.801	-0.994	10
22	0.555	1.546	-1.297	3
40	0.522	1.295	-0.475	5

Recipient row

canopy_cover	roughness	water_def
0.517	1.748	-0.763

Step 1: select a recipient row of data

Step 2: find k nearest neighbors

Step 3: rank each potential donor by distance in *auxiliary data* space

KBAABB Imputation

Potential donors

biomass	canopy_cover	roughness	water_def	Probability
14	0.615	2.259	-0.329	0.0315
33	0.500	2.083	-1.864	0.0016
28	0.547	1.877	-0.748	0.0016
22	0.505	2.452	-0.495	0.2325
7	0.537	1.802	-0.742	0.6321
17	0.495	1.907	-0.203	0.0043
21	0.474	1.955	-0.289	0.0002
10	0.391	1.801	-0.994	0.0001
22	0.555	1.546	-1.297	0.0855
40	0.522	1.295	-0.475	0.0116

Recipient row

canopy_cover	roughness	water_def
0.517	1.748	-0.763

Step 1: select a recipient row of data

Step 2: find k nearest neighbors

Step 3: rank each potential donor by distance in *auxiliary data* space

Step 4: impute based on probability derived from rank and the ABB

KBAABB Imputation

Recipient row

canopy_cover	roughness	water_def	biomass
0.517	1.748	-0.763	22

Potential donors

biomass	canopy_cover	roughness	water_def	Probability
14	0.615	2.259	-0.329	0.0315
33	0.500	2.083	-1.864	0.0016
28	0.547	1.877	-0.748	0.0016
22	0.505	2.452	-0.495	0.2325
7	0.537	1.802	-0.742	0.6321
17	0.495	1.907	-0.203	0.0043
21	0.474	1.955	-0.289	0.0002
10	0.391	1.801	-0.994	0.0001
22	0.555	1.546	-1.297	0.0855
40	0.522	1.295	-0.475	0.0116

Step 1: select a recipient row of data

Step 2: find k nearest neighbors

Step 3: rank each potential donor by distance in *auxiliary data* space

Step 4: impute based on probability derived from rank and the ABB

KBAABB Imputation

Recipient row

canopy_cover	roughness	water_def	biomass
0.517	1.748	-0.763	22

Potential donors

biomass	canopy_cover	roughness	water_def	Probability
14	0.615	2.259	-0.329	0.0315
33	0.500	2.083	-1.864	0.0016
28	0.547	1.877	-0.748	0.0016
22	0.505	2.452	-0.495	0.2325
7	0.537	1.802	-0.742	0.6321
17	0.495	1.907	-0.203	0.0043
21	0.474	1.955	-0.289	0.0002
10	0.391	1.801	-0.994	0.0001
22	0.555	1.546	-1.297	0.0855
40	0.522	1.295	-0.475	0.0116

Step 1: select a recipient row of data

Step 2: find k nearest neighbors

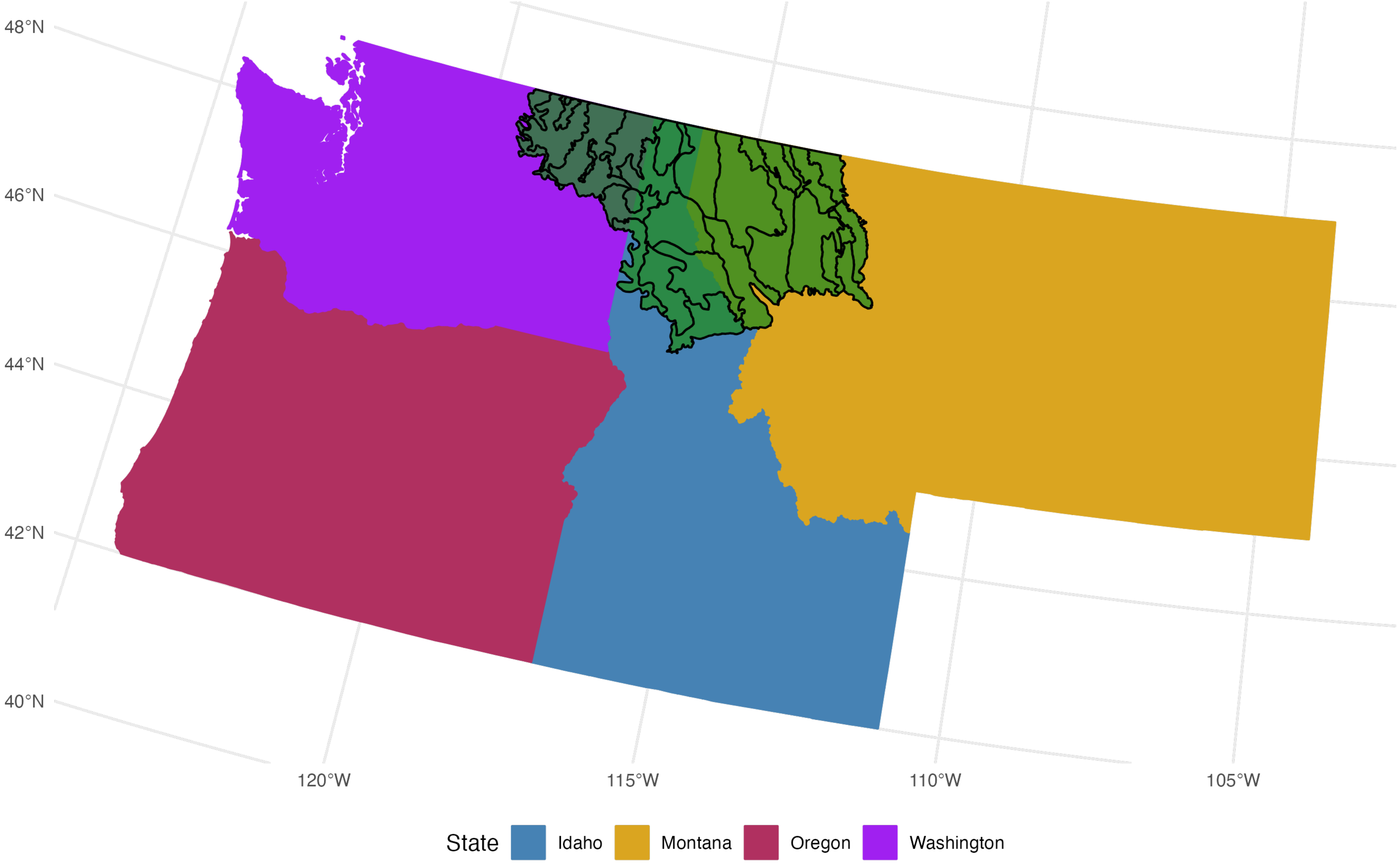
Step 3: rank each potential donor by distance in *auxiliary data* space

Step 4: impute based on probability derived from rank and the ABB

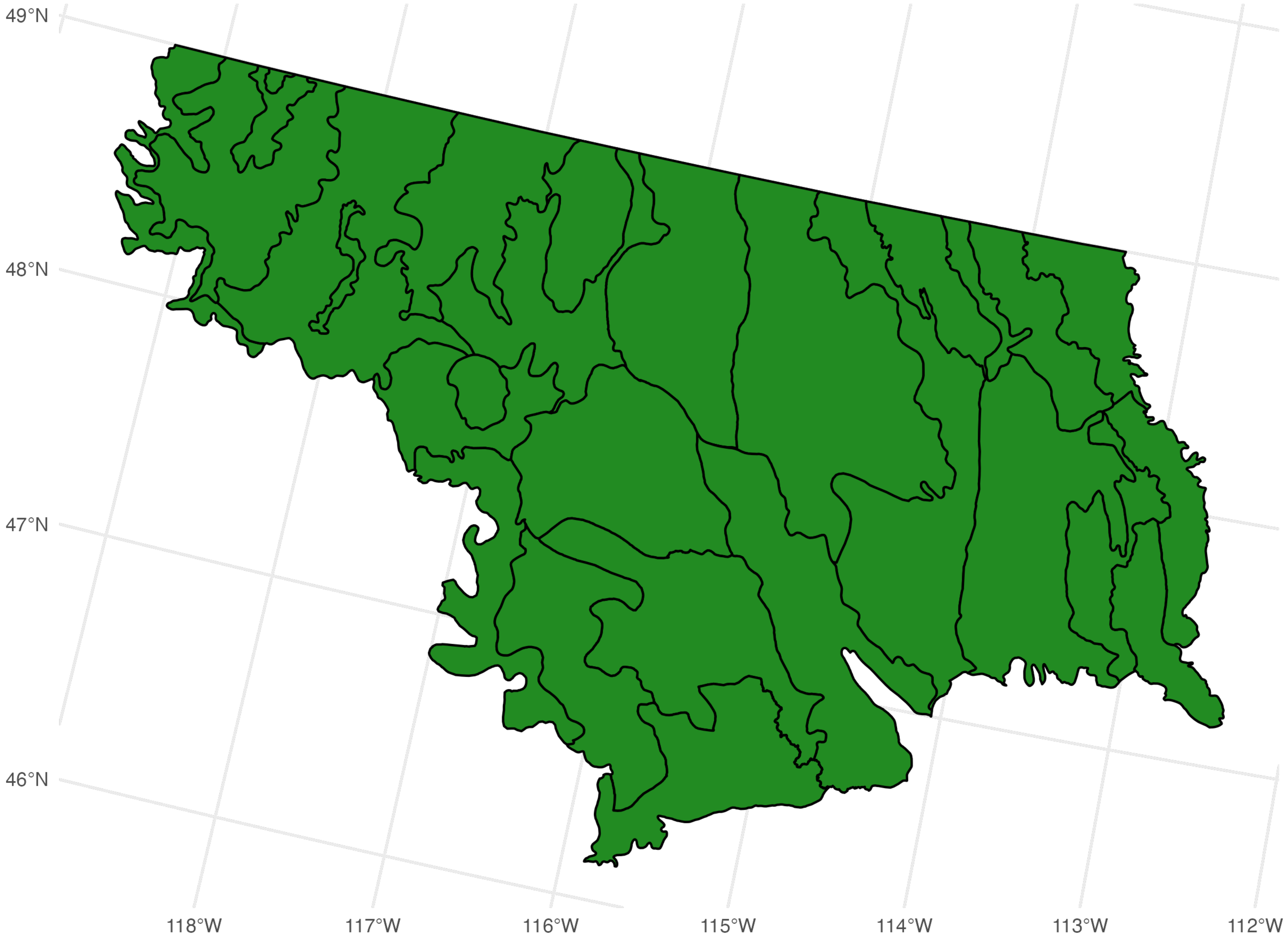
Step 5: repeat for each recipient to generate artificial population

Data Application

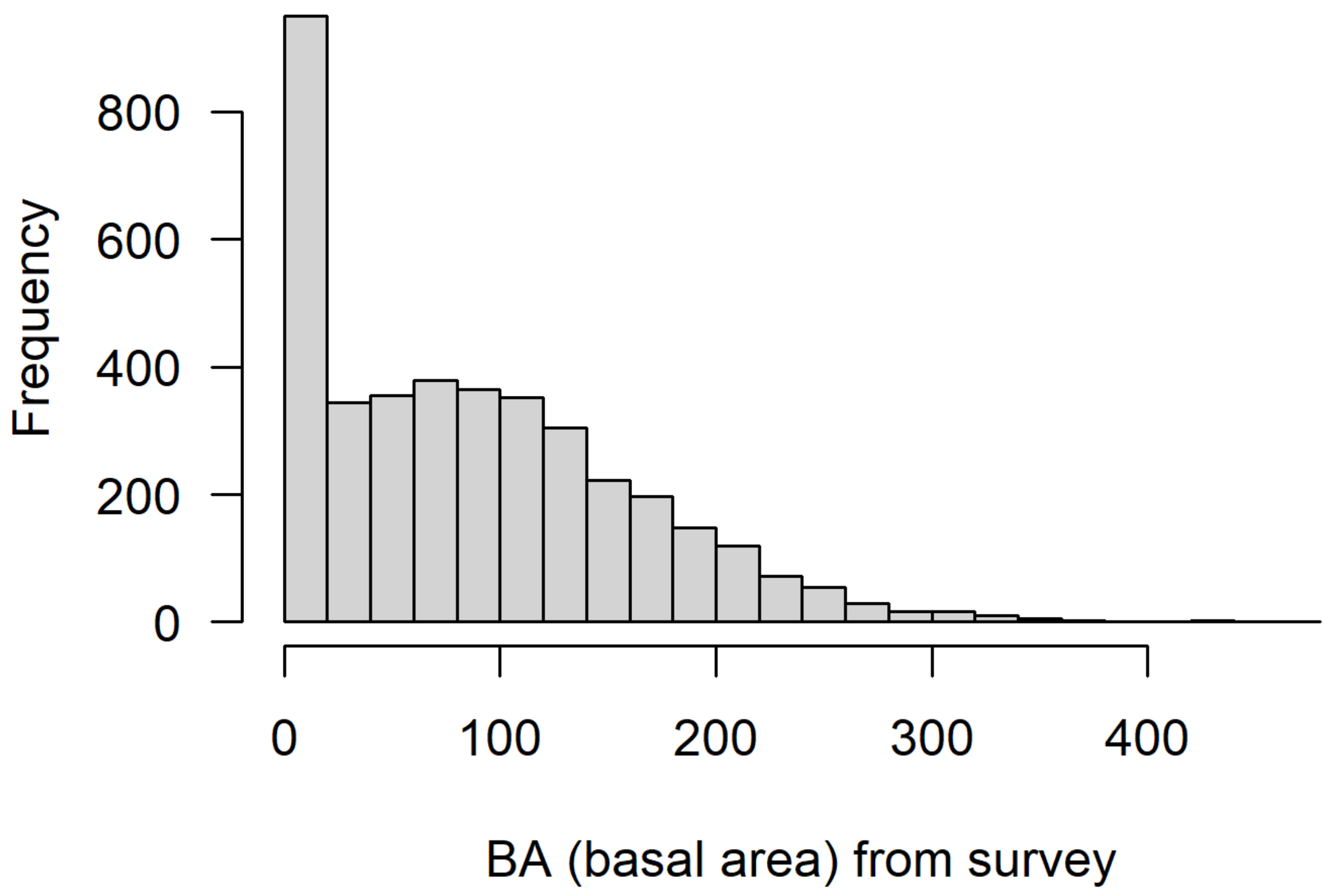
Study region: Ecoprovince M333, Areas of interest: Ecosubsections

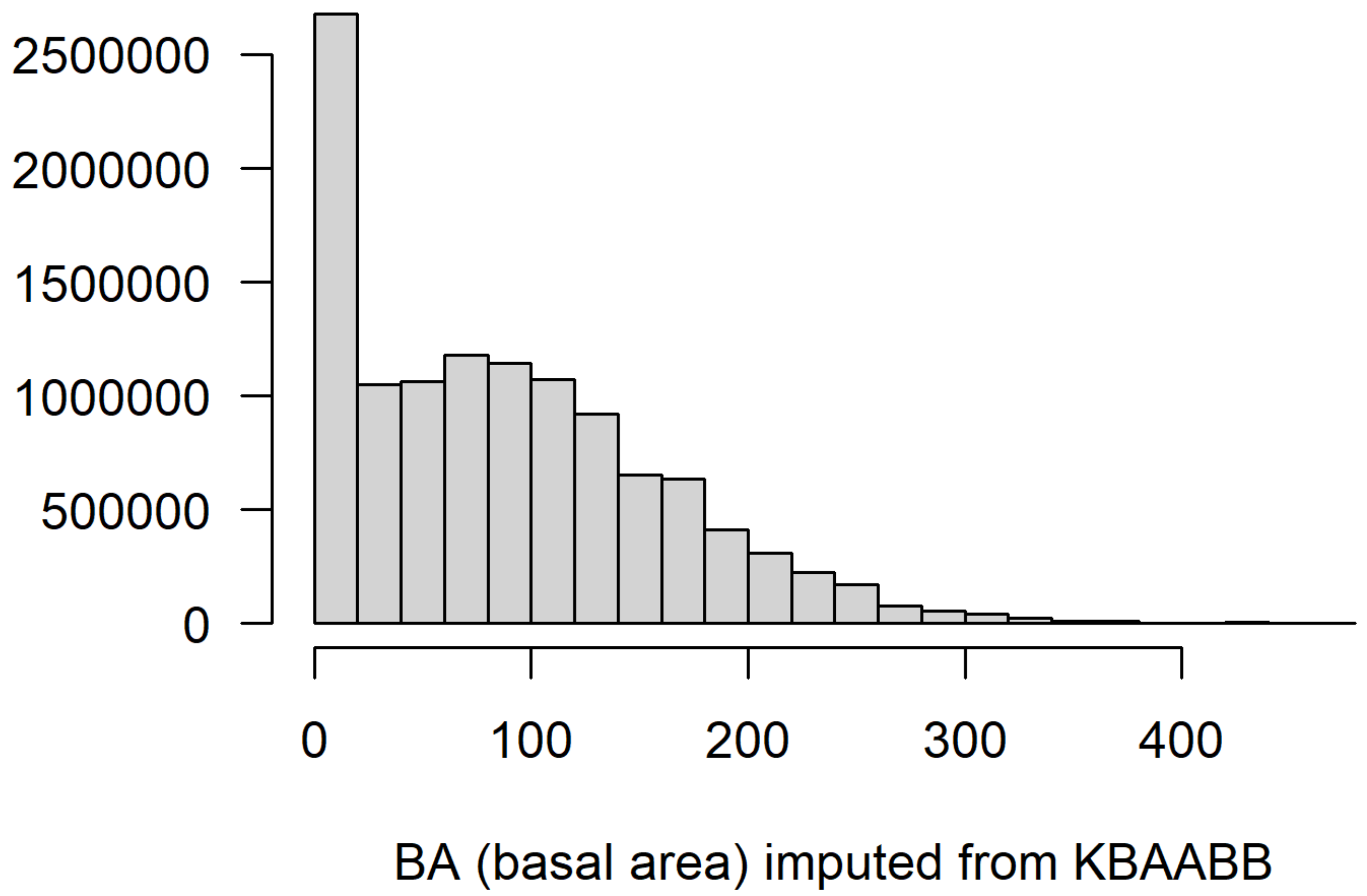


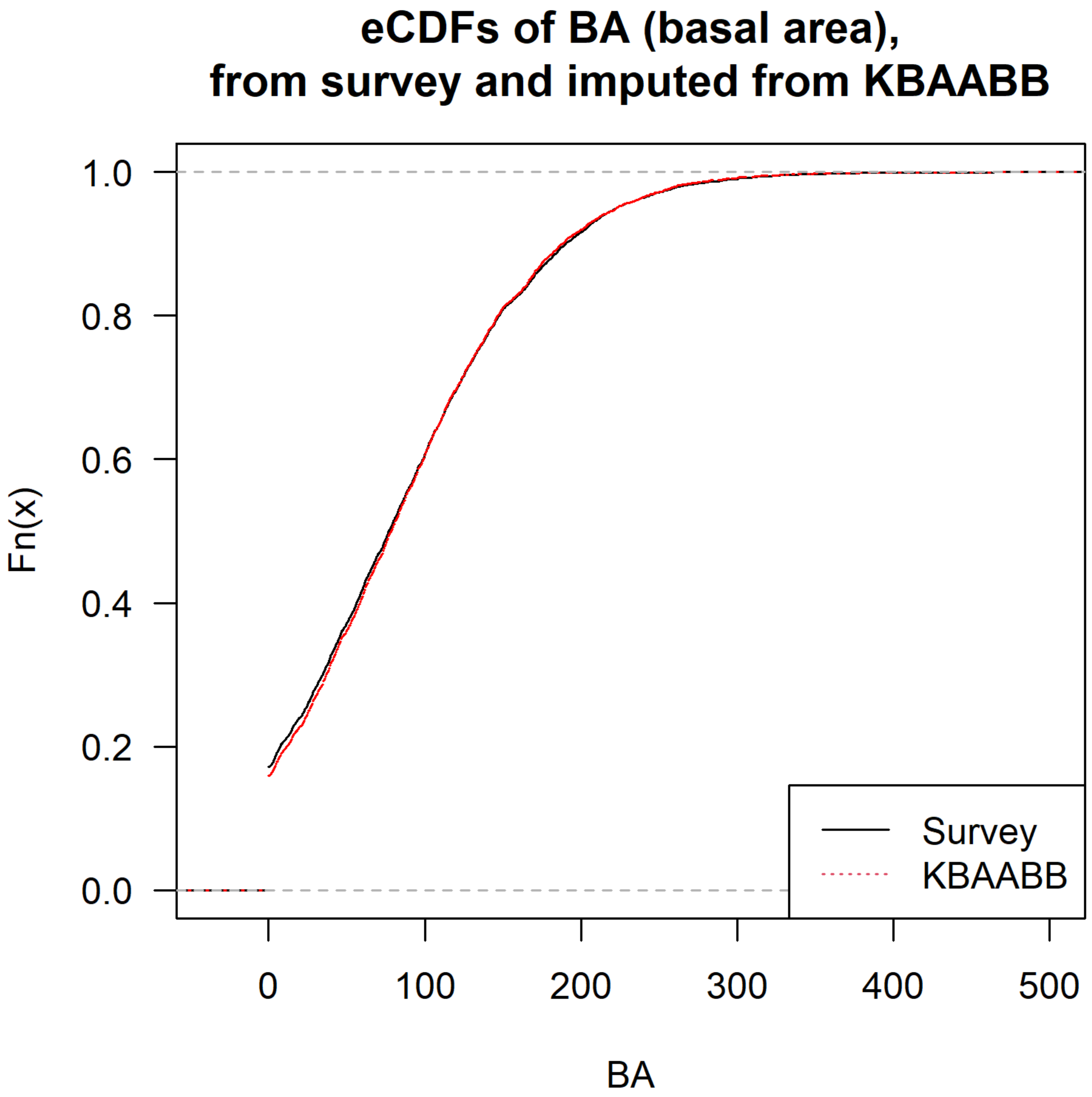
Study region: Ecoprovince M333, Areas of interest: Ecosubsections



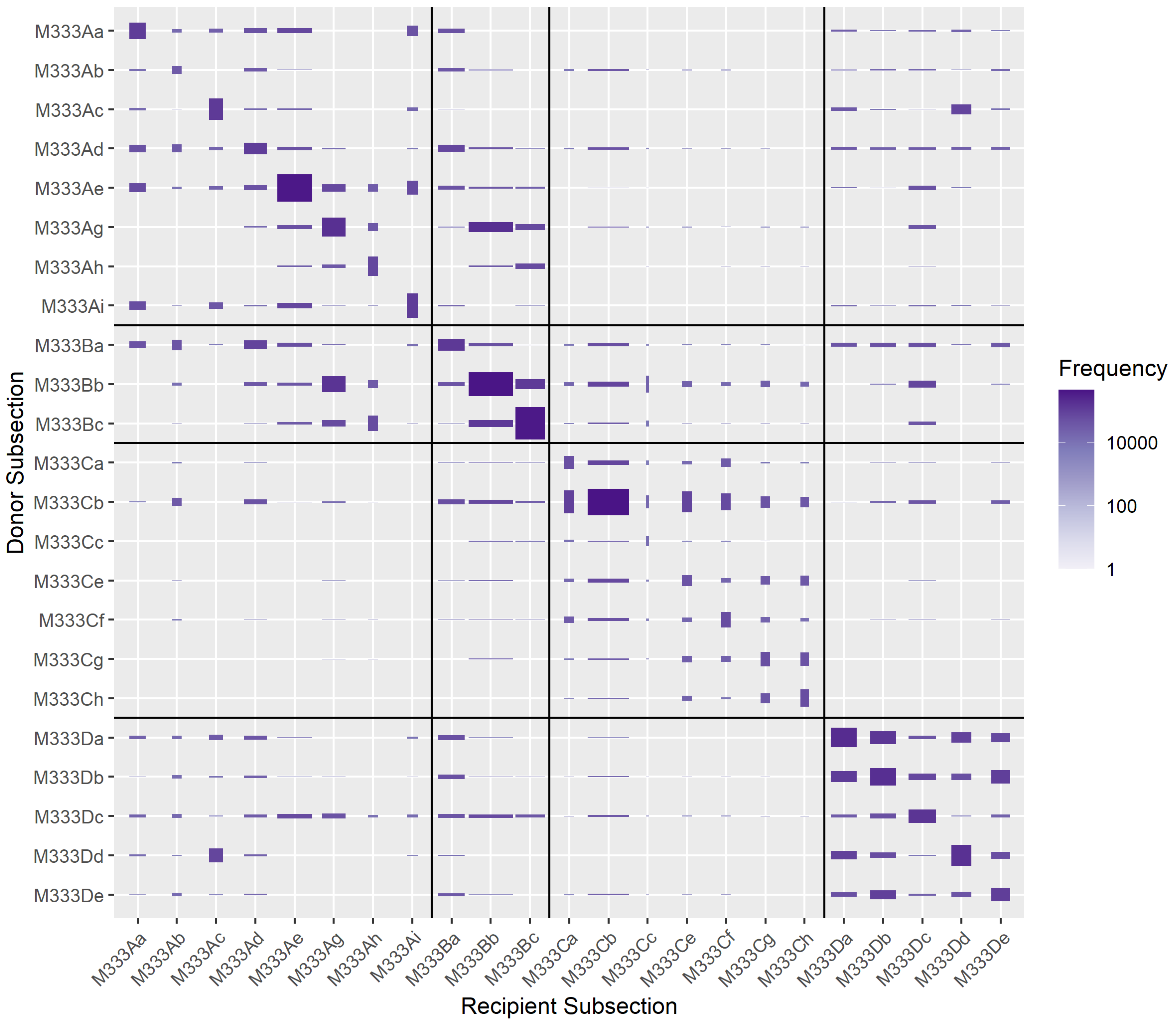
KBAABB Population Characteristics



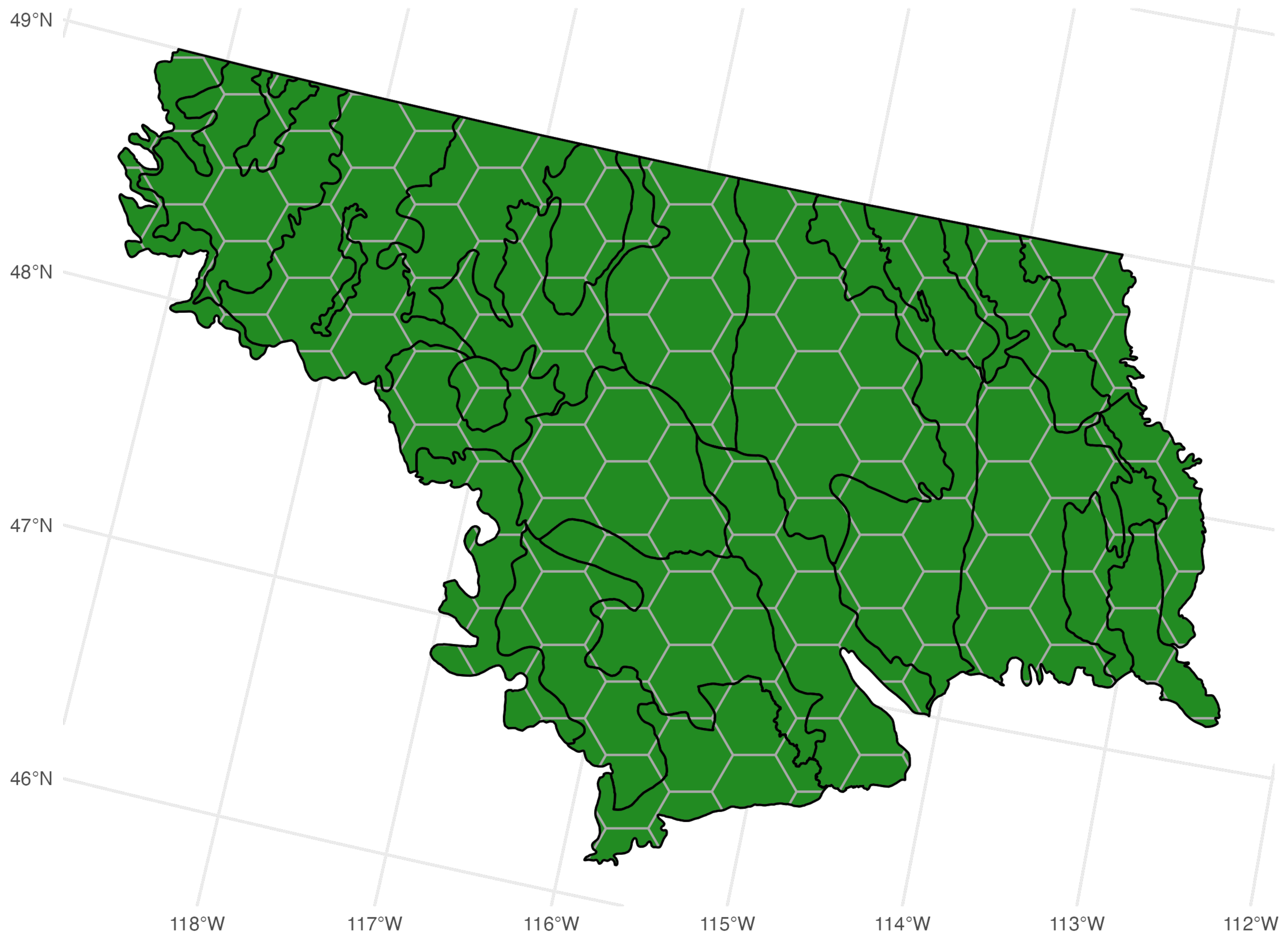




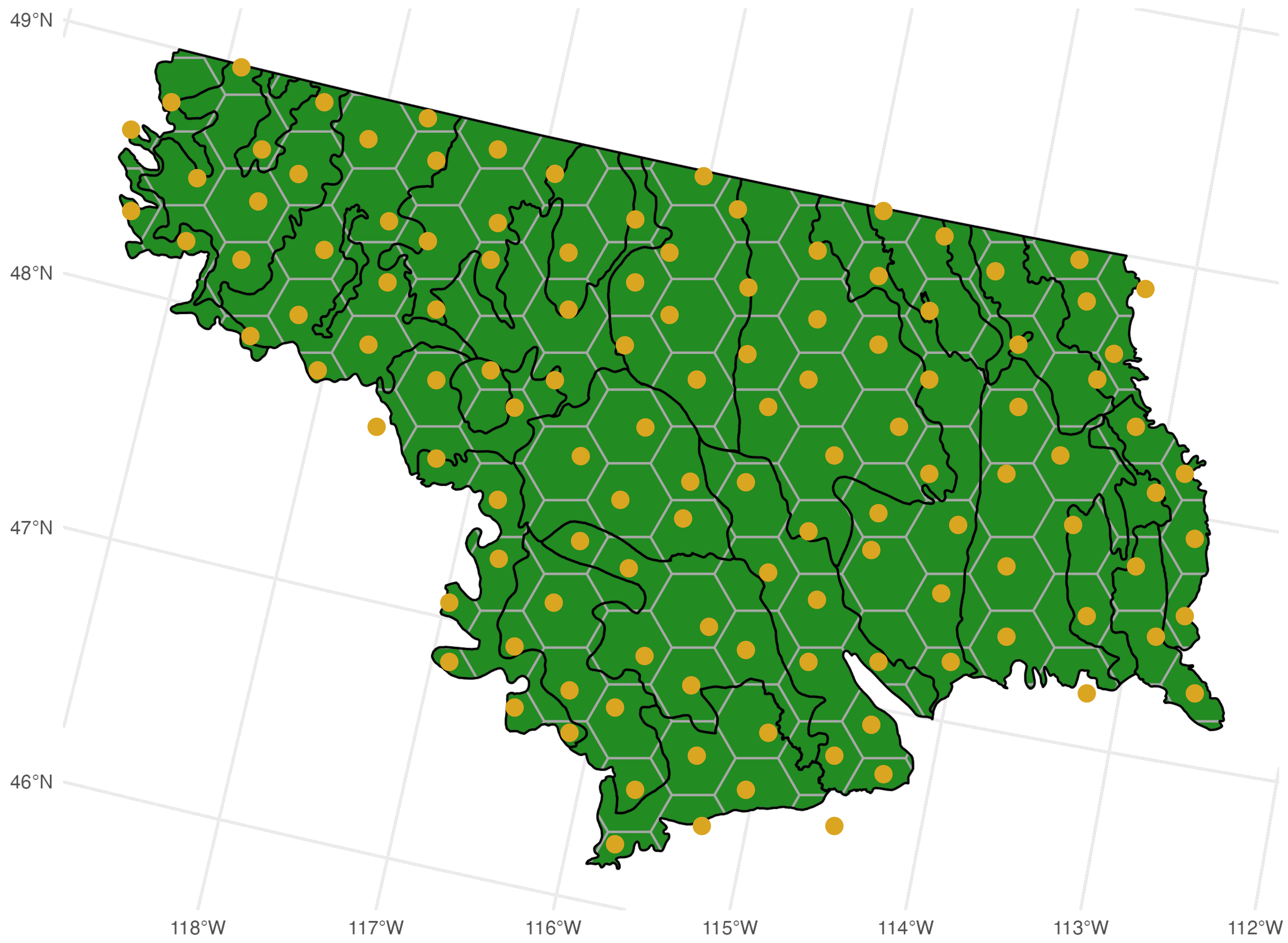
KBAABB Population Characteristics



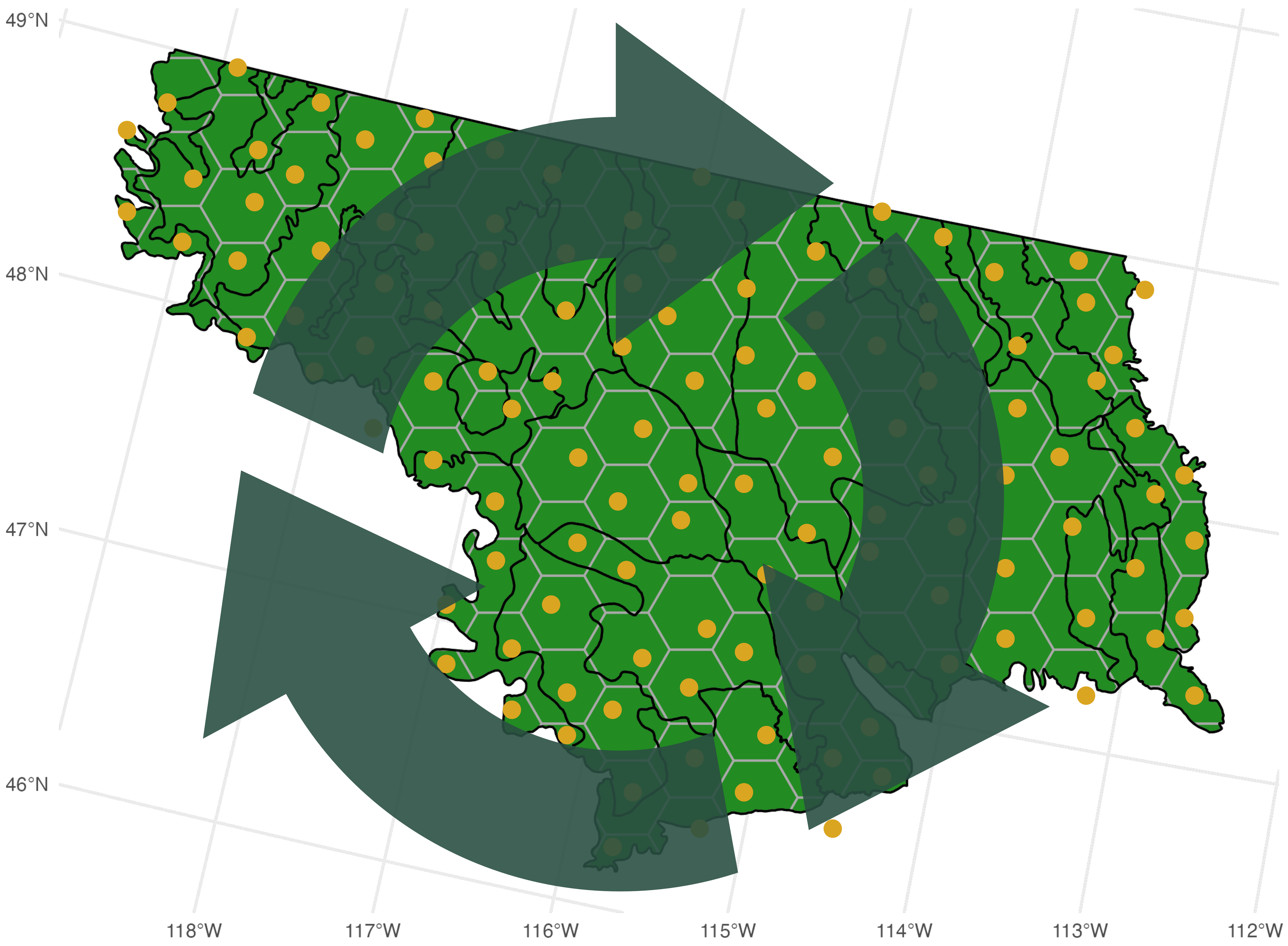
Quasi-systematic sampling design: hexagonal sampling frames



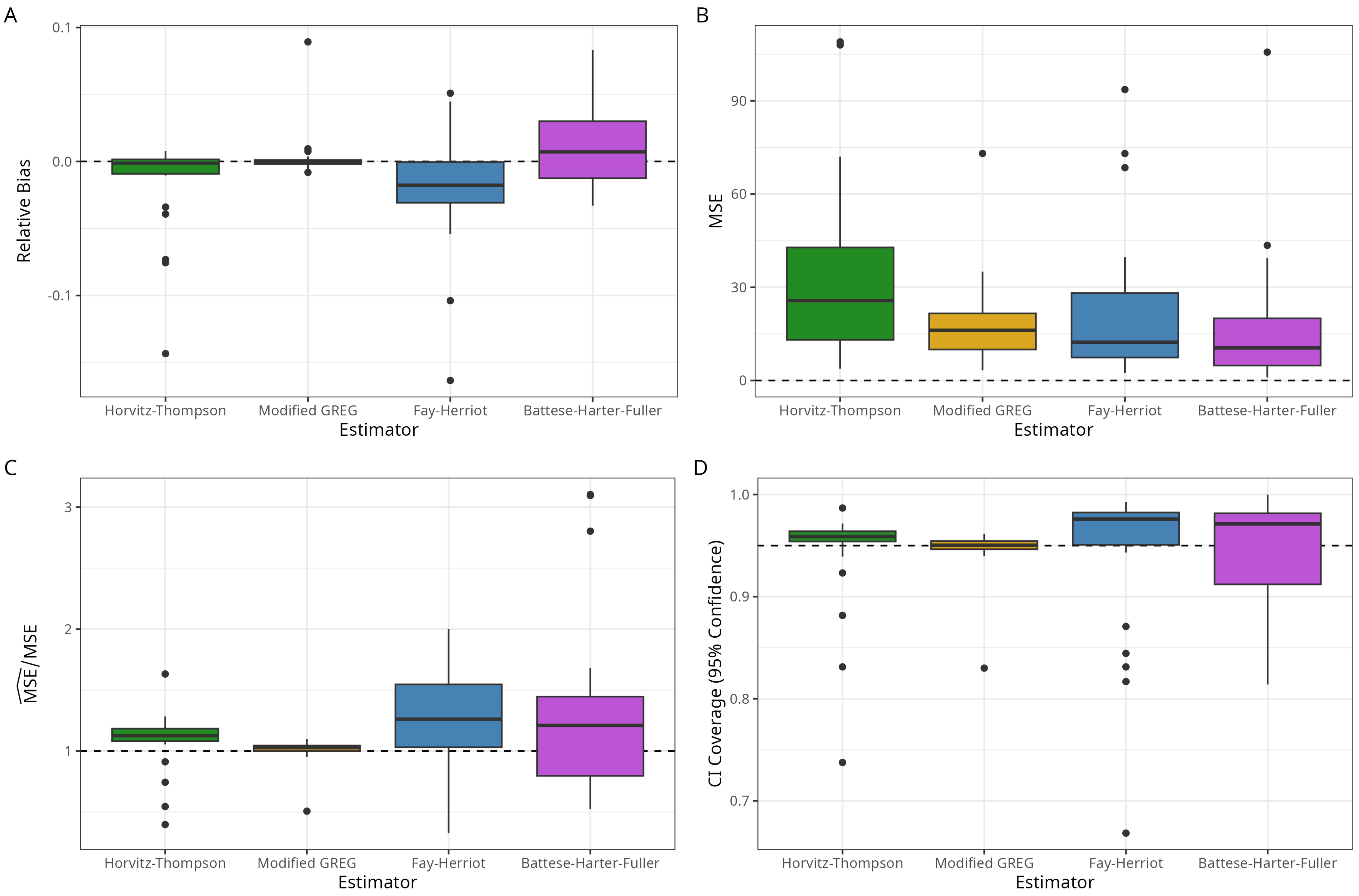
Quasi-systematic sampling design: random sample location within frame



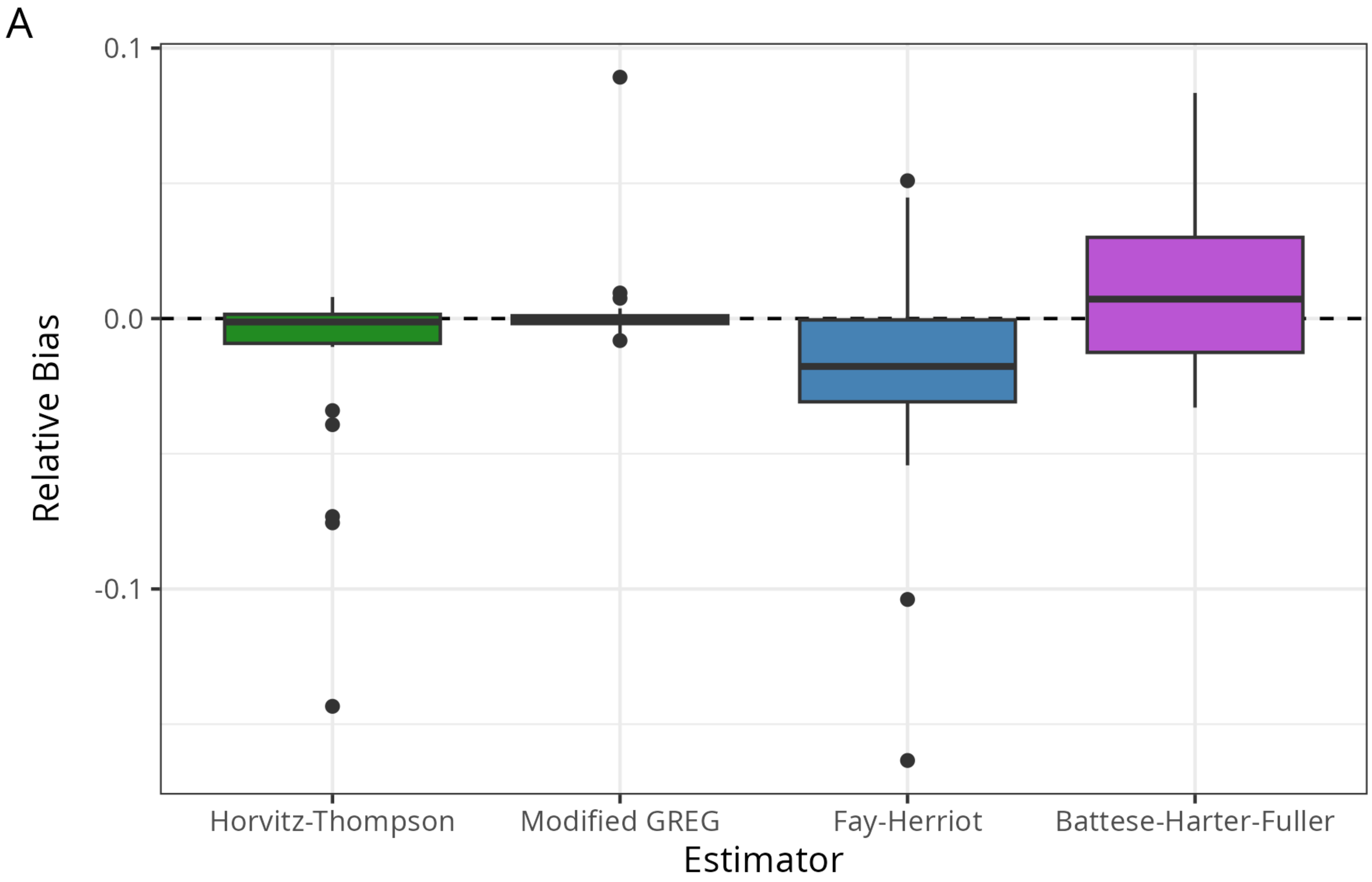
Repeat taking design-based samples from the artificial population

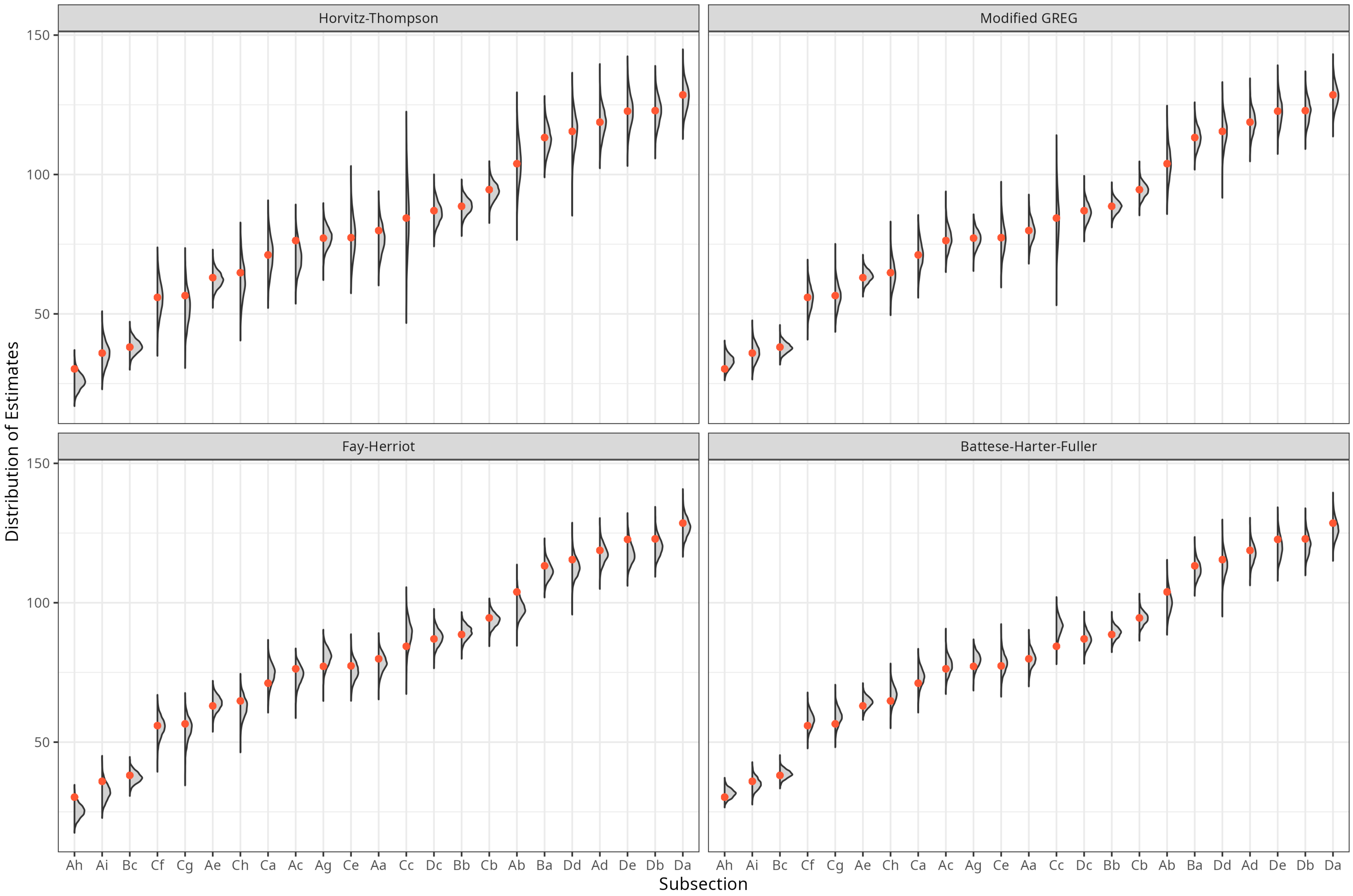


Assessing Estimators on KBAABB Samples



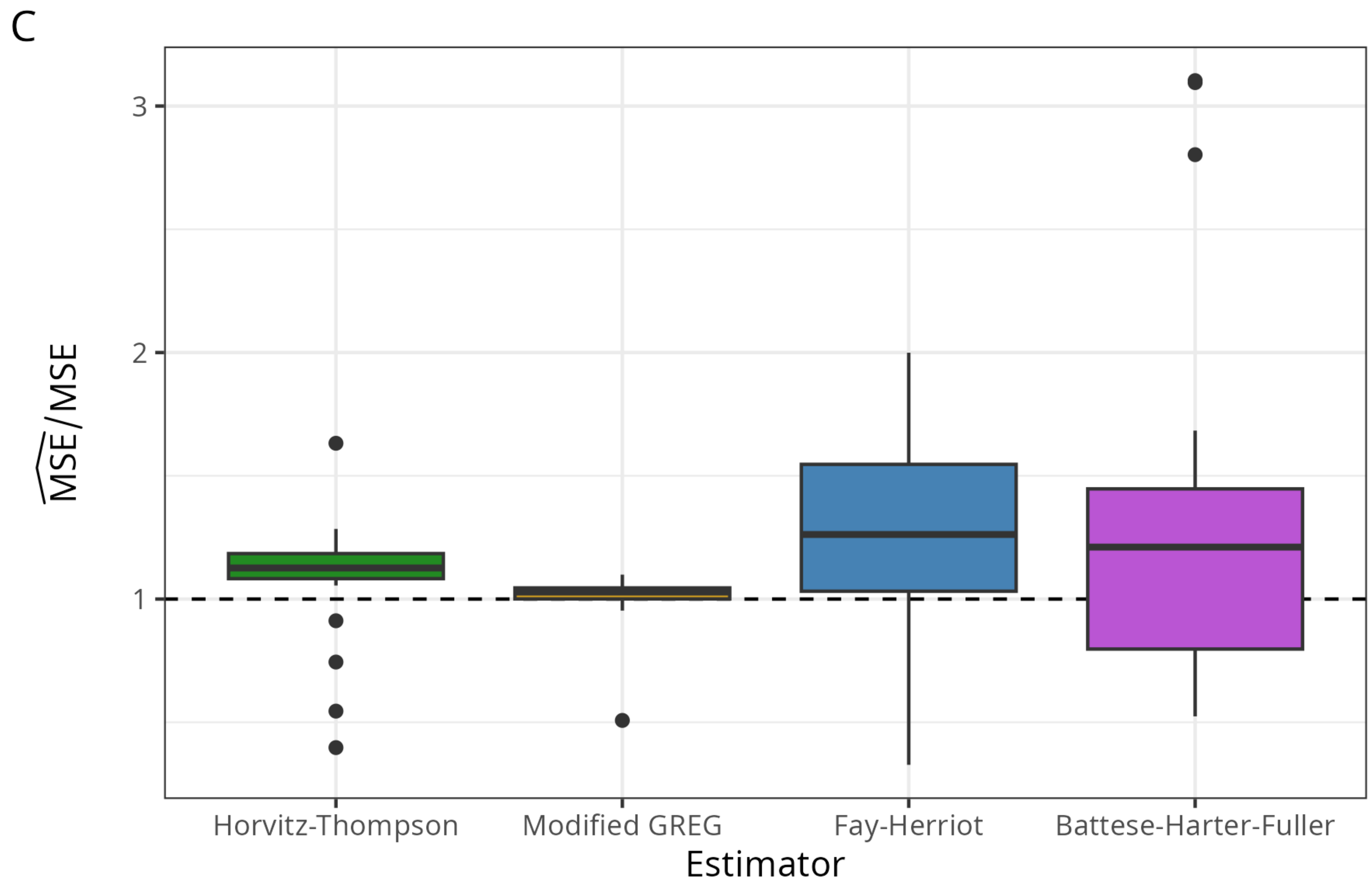
Bias: What's going on here?

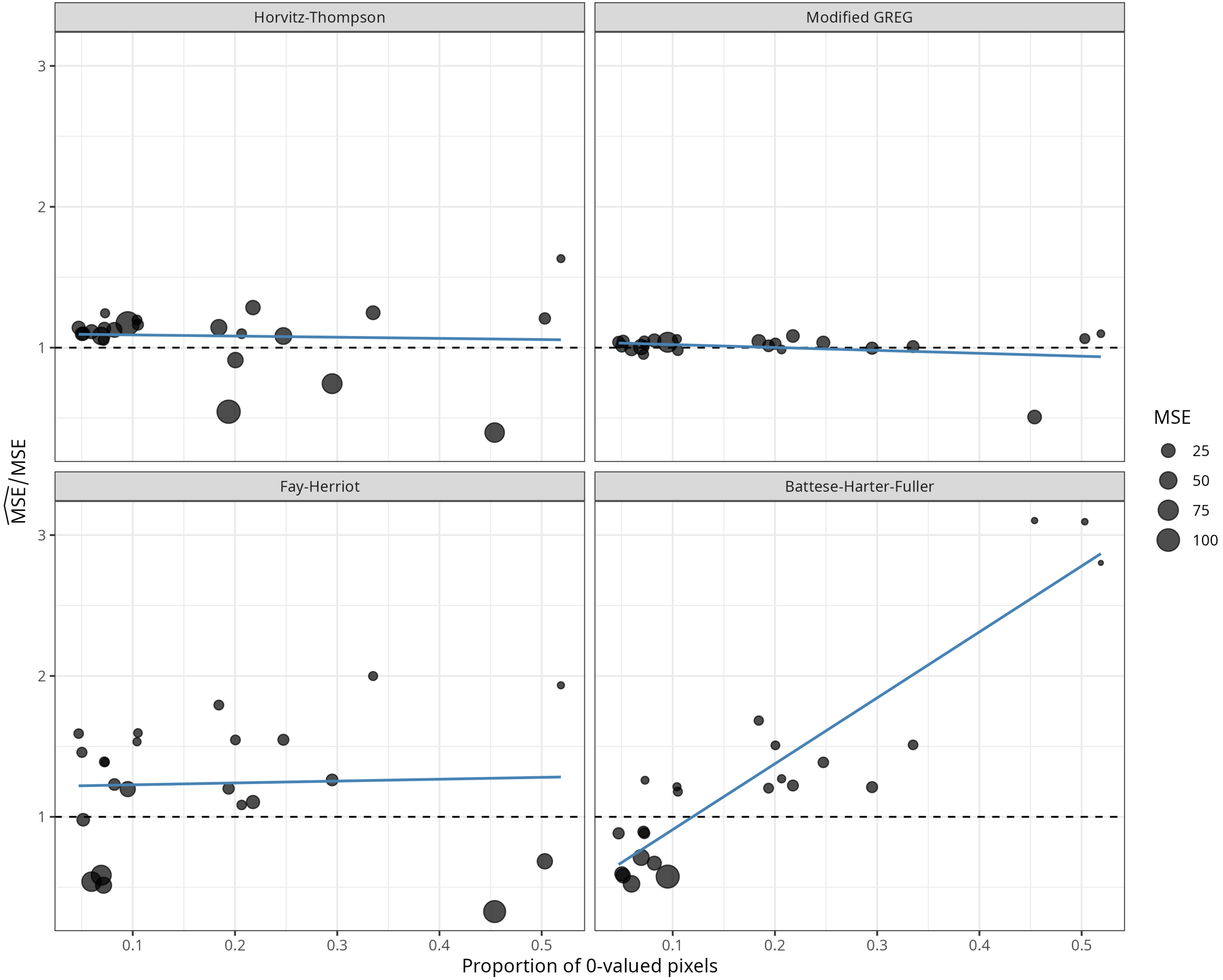




Horvitz-Thompson
Modified GREG
Fay-Herriot
Battese-Harter-Fuller
150
100
50
Distribution of Estimates
Ah Ai Bc Cf Cg Ae Ch Ca Ac Ag Ce Aa Cc Dc Bb Cb Ab Ba Dd Ad De Db Da
Subsection

MSE Estimate Behavior





Horvitz-Thompson
Modified GREG
Fay-Herriot
Battese-Harter-Fuller
$\widehat{\text{MSE}}$/MSE
Proportion of 0-valued pixels
MSE
25
50
75
100

kbaabb R Package(2)

- The `kbaabb` R package allows users to create an imputed population based on the KBAABB methodology.

- Currently under development, use at your own risk!

- Available on GitHub: www.github.com/graysonwhite/kbaabb/

- Install with:

```
# install.packages(‘‘devtools’’)

devtools::install_github(‘‘graysonwhite/kbaabb’’)
```

Future Work

1. Spatial smoothing for more realistic artificial populations
2. National application of methodology
3. Software improvements

Image credit: USFS Forest Atlas

Thank You! Questions?

References

1. White, Grayson W. et al. (2024). Assessing small area estimates via bootstrap-weighted k-Nearest-Neighbor artificial populations. arXiv: 2306.15607 [stat.ME].

2. White, Grayson W. et al. (2024). kbaabb: Generates an Artificial Population Based on the KBAABB Methodology. R package version 0.0.0.9000.