A national application of the KBAABB imputation methodology for formal assessment of small area estimators

Grayson W. White[1], Andrew O. Finley[1], Jerzy A. Wieczorek[2], Kelly S. McConville[3], Tracey S. Frescino[4]

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

PSAE Meeting 2024 Tucson, Arizona, USA April 4, 2024

Motivation and Methodology

When estimating parameters about some forest attribute of interest, we have so many choices!

When estimating parameters about 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 parameters about 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 parameters about 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 parameters about 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 parameters about some forest attribute of interest, we have so many choices!

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

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

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

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.

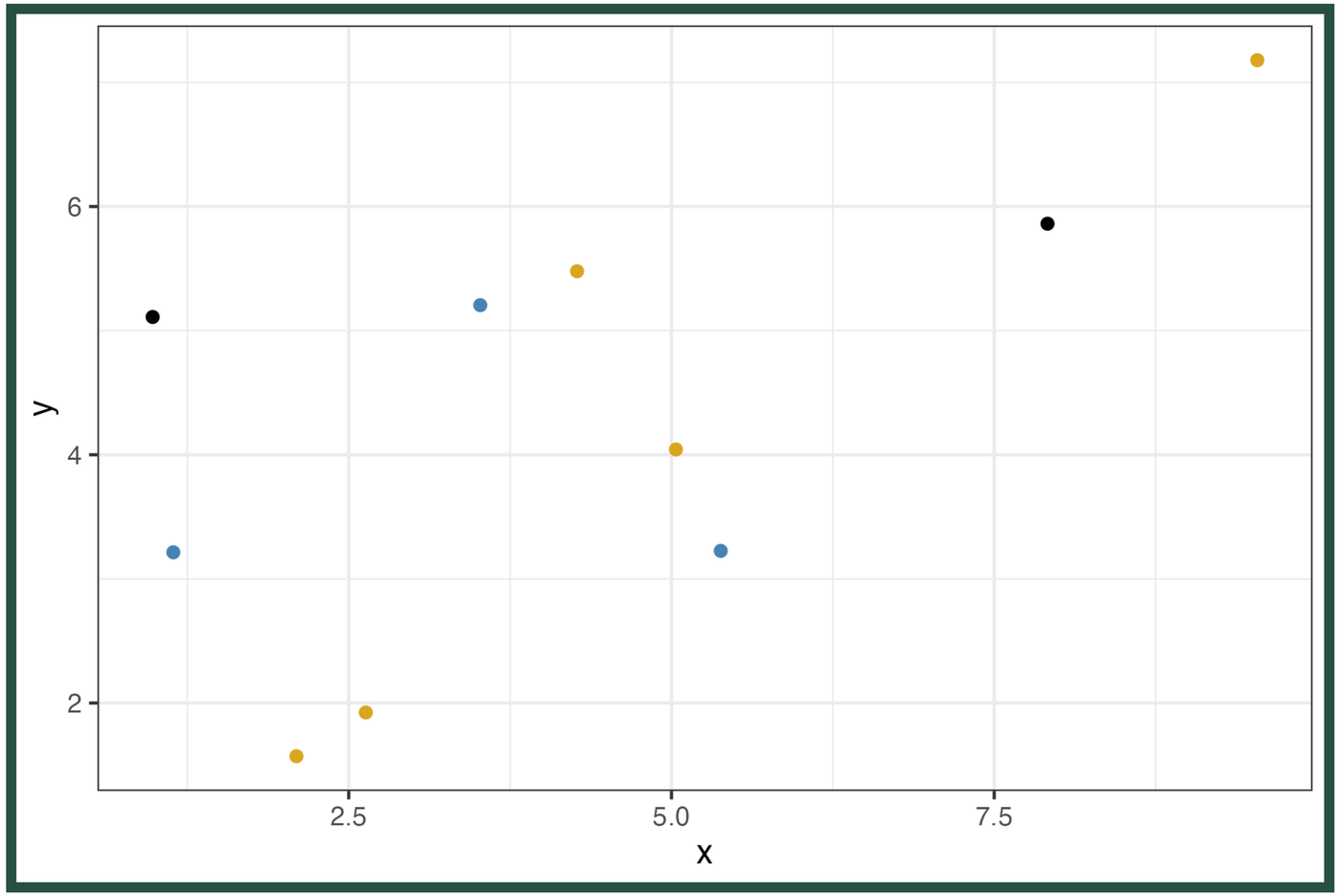
In order to help assess estimators, we have proposed the KBAABB methodology[1]. We use ideas from the following for our artificial population generation:

- ***k*NN[2]**
- **Hot deck imputation[3]**
- **Approximate Bayesian bootstrap (ABB)[4]**

In order to help assess estimators, we have proposed the KBAABB methodology[1]. We use ideas from the following for our artificial population generation:

*k*NN[2]

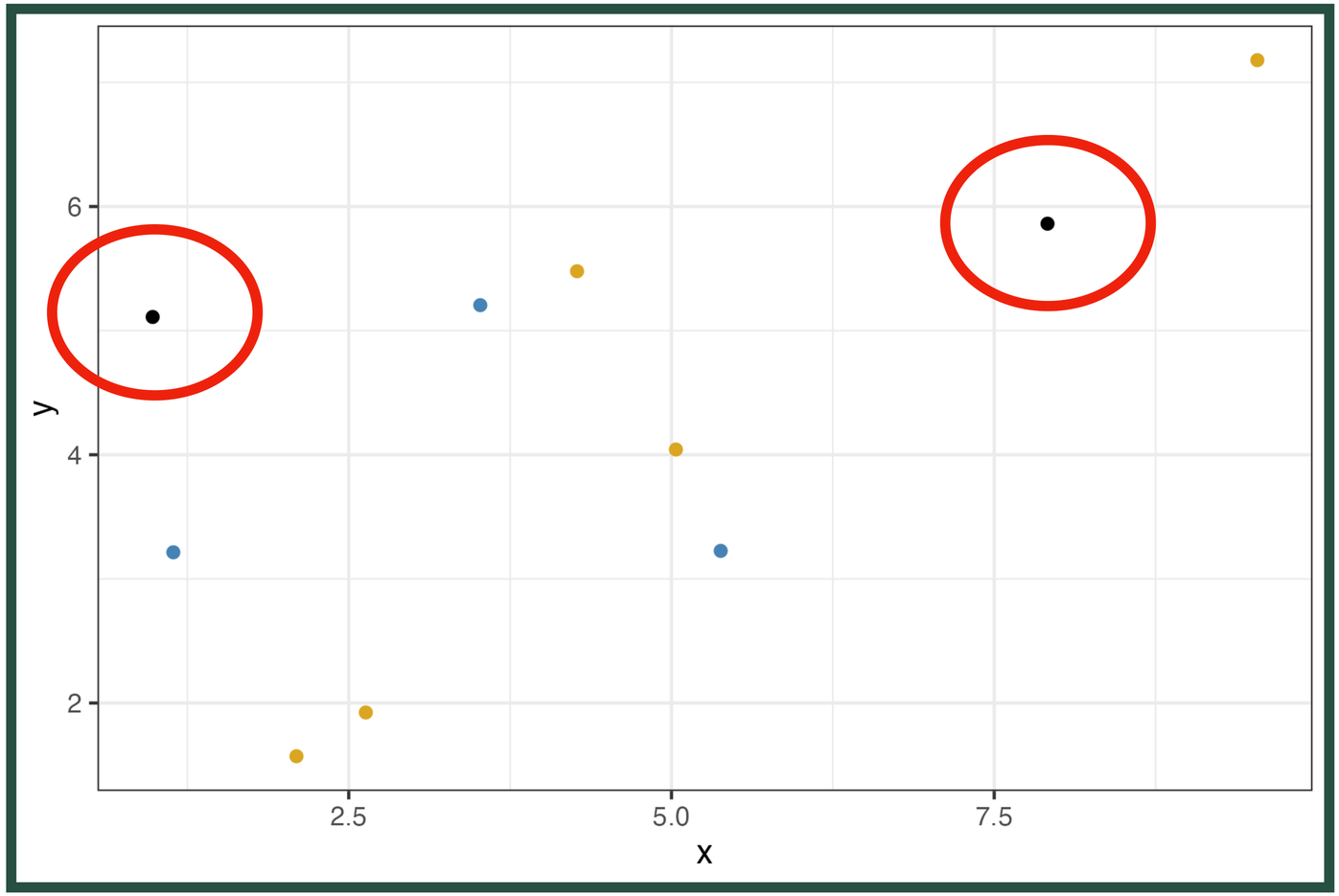
A non-parametric algorithm for finding the k "closest" observations to a given observation.



In order to help assess estimators, we have proposed the KBAABB methodology[1]. We use ideas from the following for our artificial population generation:

*k*NN[2]

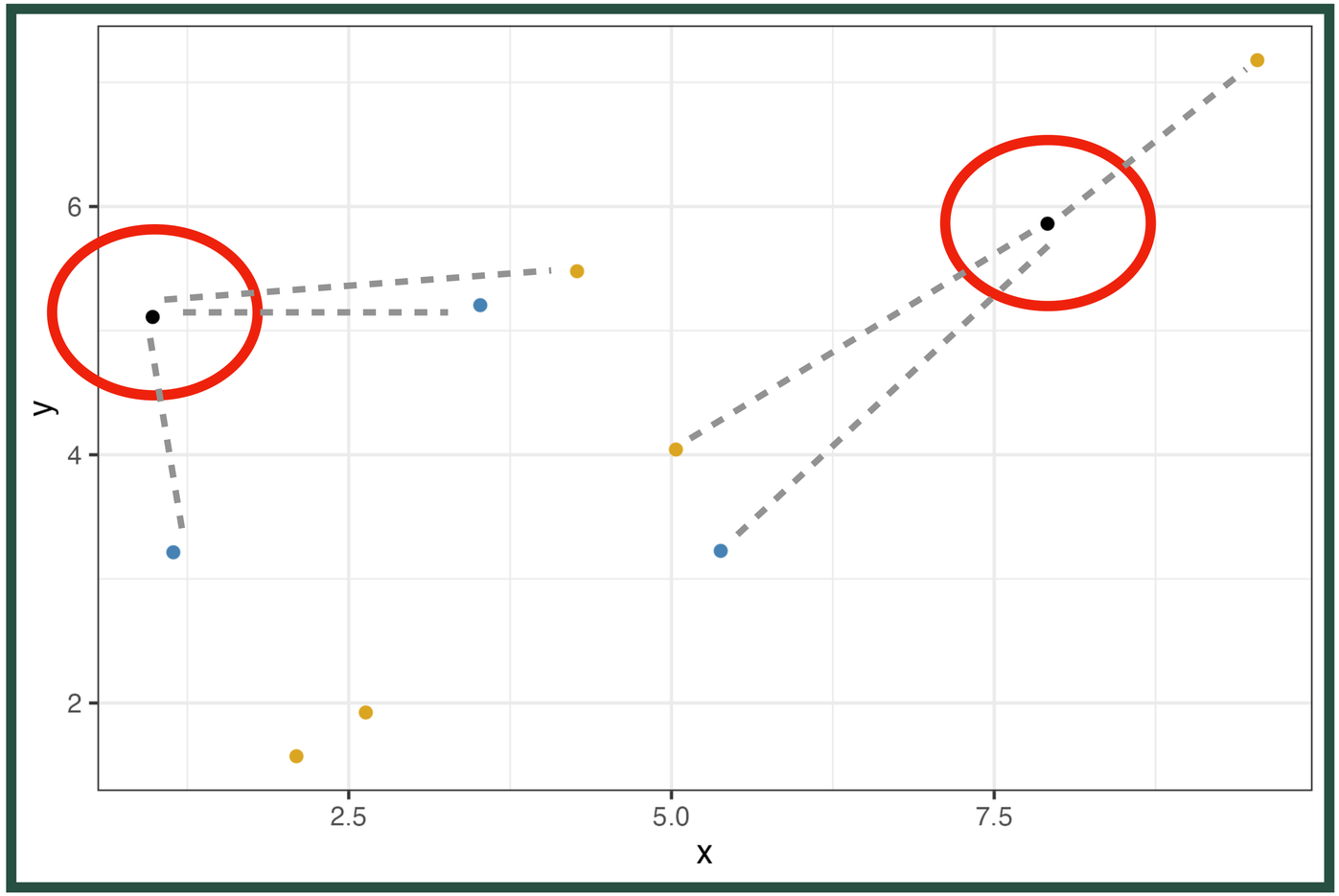
A non-parametric algorithm for finding the k "closest" observations to a given observation.



In order to help assess estimators, we have proposed the KBAABB methodology[1]. We use ideas from the following for our artificial population generation:

*k*NN[2]

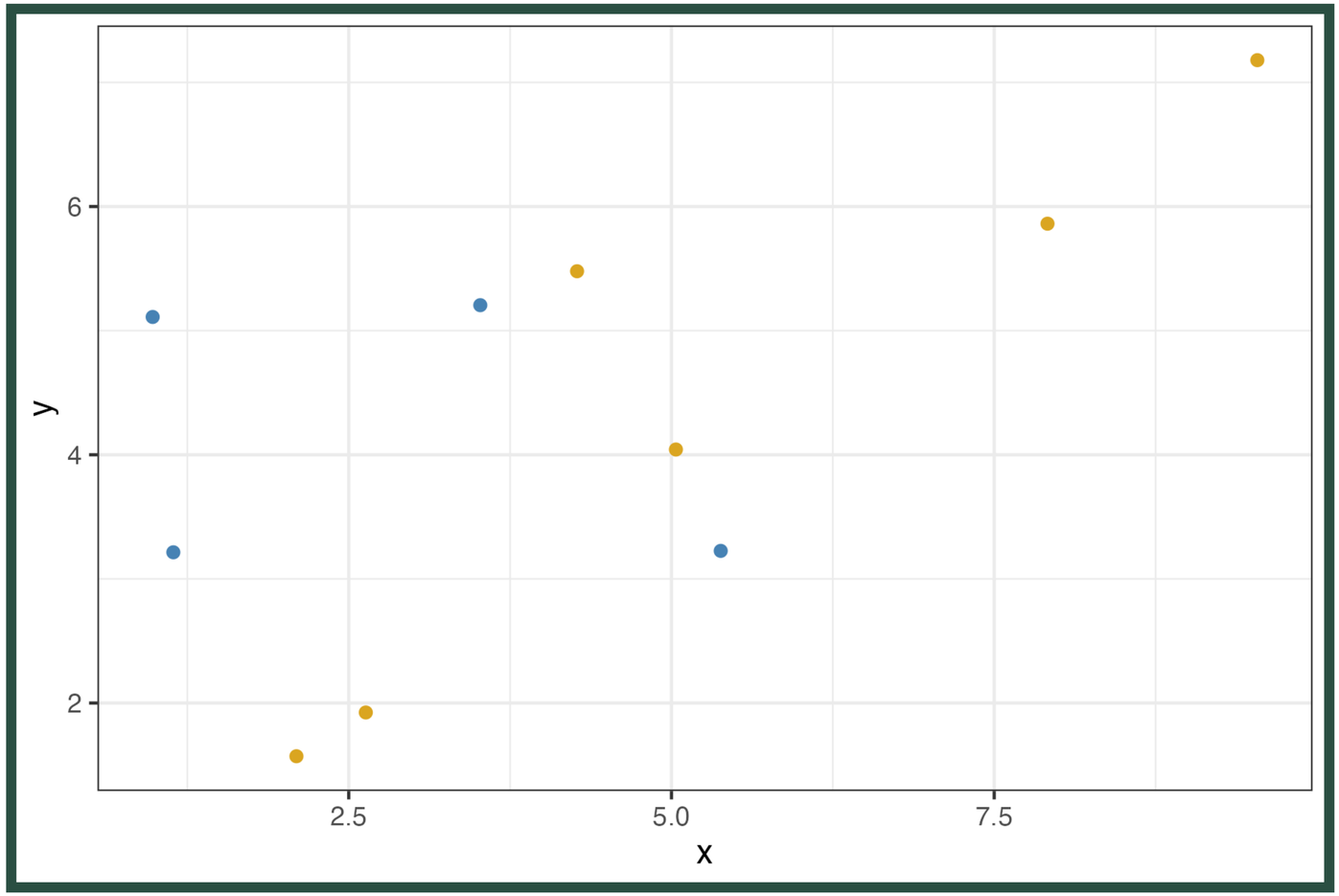
A non-parametric algorithm for finding the k "closest" observations to a given observation.



In order to help assess estimators, we have proposed the KBAABB methodology[1]. We use ideas from the following for our artificial population generation:

*k*NN[2]

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



In order to help assess estimators, we have proposed the KBAABB methodology[1]. We use ideas from the following for our artificial population generation:

*k*NN[2]

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

Hot deck imputation[3]

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

Approximate Bayesian bootstrap (ABB)[4]

In order to help assess estimators, we have proposed the KBAABB methodology[1]. We use ideas from the following for our artificial population generation:

*k*NN[2]

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

Hot deck imputation[3]

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

Approximate Bayesian bootstrap (ABB)[4]

A hot deck imputation method that results in non-uniform probability of selection.

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 and provide a computational shortcut to the ABB.
- 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

Recipient row

canopy_cover	roughness	water_def
0.517	1.748	-0.763

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

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

Recipient row

canopy_cover	roughness	water_def
0.517	1.748	-0.763

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

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

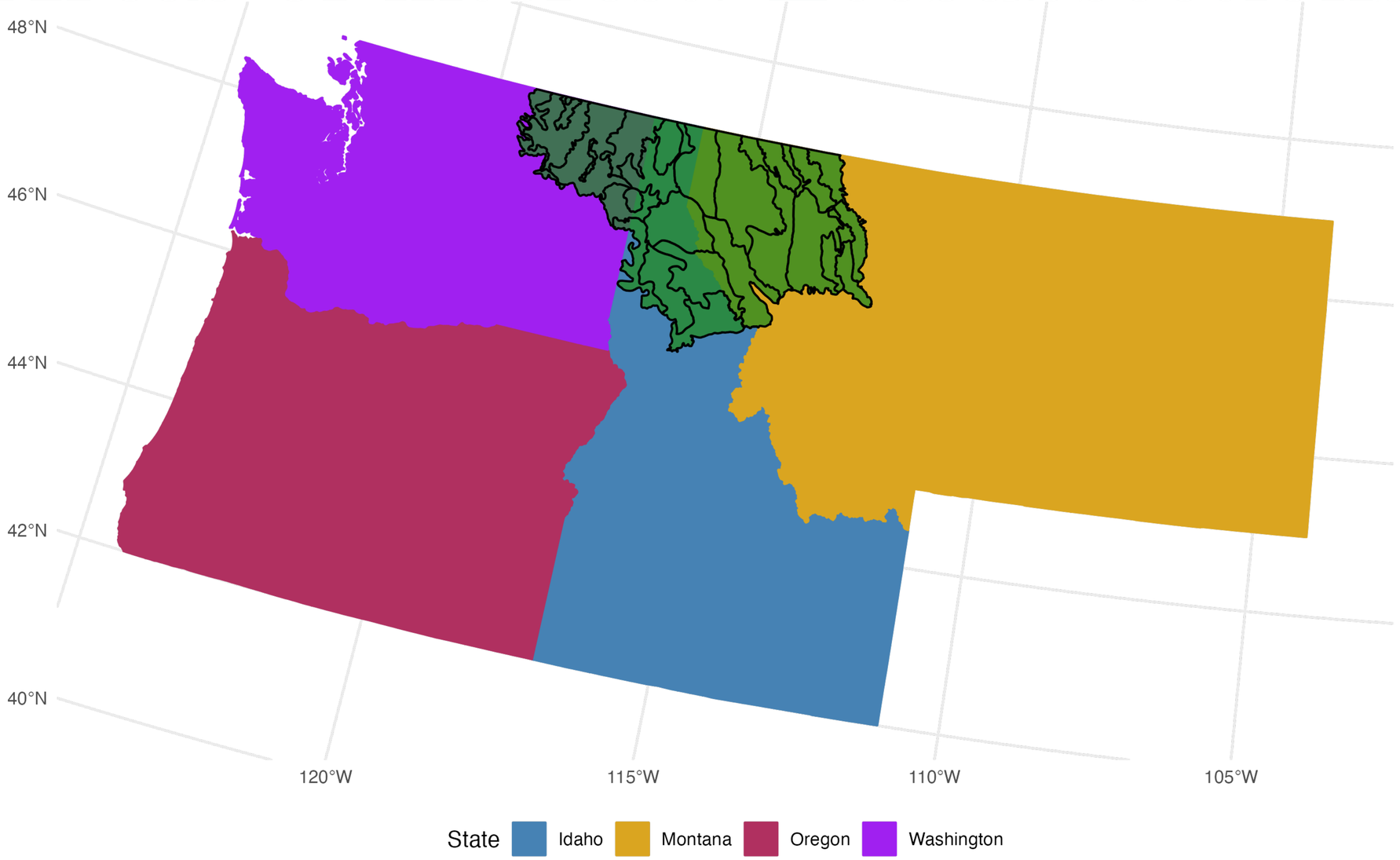
Step 3: rank each potential donor by distance in *auxiliary data* space

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

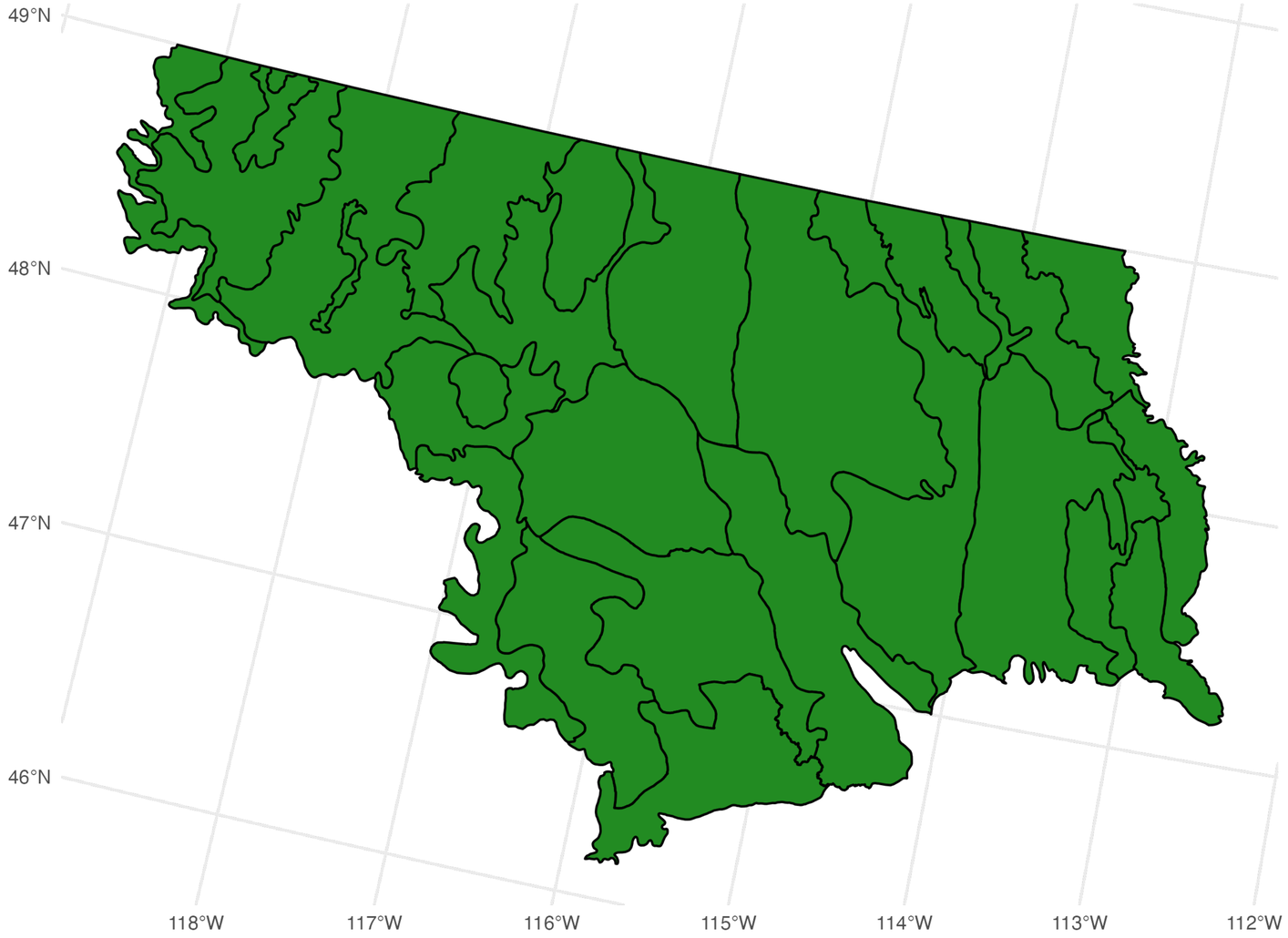
Step 5: repeat for each recipient to generate artificial population

Application of KBAABB

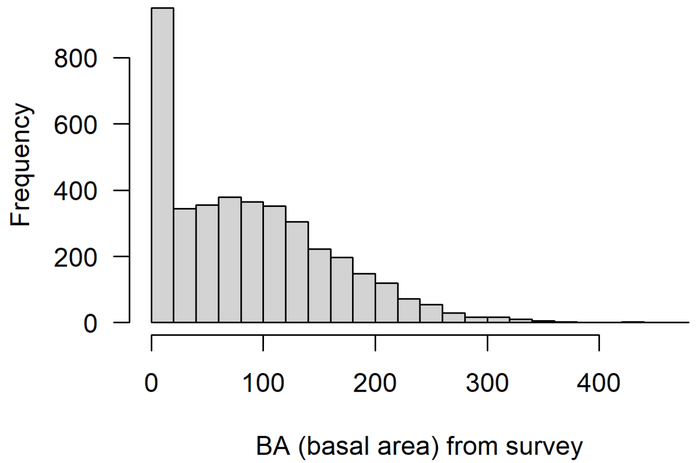
Study region: Ecoprovince M333, Areas of interest: Ecosubsections

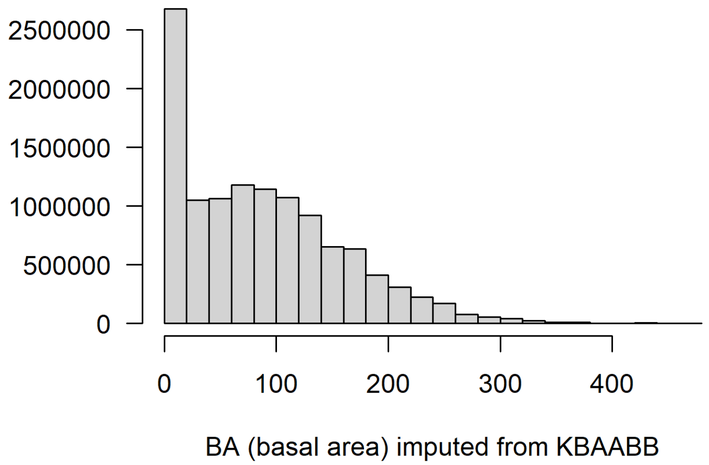


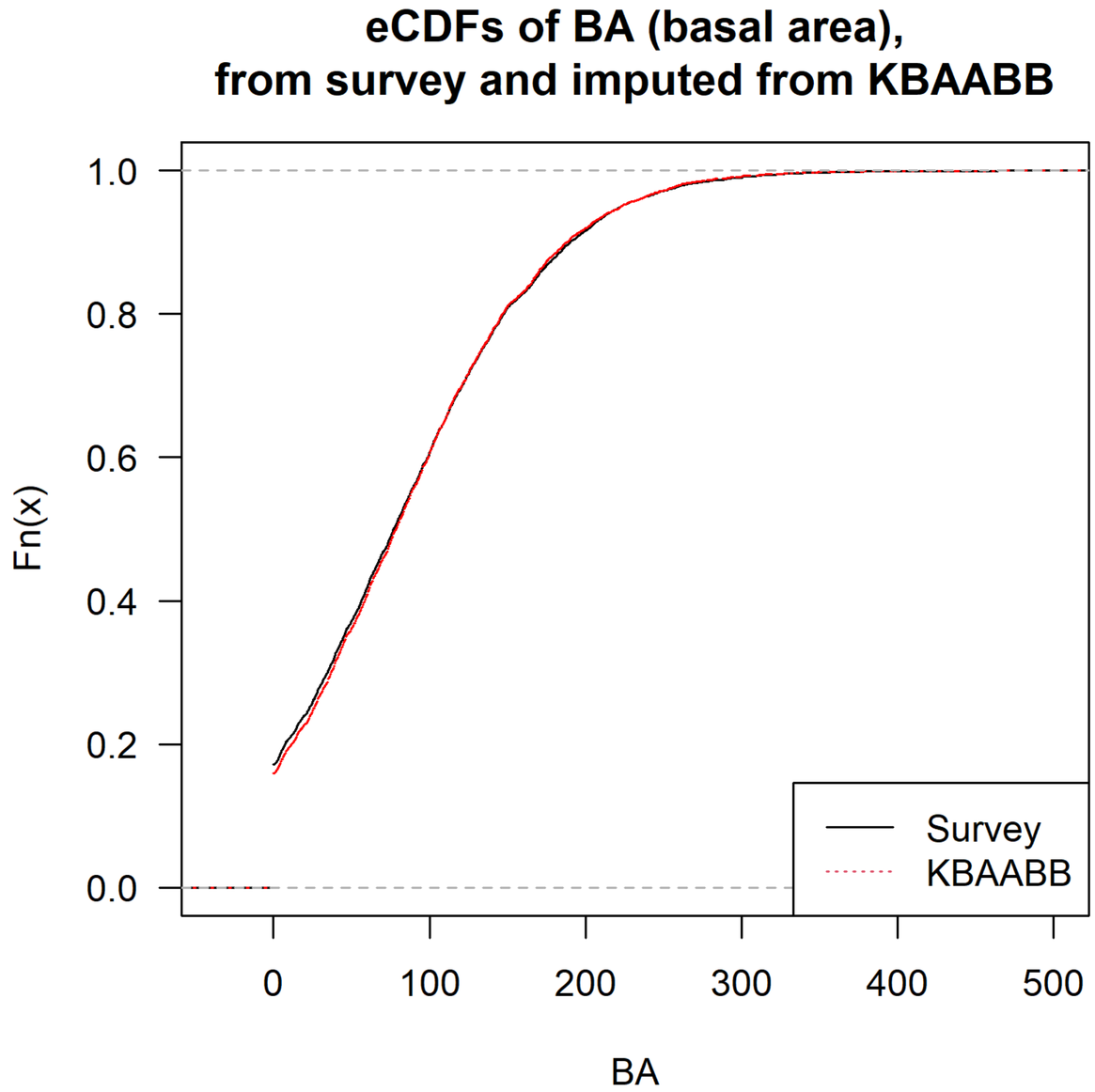
Study region: Ecoprovince M333, Areas of interest: Ecosubsections



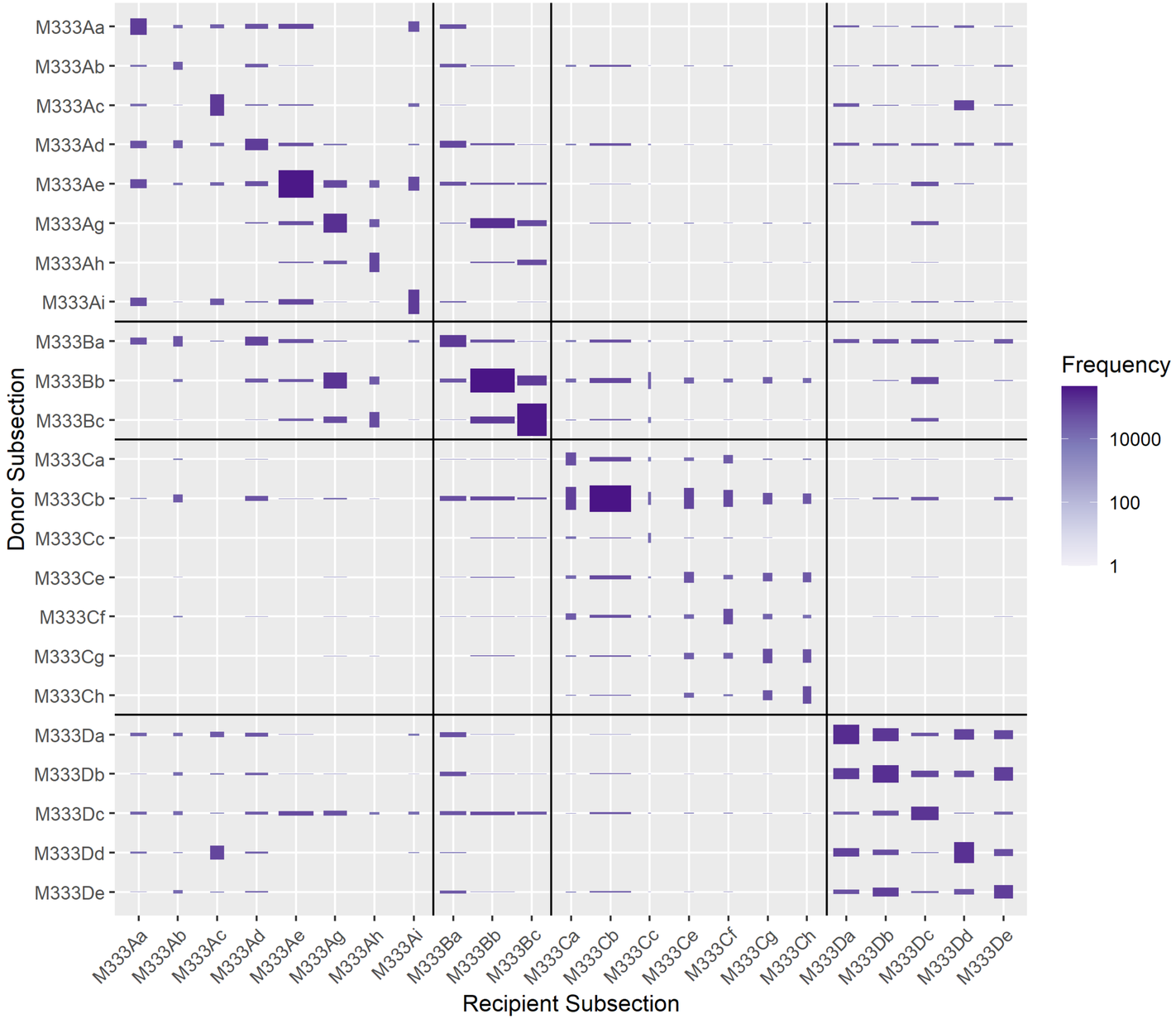
KBAABB Population Characteristics



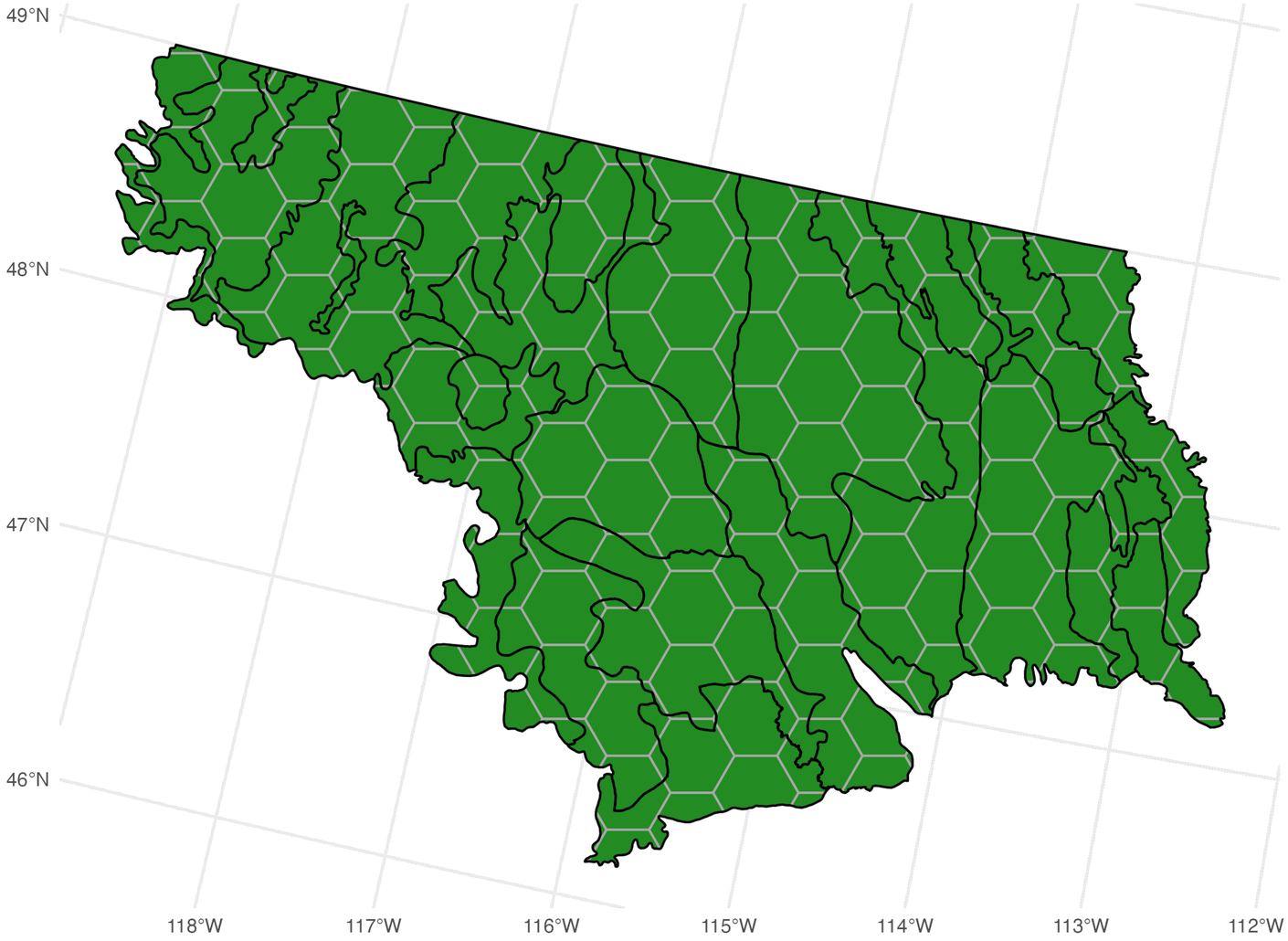




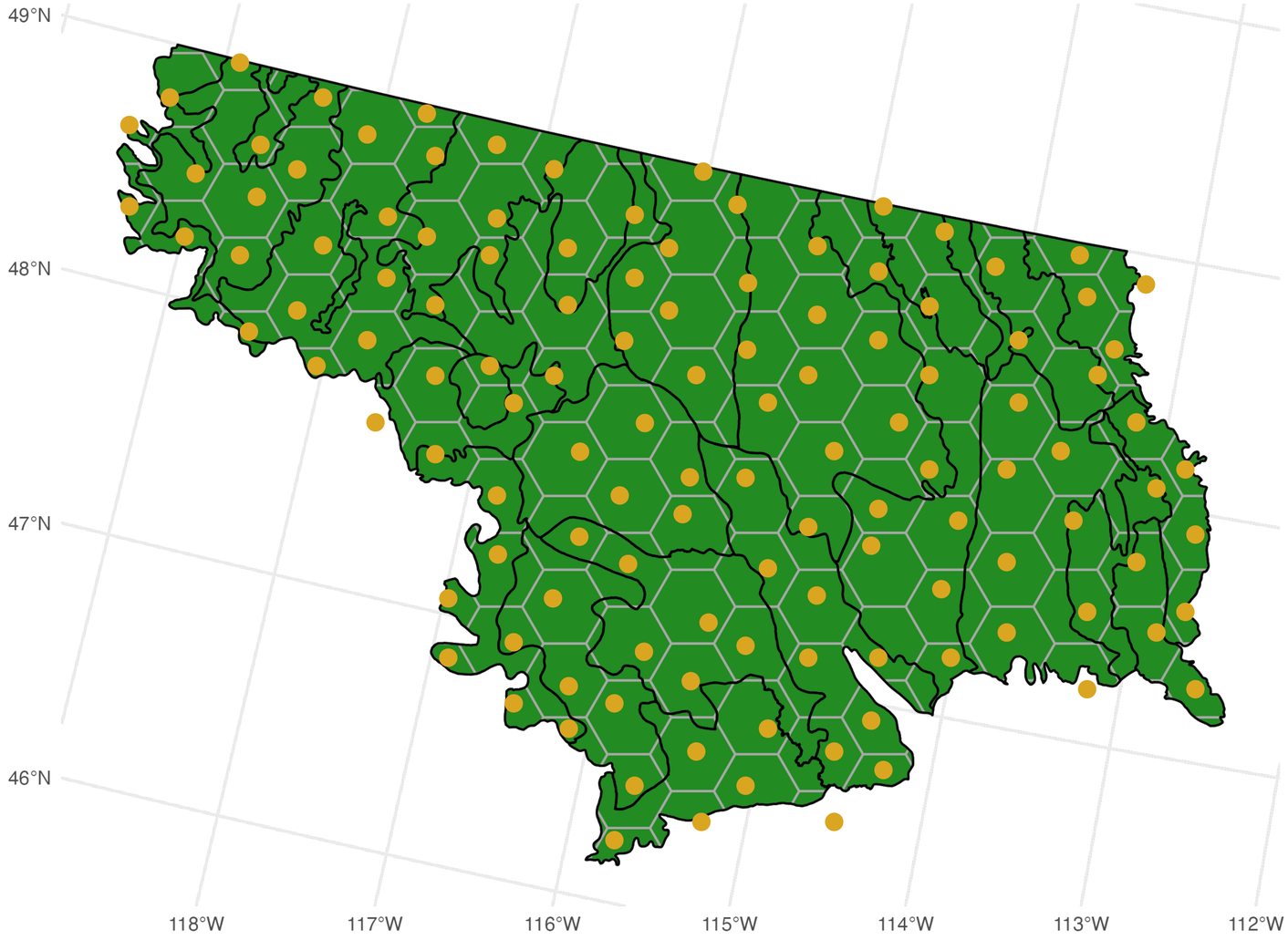
KBAABB Population Characteristics



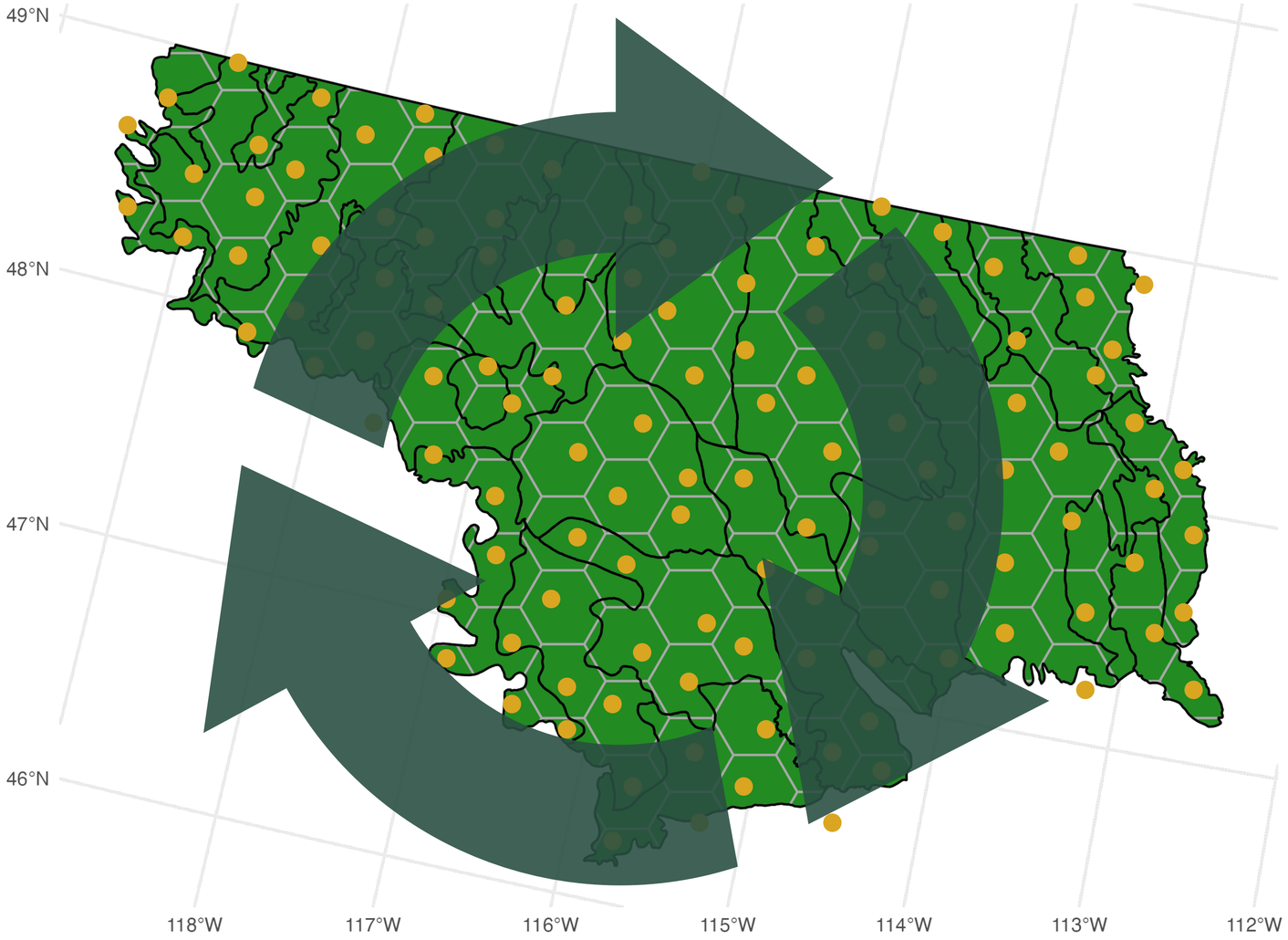
Sampling from KBAABB Population: Overlay FIA's Hexagonal Grid



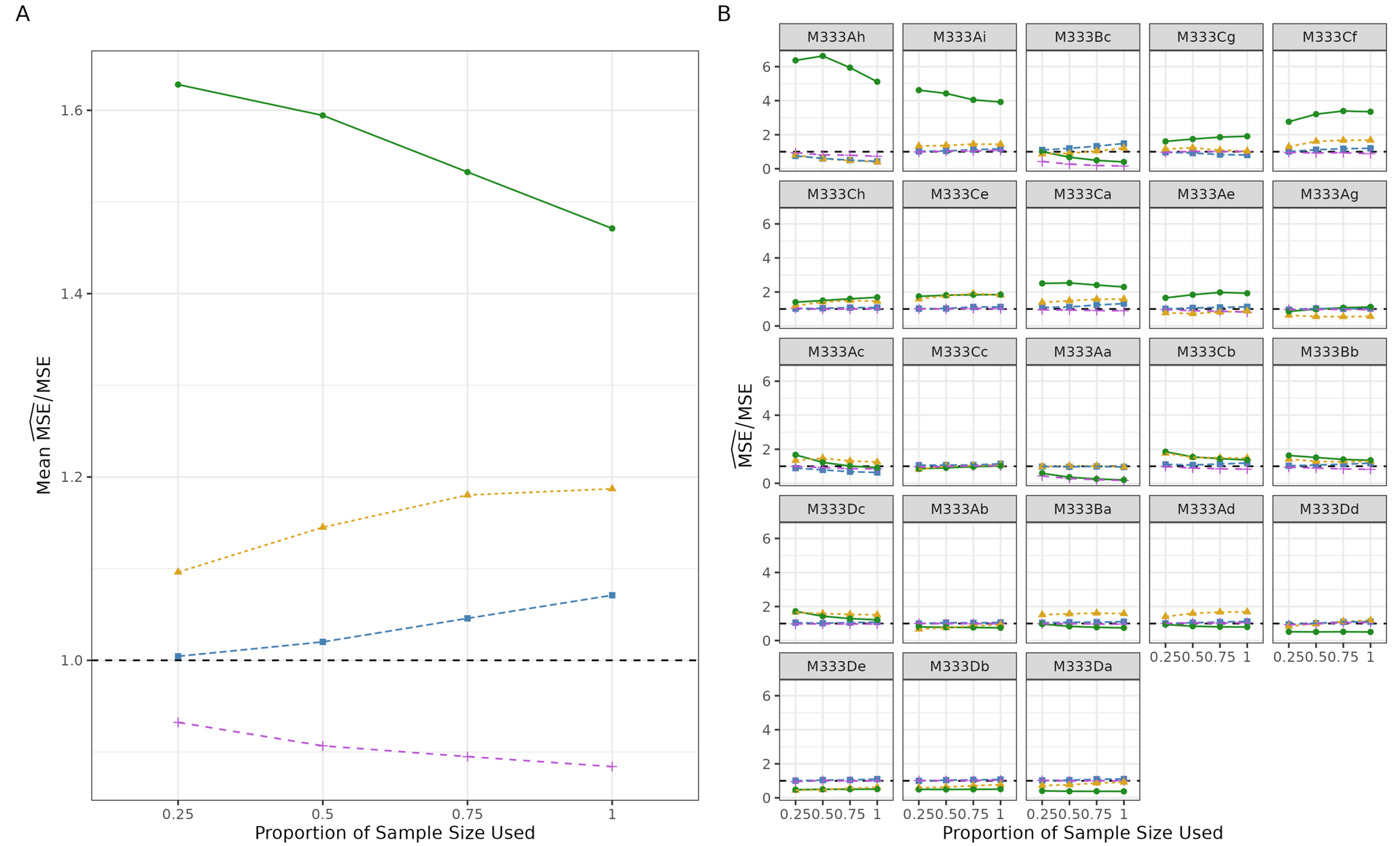
Sampling from KBAABB Population: Randomly Select 90m Pixel Within Each Hex



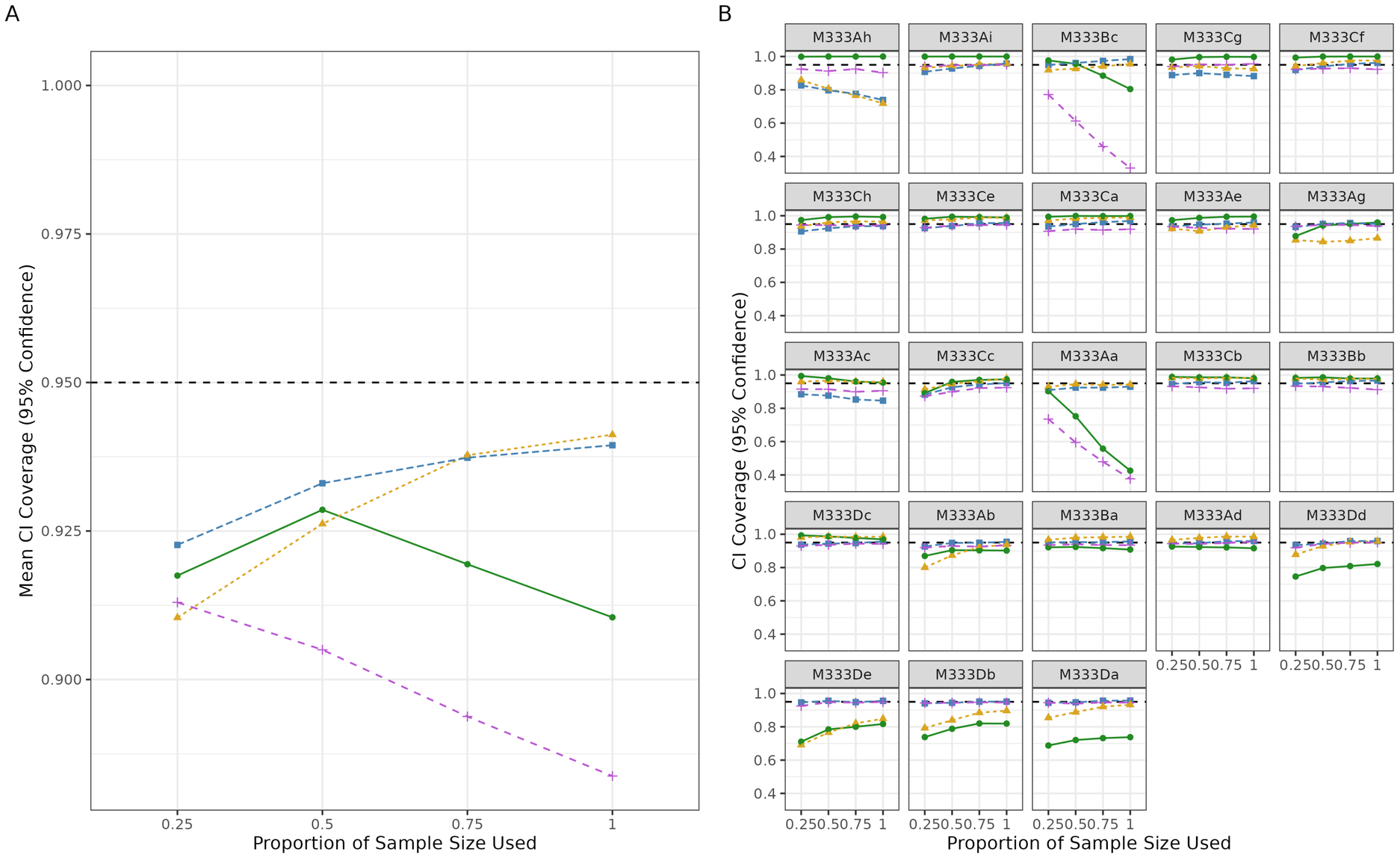
Sampling from KBAABB Population: Repeat Pixel Selection Many Times



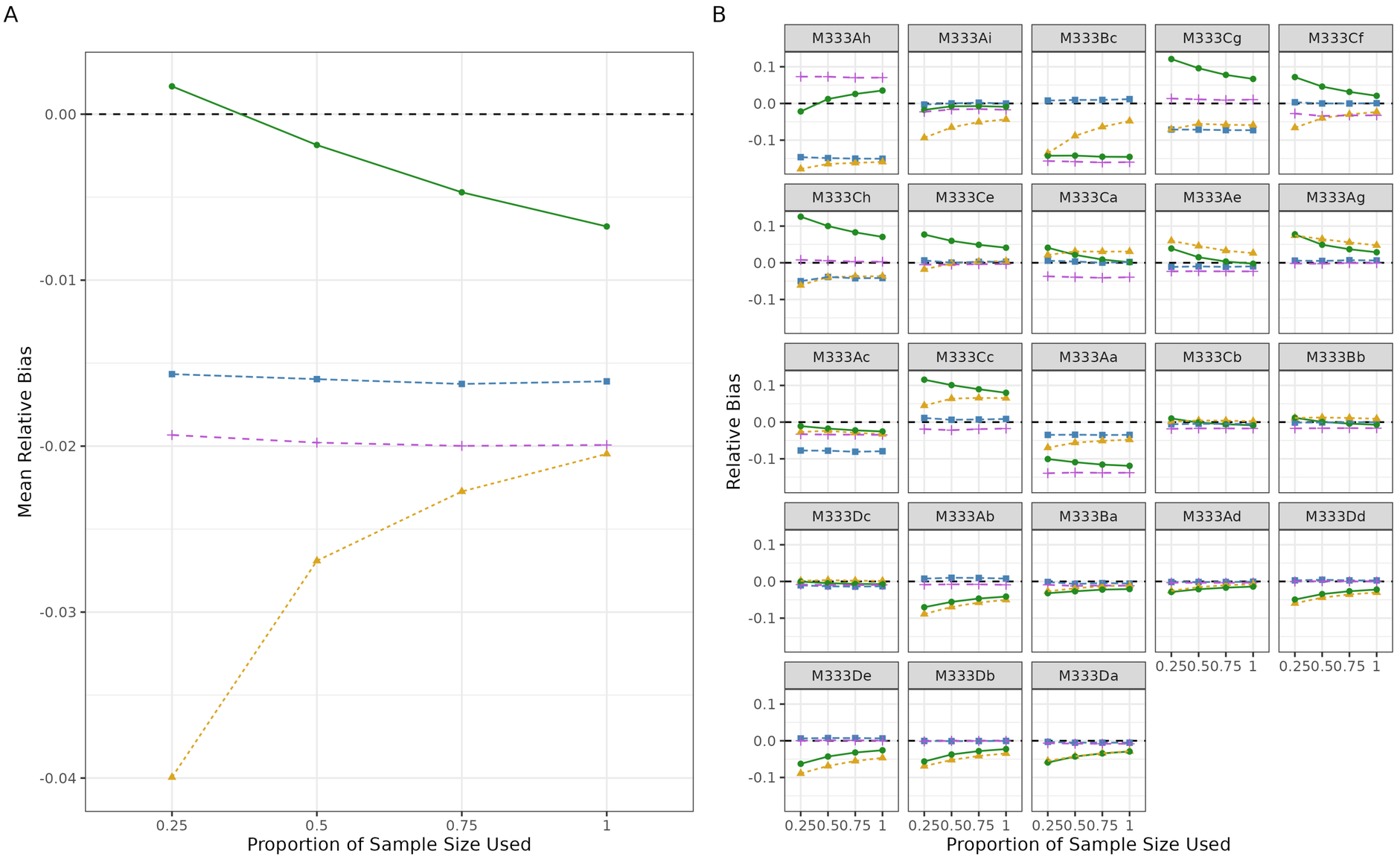
Assessing Estimators on KBAABB Samples



Assessing Estimators on KBAABB Samples



Assessing Estimators on KBAABB Samples



`kbaabb` R Package[5]

- 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")
```

National Scale KBAABB

What's Next? National KBAABB

We are planning a national application of KBAABB, with improvements to current methodology and extensive testing of the artificial population

+

Image credit: USFS Forest Atlas

National KBAABB: Why?

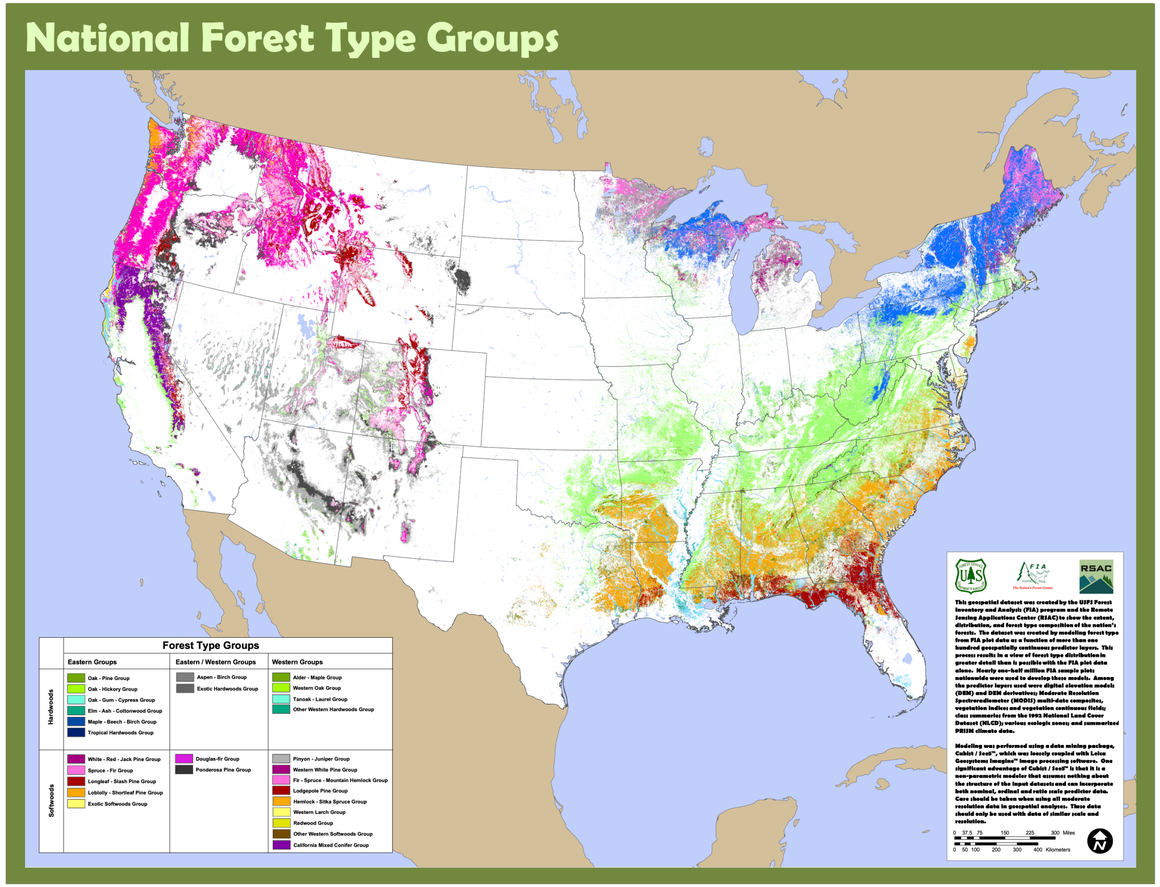
National KBAABB allows for comparison of small area estimators across scales (e.g. counties, national forests, recently burned areas, etc.) and regions of the county.

+

Image credit: USFS Forest Atlas

National KBAABB: How?

At a national scale, imputation becomes more difficult. We plan to stratify by variables such as forest type group and possibly include spatial dependence to ensure a sensible population.



National KBAABB: Goals

- With our national artificial population, we hope to assess a wide variety of novel small area estimators, including those funded by PSAE.

National KBAABB: Goals

- With our national artificial population, we hope to assess a wide variety of novel small area estimators, including those funded by PSAE.
- Further, we plan to provide the artificial population (and code to reproduce the population) as a raster layer for future estimator comparison and other endeavors.

National KBAABB: Goals

- With our national artificial population, we hope to assess a wide variety of novel small area estimators, including those funded by PSAE.
- Further, we plan to provide the artificial population (and code to reproduce the population) as a raster layer for future estimator comparison and other endeavors.
- We hope this national layer and simulation study helps the USFS FIA Program in best understanding when and where to use particular small area estimators.

Thank You! Questions?

References

1. Wieczorek, Jerzy A. et al. (2023). Assessing small area estimates via artificial populations from KBAABB: a kNN-based approximation to ABB. arXiv: 2306.15607 [stat.ME].

2. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016 Jun;4(11):218. doi: 10.21037/atm.2016.03.37. PMID: 27386492; PMCID: PMC4916348.

3. Andridge RR, Little RJ. A Review of Hot Deck Imputation for Survey Non-response. Int Stat Rev. 2010 Apr;78(1):40-64. doi: 10.1111/j.1751-5823.2010.00103.x. PMID: 21743766; PMCID: PMC3130338.

4. Rubin, D. B., Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81(394), 366–374. https://doi.org/10.1080/01621459.1986.10478280.

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