How do we assess the performance of our small area estimators?

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

SAE 2022: Small Area Estimation, Surveys and Data Science
University of Maryland, College Park, USA
May 23, 2022

[1] Colby College, [2] Harvard University, [3] U.S. Forest Service

(more concretely)

How do we assess the performance of our small area estimators… when we have *complete* unit-level auxiliary data?

Context

About FIA

- FIA: the **Forest Inventory and Analysis** Program of the U.S. Forest Service
- Mission: keep a current, comprehensive inventory of US forest resources
- Survey data: Ground crews visit samples of forested areas to take measurements of timber supply, forest health, etc.
- Auxiliary data: “Wall-to-wall” satellite measurements, elevation, temperature and precipitation records, etc.

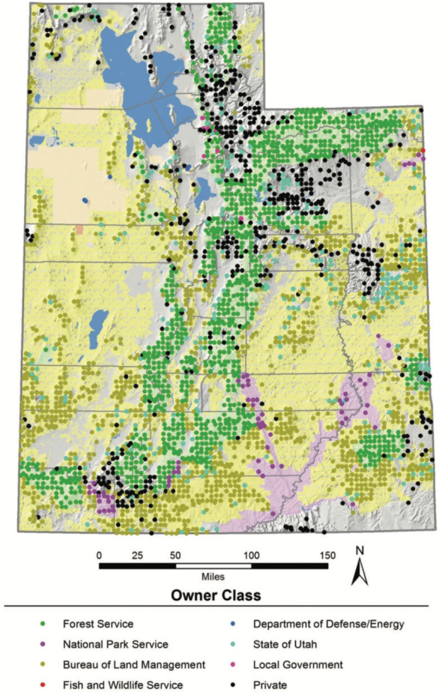
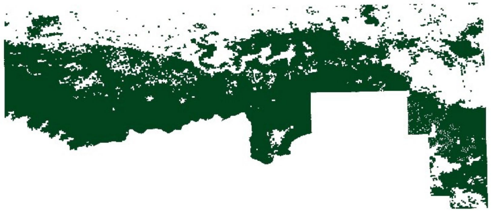


Figure 3—Distribution of inventory plots on forest land by owner class, Utah, 2003–2012. (Note: plot locations are approximate; some plots on private land were randomly swapped.)

About FIA

- FIA: the **Forest Inventory and Analysis** Program of the U.S. Forest Service
- Mission: keep a current, comprehensive inventory of US forest resources
- Survey data: Ground crews visit samples of forested areas to take measurements of timber supply, forest health, etc.
- Auxiliary data: "Wall-to-wall" satellite measurements, elevation, temperature and precipitation records, etc.



Forest or Non-Forest

Elevation

Why SAEs at FIA?

- Sampling design originally developed for estimates at US State level
- Interest in smaller geographies & subgroups, incl. ests for custom domains on request
- Wildfire risk assessment: very small domains and fast turn-around times
  - A wildfire in progress may contain only 1-2 survey observations, if any
  - Real-time estimates could help prioritize firefighter response (e.g. fire A risks damaging more commercially-productive timber supply than fire B)

USDA United States Department of Agriculture

Utah's Forest Resources, 2003–2012

Charles E. Werstak, Jr., John D. Shaw, Sara A. Goeking, Chris Witt, Jim Menlove, Michael T. Thompson, R. Justin DeRose, Michael C. Amacher, Sarah Jovan, Todd A. Morgan, Colin B. Sorenson, Steven W. Hayes, and Chelsea P. McIver

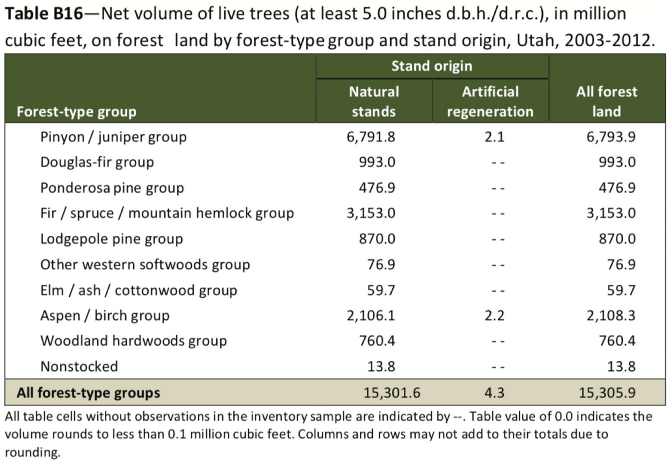


Table B16—Net volume of live trees (at least 5.0 inches d.b.h./d.r.c.), in million cubic feet, on forest land by forest-type group and stand origin, Utah, 2003-2012.

	Stand origin		
Forest-type group	Natural stands	Artificial regeneration	All forest land
Pinyon / juniper group	6,791.8	2.1	6,793.9
Douglas-fir group	993.0	- -	993.0
Ponderosa pine group	476.9	- -	476.9
Fir / spruce / mountain hemlock group	3,153.0	- -	3,153.0
Lodgepole pine group	870.0	- -	870.0
Other western softwoods group	76.9	- -	76.9
Elm / ash / cottonwood group	59.7	- -	59.7
Aspen / birch group	2,106.1	2.2	2,108.3
Woodland hardwoods group	760.4	- -	760.4
Nonstocked	13.8	- -	13.8
All forest-type groups	15,301.6	4.3	15,305.9

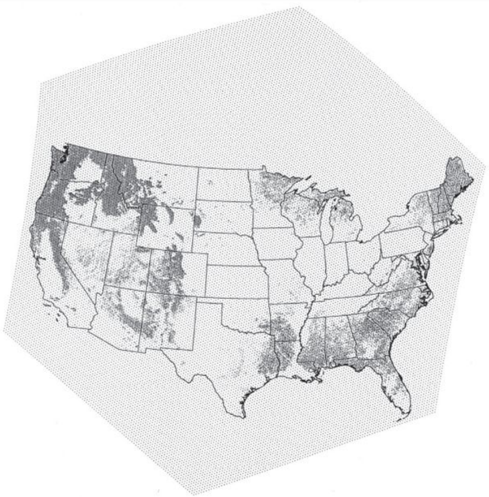
All table cells without observations in the inventory sample are indicated by --. Table value of 0.0 indicates the volume rounds to less than 0.1 million cubic feet. Columns and rows may not add to their totals due to rounding.

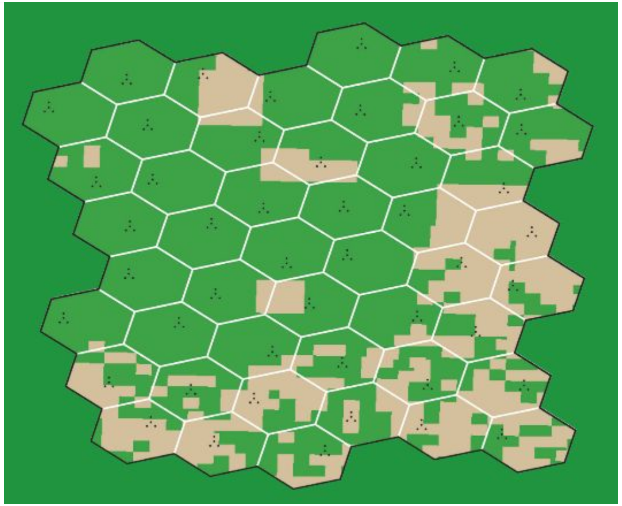
How can we assess new SAEs?

- Dorfman (2018), “Towards a Routine External Evaluation Protocol for SAE,” lists many approaches
  - **Supplementary validation sample** –
    requires advance planning and $$$
  - **Cross-validation** and checking **model-fit diagnostics** –
    requires having enough survey data in each domain
- In today’s talk, we focus on 2 approaches to **design-based simulations** when we have (essentially) **complete unit-level auxiliary data**, as FIA does
  - Generate data that mimics our real situation, and evaluate SAEs for specific scenarios

Overview of FIA survey design

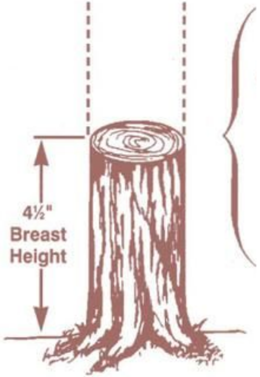
- Sampling frame: cover USA by a tessellation of ~6000-acre hexagons
- Sample: FIA chose one location at random in each hexagon, and field crews have defined a permanent **sample plot** at each chosen location



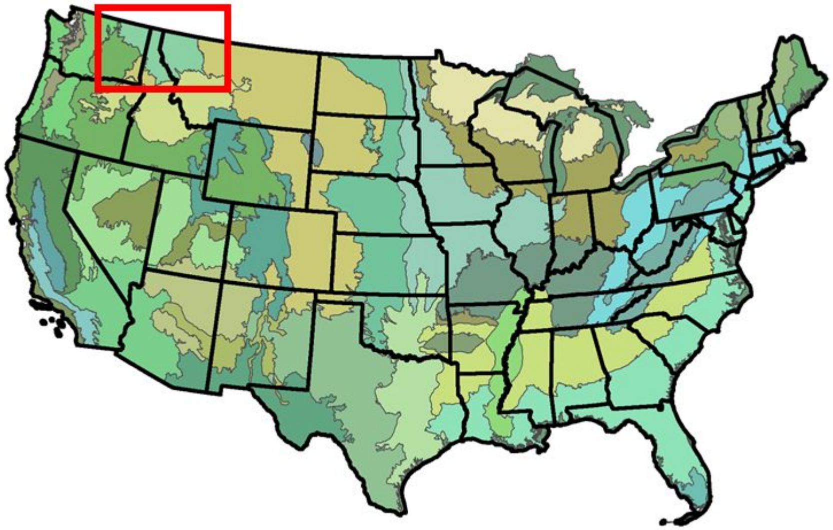


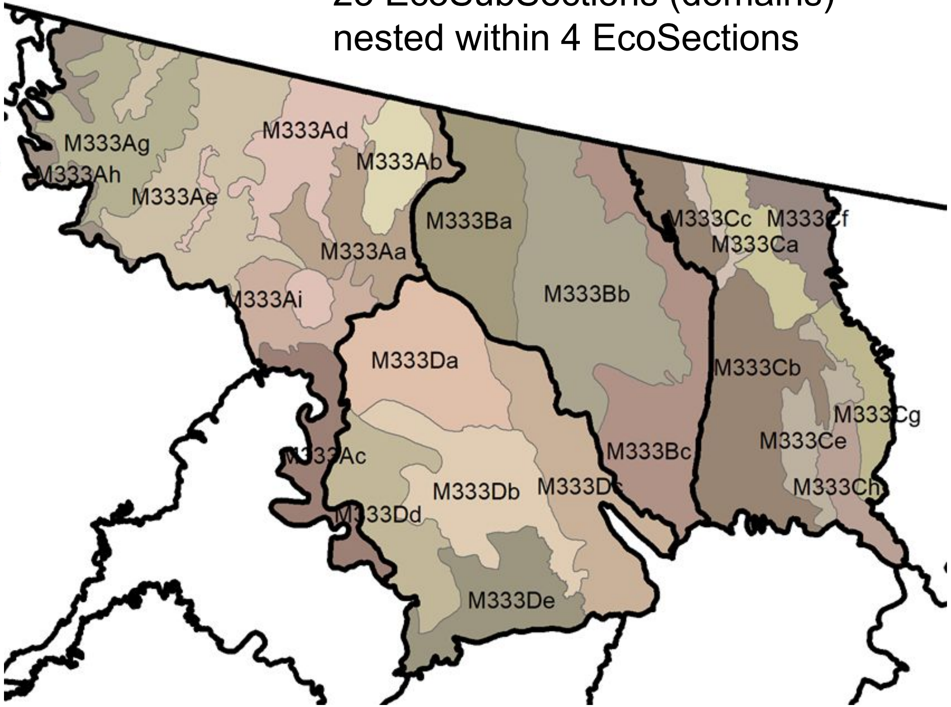
Overview of FIA survey design

- Measurements: field crews record **attributes** (number of trees per acre, total basal area of trees, etc.)
- Panel design: revisit all forested plots every 5-10 years (varies by State)
- Estimation: for the Interior West, post-stratify by forest/non-forest class (from a satellite-based layer)

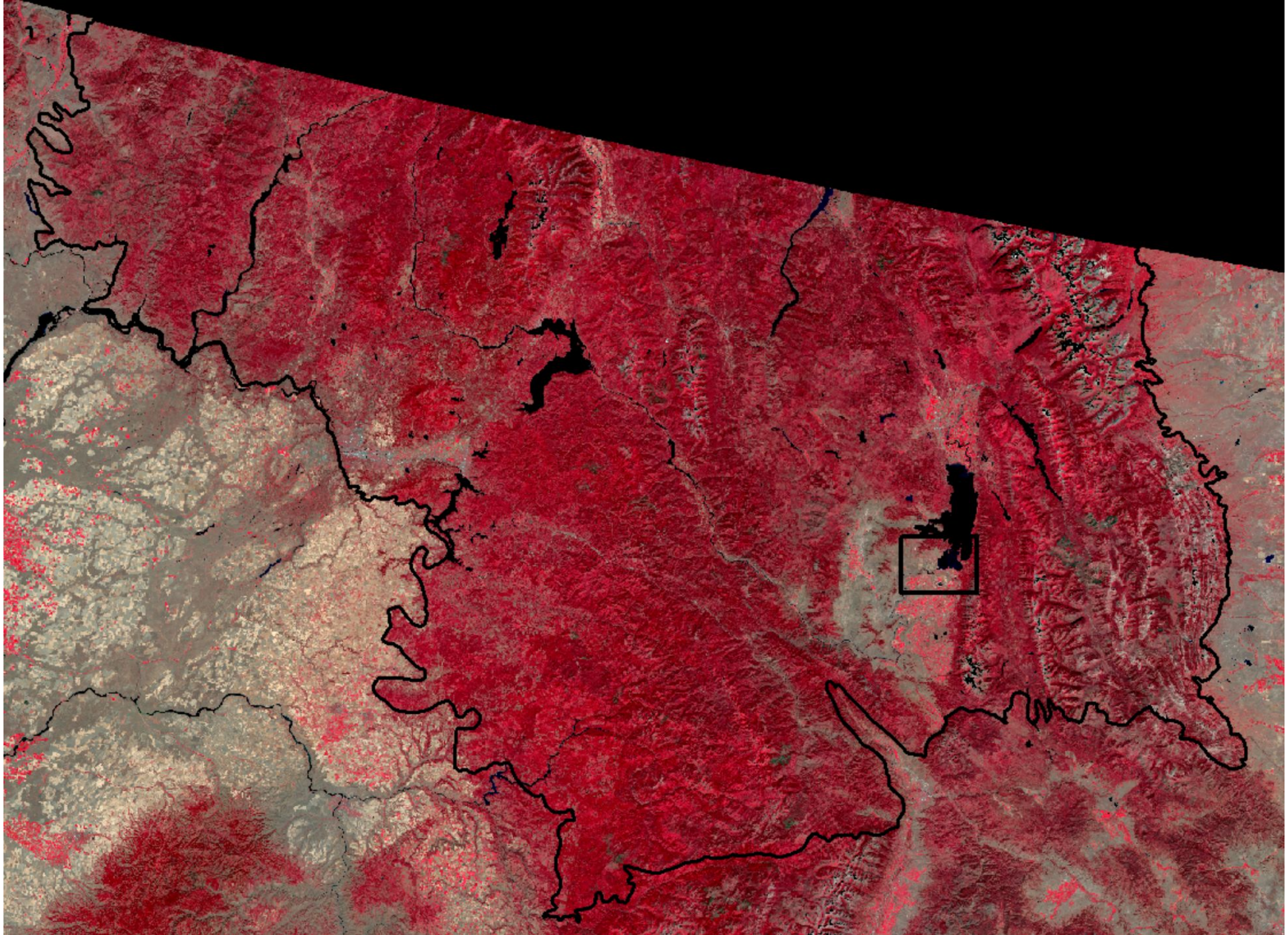


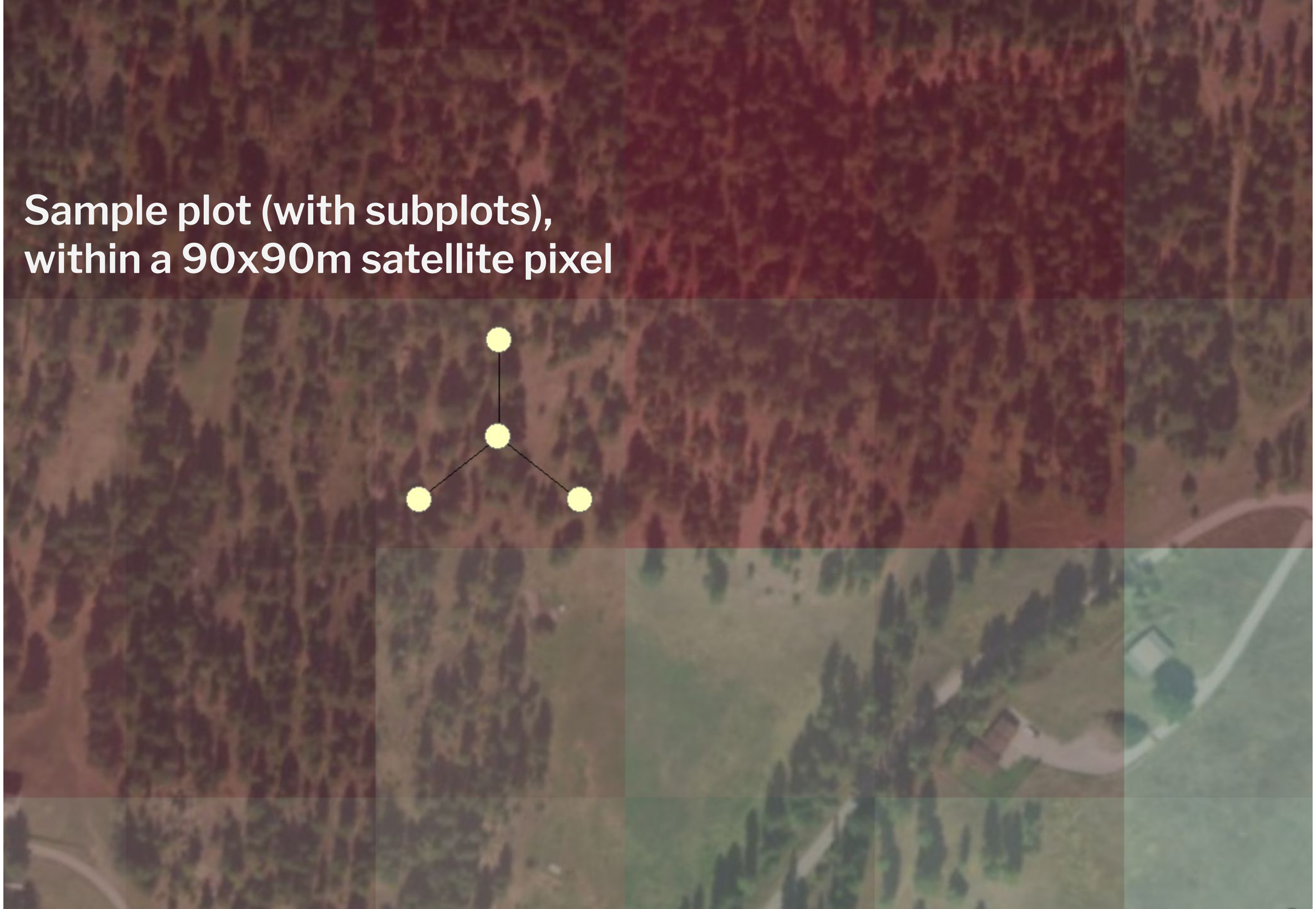
Our region of interest: EcoProvince M333





23 EcoSubSections (domains) nested within 4 EcoSections

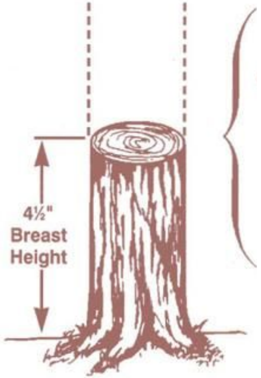




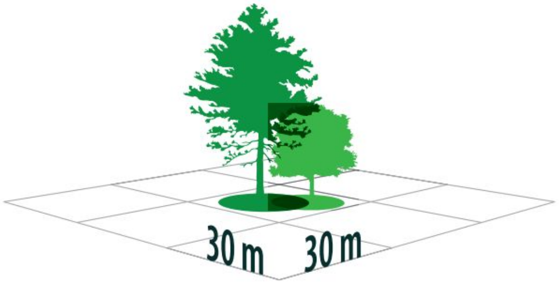
Sample plot (with subplots), within a 90x90m satellite pixel

Summary of available data

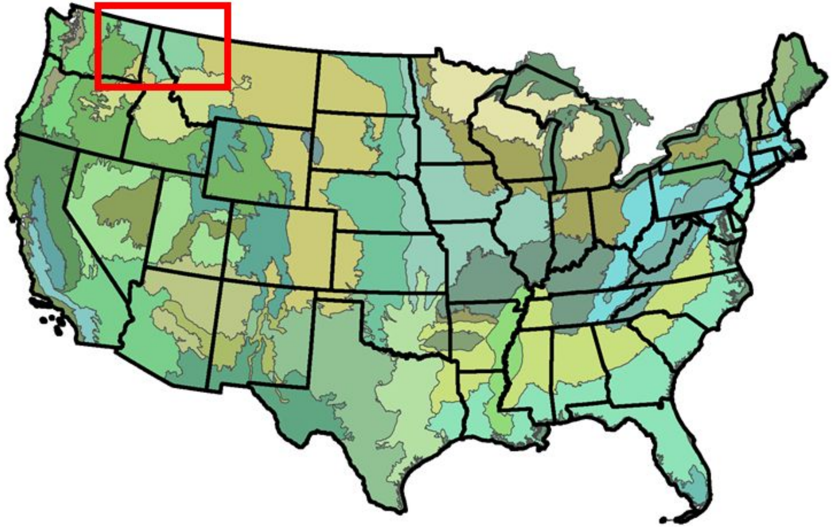
- Y = surveys (from field crews) = “plot-level data”
- X = auxiliary data (from satellites etc.) = “pixel-level data”
  - Pixel size chosen to approx. match sample plots from surveys
  - Landsat: remote sensing satellite data
  - Climate records: mean annual precipitation, temperature, etc.
  - Topographic records: elevation, eastness/northness, etc.
- Used in today’s simulation-study examples:
  - Y = `BA` (basal area) from surveys, or
    “Y” = `EVI` (enhanced vegetation index) from Landsat
  - X = `tcc` (tree canopy cover),
    `def` (mean annual climatic water deficit),
    `tri` (terrain ruggedness index),
    `tnt` (tree/non-tree classification)



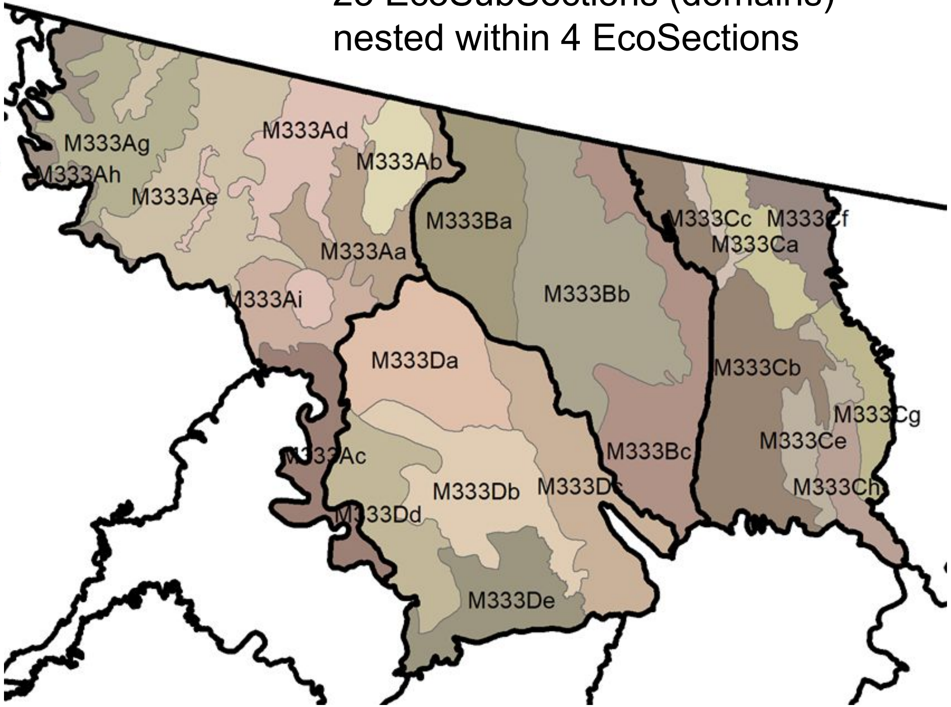
EXAMPLE
TCC Value = 65% of 30 meter pixel or cell



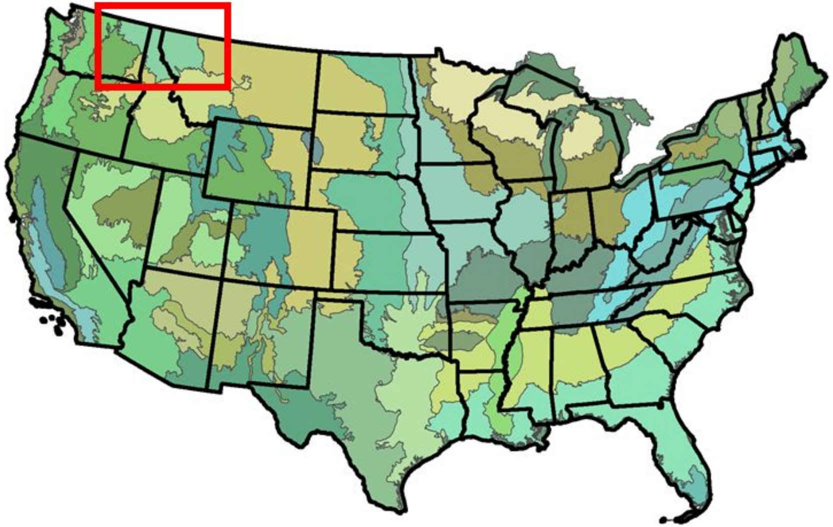
Our region of interest: EcoProvince M333



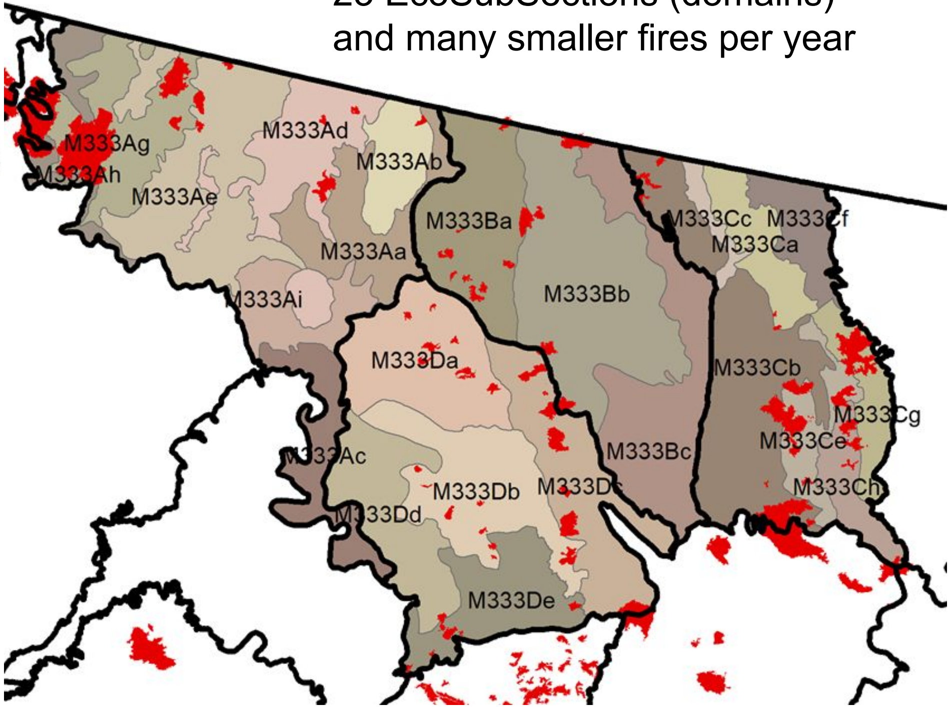
23 EcoSubSections (domains) nested within 4 EcoSections



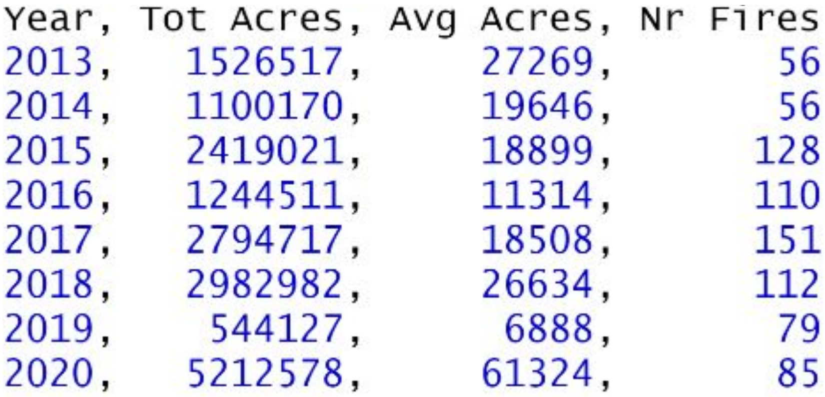
Our region of interest: EcoProvince M333 (with fires)



23 EcoSubSections (domains)
and many smaller fires per year

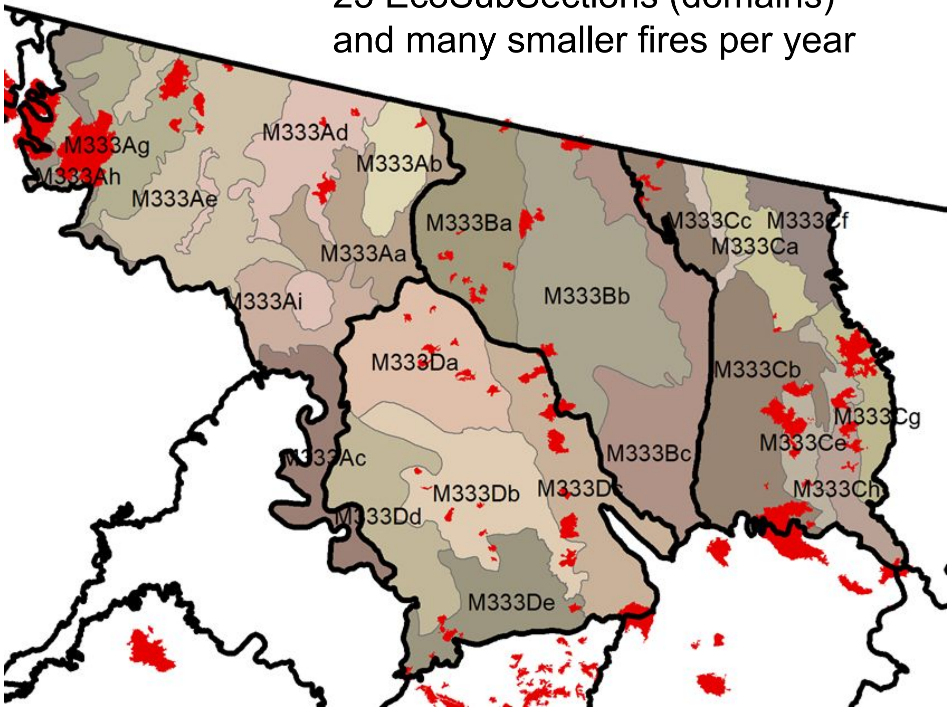


Our region of interest: EcoProvince M333 (with fires)



```
Year, Tot Acres, Avg Acres, Nr Fires
2013,   1526517,     27269,       56
2014,   1100170,     19646,       56
2015,   2419021,     18899,      128
2016,   1244511,     11314,      110
2017,   2794717,     18508,      151
2018,   2982982,     26634,      112
2019,    544127,      6888,       79
2020,   5212578,     61324,       85
```

23 EcoSubSections (domains)
and many smaller fires per year



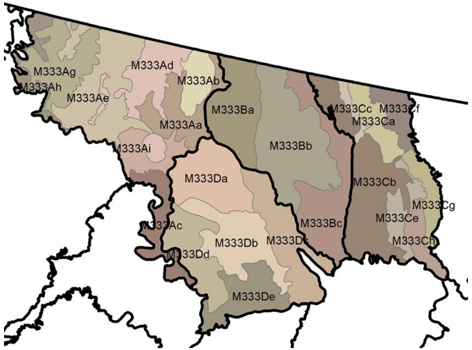
Design-based sim. populations

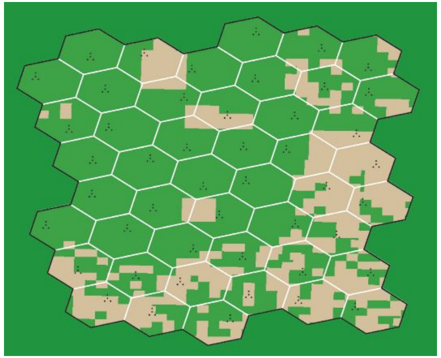
(more concretely)

How do we assess the performance of our small area estimators… when we have *complete* unit-level auxiliary data X… but relatively *few* survey responses Y?

Two approaches to design-based sim pops

- In both cases, we start with wall-to-wall coverage of auxiliary data
- We mimic the real sampling design (partition land into 6000-acre hexagons and sample one pixel in each hexagon) and repeat 2500 times





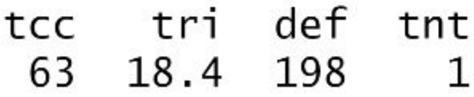
- Then we choose the response variables in 2 ways…

Two approaches to design-based sim pops

- We choose the response variables in 2 ways…
  - Auxiliary “response”: choose one X variable to treat as the response Y
    - Choose a variable that isn’t “too easy” to predict from other Xs!
    - e.g. Y = Enhanced Vegetation Index (nonlinear fn. of satellite bands)
  - Survey response: using a form of “*k*NN hot deck” imputation,
    combine real survey responses Y with simulated samples of auxiliary X
    - This is NOT imputation for creating estimates to publish!
      Just to create a simulated pop we can use to evaluate models

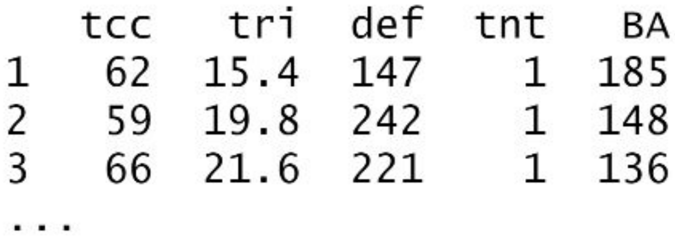
*k*NN hot deck, illustrated

Example row of recipient data



```
tcc   tri  def  tnt
 63  18.4  198    1
```

Example of nearest-neighbor donor rows

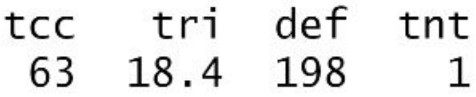


```
   tcc   tri  def  tnt   BA
1   62  15.4  147    1  185
2   59  19.8  242    1  148
3   66  21.6  221    1  136
...
```

1. Find *k* best matches in donor dataset

*k*NN hot deck, illustrated

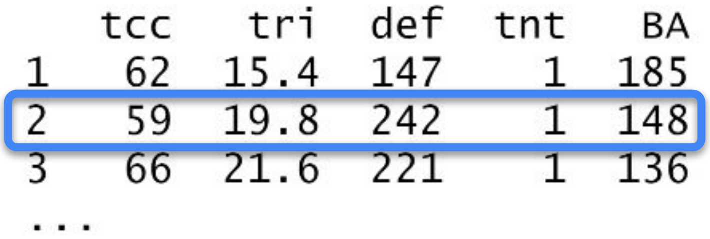
Example row of recipient data



```
tcc   tri  def  tnt
 63  18.4  198    1
```

1. Find *k* best matches in donor dataset
2. Choose one at random

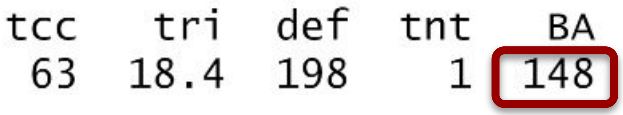
Example of nearest-neighbor donor rows



```
   tcc   tri  def  tnt   BA
1   62  15.4  147    1  185
2   59  19.8  242    1  148
3   66  21.6  221    1  136
...
```

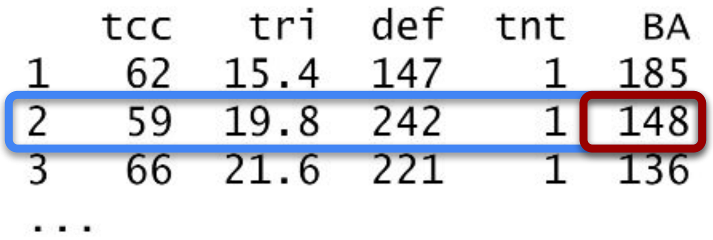
*k*NN hot deck, illustrated

Example row of recipient data



```
tcc    tri   def   tnt    BA
 63   18.4   198     1   148
```

Example of nearest-neighbor donor rows



```
     tcc    tri   def   tnt    BA
1     62   15.4   147     1   185
2     59   19.8   242     1   148
3     66   21.6   221     1   136
...
```

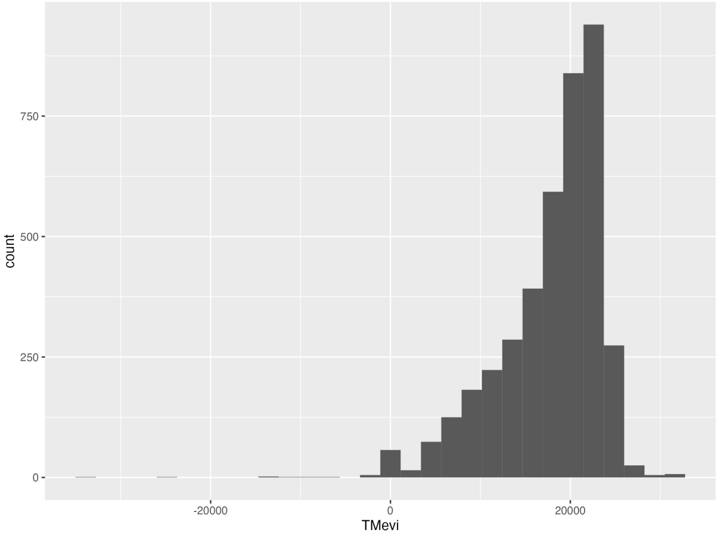
1. Find *k* best matches in donor dataset
2. Choose one at random
3. Impute its Y-value to the recipient row

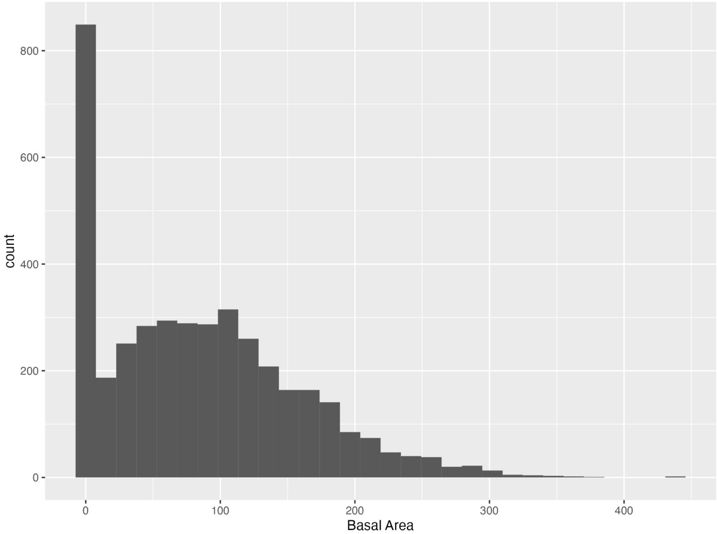
Details of *k*NN hot deck

- Choose a set of auxiliary X variables to use for matching
  - Not too correlated nor conceptually equivalent, to avoid double-counting
  - Transform Xs if necessary (e.g. if so skewed that their outliers never get used as matches)
  - Center and scale each X
- Stratify by tree / non-tree
- For each recipient (simulation) row, find its *k* nearest neighbors in the donor (real survey) dataset: sample plots whose associated pixels had most-similar Xs
  - For us, *k*=10 gave fairly realistic variability in outputs
  - We used Euclidean distance on the standardized variables
- Choose one of these *k* NNs uniformly at random (“hot deck”)
- Impute its survey Y-values to the simulated sample row
- “True” pop. means in each domain:
  average-across-2500-reps of domain means

Alternatives to *k*NN hot deck

- Why not just always use the simple auxiliary-as-“response” approach?
  - There may be no auxiliary X whose distribution is realistically similar to Y of interest
  - The most realistic X may be a function of other Xs and hence “too easy” to model



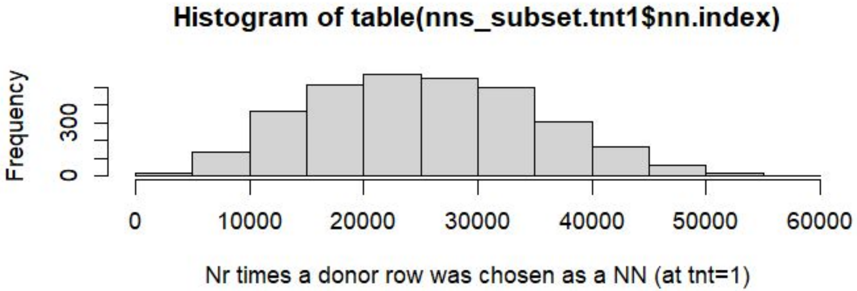


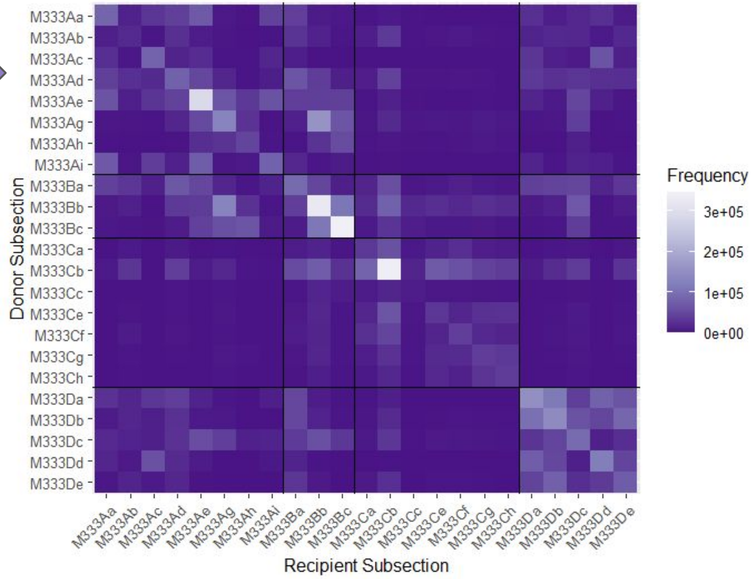
Alternatives to *k*NN hot deck

- Why not use a simpler imputation approach?
  - Simple random distributions such as Y ~ N(0,1) won’t reflect our real scenarios
  - If we fit a regression model and generated Ys from it, our sim study would be over-optimistic when evaluating SAEs built on similar regression models
- *k*NN matching with random hot deck should keep realistic associations between Ys and their Xs, but avoids using the same model form as the SAE models we are evaluating

Quality checks on the *k*NN hot deck process

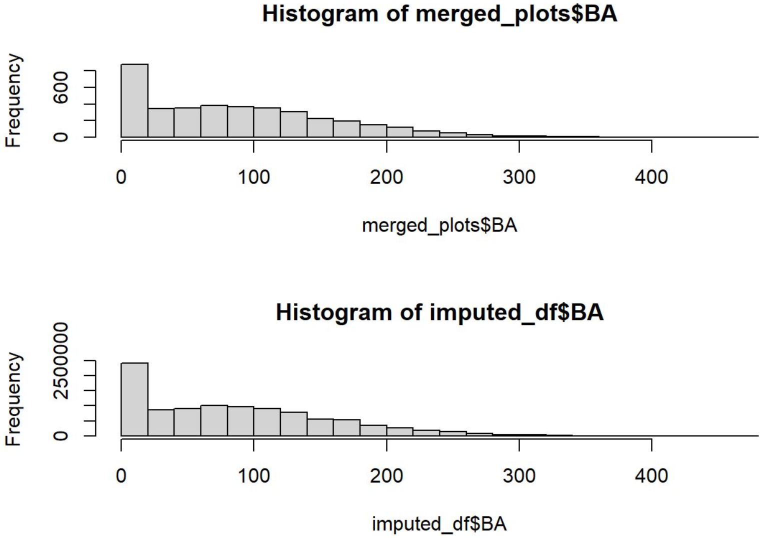
- Frequency of reuse: are we just reusing the same few donors over and over?
- Distribution of donors across domains: does donor nearly always come from same domain as recipient?





Quality checks on the *k*NN hot deck process

- Distributions of Ys: do histograms and paired scatterplots of Ys or Ys-vs-Xs look similar in simulated samples as in real survey dataset?



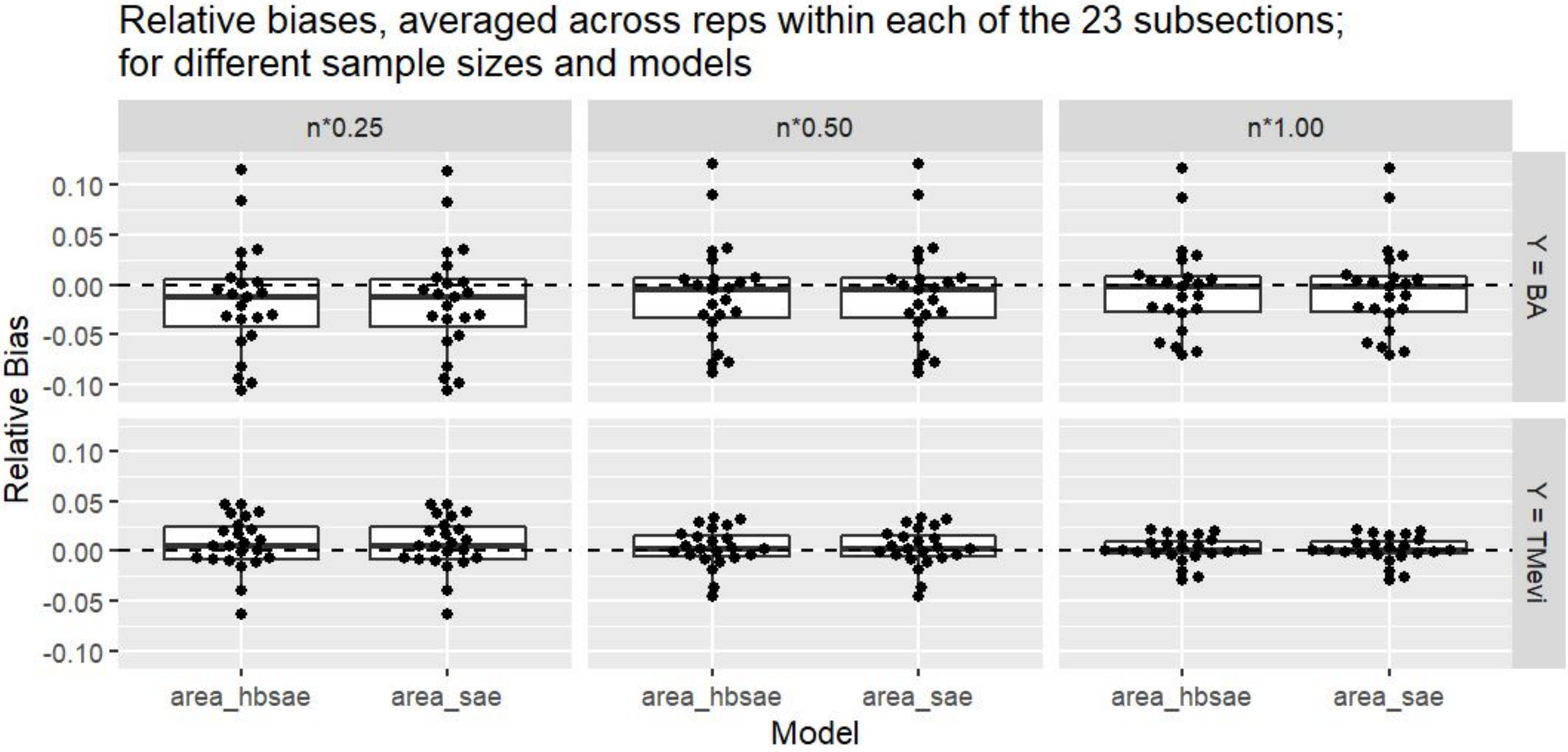
Proof-of-concept simulation study design

Proof-of-concept simulation study

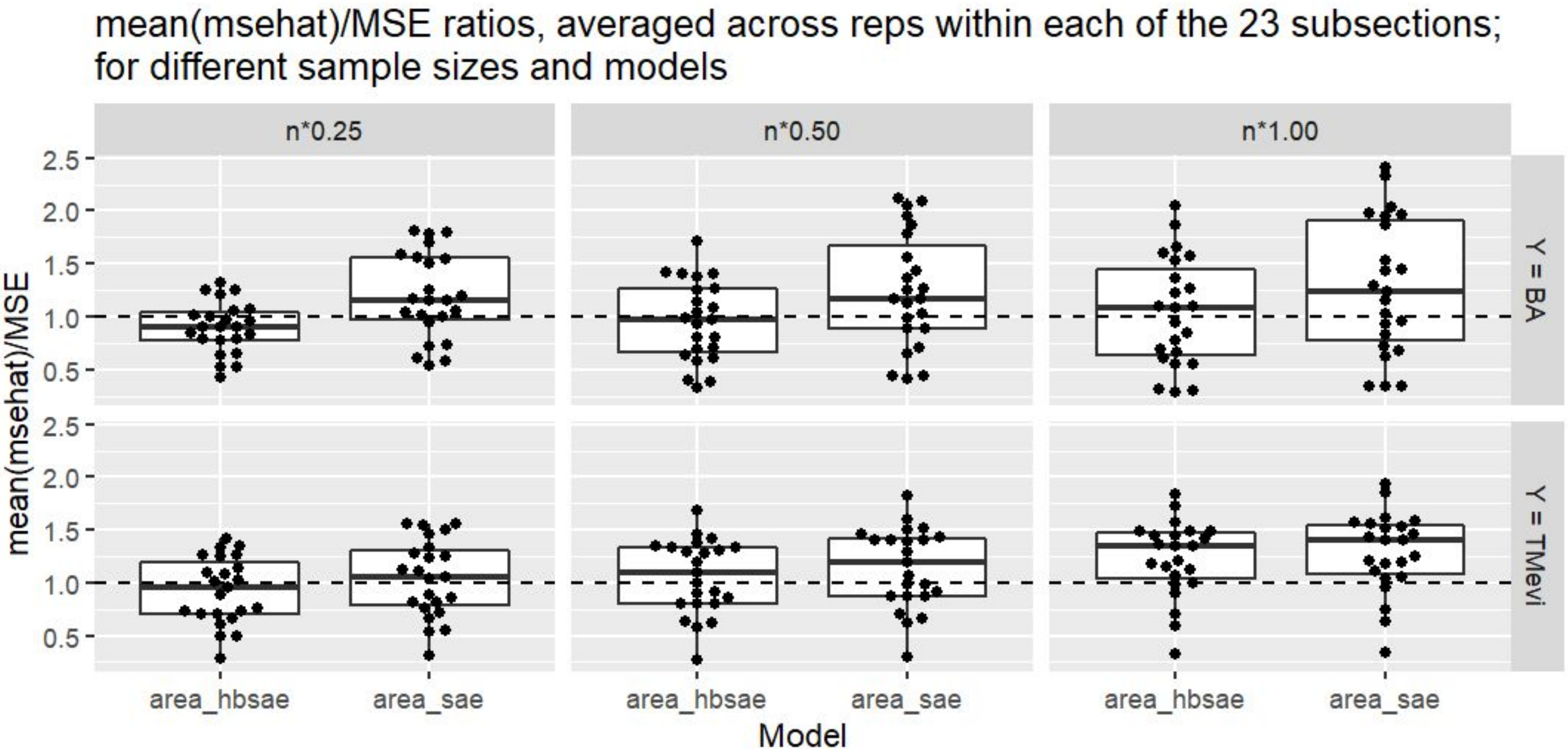
- Used one auxiliary response Y=EVI and one *k*NN hot deck response Y=BA
- Turned sample size “dial” to quarter, half, or original sizes per domain
- Fit two area-level Fay-Herriot models with REML, using:
    - `sae` R package (Molina & Marhuenda 2020), and
    - `hbsae` R package (Boonstra 2022)
- Calculated Relative Bias, MSE Ratios, and 95% CI Coverage, averaged across 2500 reps within each of 23 domains
    - Similar to evaluations in Wieczorek, Nugent, and Hawala (2012)

- Within each response, effects of dial and model were as expected
- Across the two responses, results differed enough to justify having both sim-pop approaches in our toolkit

Relative bias, per domain

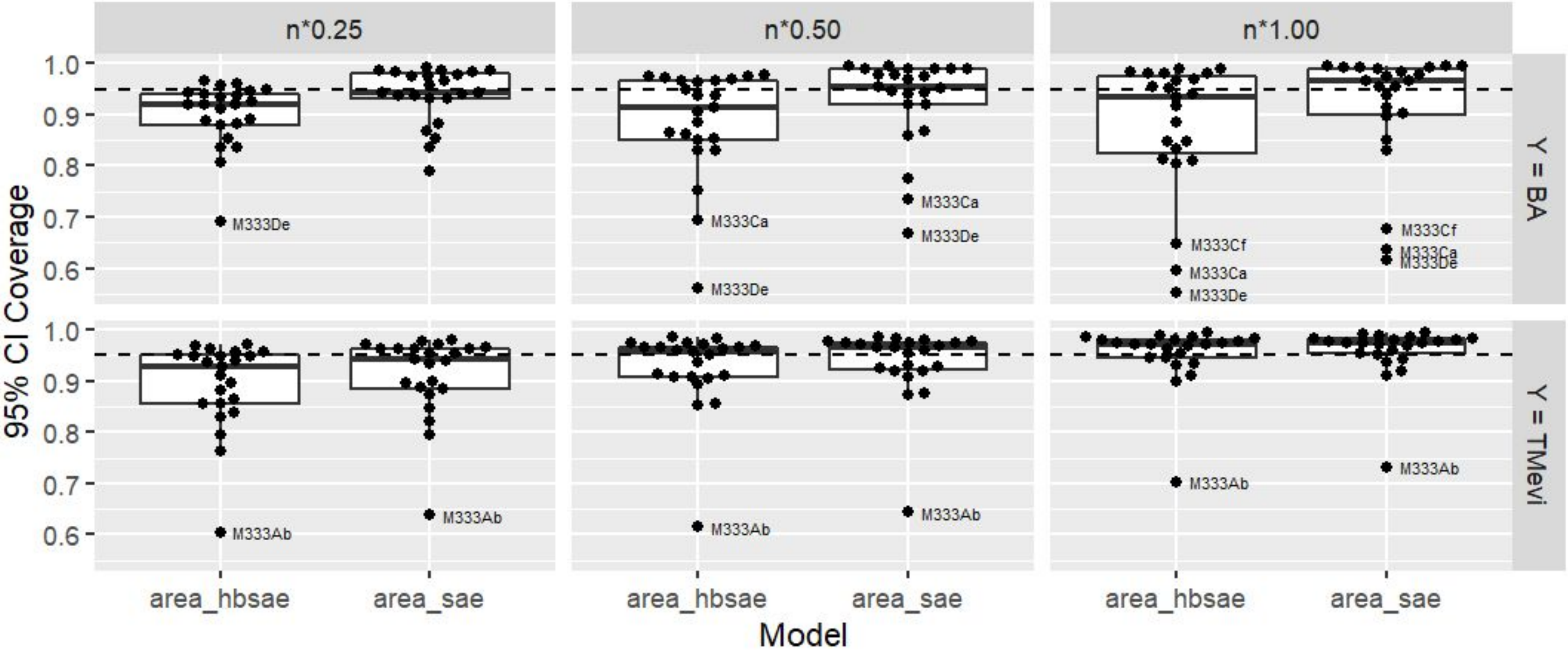


mean(MSEhat) / MSE, per domain



95% CI coverage, per domain

95% CI coverage, averaged across reps within each of the 23 subsections;
for different sample sizes and models



Proof-of-concept sim study’s conclusions

Again:

- Within each response, effects of dial and model were as expected
- Across the two responses, results differed enough to justify having both (auxiliary-“response” and *k*NN hot deck) approaches in our toolkit

Next steps

- Implement more dials, models, and quality metrics
  - Dials: reduced nr of domains? Fit one wildfire at a time vs several at once?
  - Models: compare with poststrat, GREG, unit-level, etc.
  - Metrics: Absolute relative bias; MSE; frequency of NAs; computational cost…
- Focus on dials related to wildfires, and give concrete guidance to users interested in wildfire SAEs
- Integrate into `FIESTA` R package to automate guidance around “which SAE model should I use?” for a wider range of non-expert users

Thank you

- We'd love feedback! What dials, quality metrics, etc. would you like to see? Quality checks on *k*NN hot deck? Other related design-based sims?
- Contact info
  - Speaker: Jerzy Wieczorek
    - jawieczo@colby.edu
  - To collaborate on SAE with FIA: Kelly McConville
    - kmcconville@fas.harvard.edu
- Links
  - National FIA website:
    - https://www.fia.fs.fed.us/
  - `FIESTA`: Forest Inventory Estimation and Analysis (R package)
    - https://usdaforestservice.github.io/FIESTA/
    - https://cran.r-project.org/web/packages/FIESTA/index.html

Supplemental slides

Sim study design

- Choose a sim pop (auxiliary response or *k*NN hot deck)
- Choose a “dial” and apply it at a few settings
- Choose a few models and fit them repeatedly
- Choose a few quality metrics and calculate them
    - on every model…
    - fitted to every sample from this pop…
    - at every setting of the dial…
    - by domain, averaged across simulation reps
- Compare metrics
    - How does turning this dial affect the quality of each model’s fits?
    - Which models tend to work best at the most realistic/relevant dial settings?

Some of our dials

- In today's examples:
  proportional sample-size reduction dial
  (compare {n_i vs n_i/2 vs n_i/4} in each domain *i*)
- Other "dials" (experimental settings):
  - Effect of reduced sample sizes in each domain
  - Effect of reduced number of domains
  - "Fit one wildfire at a time" vs
    "Fit a year's wildfires all at once"
  - …

Some of our models

- In today’s examples: two ways of fitting an area-level model (Fay-Herriot model using REML for the model variance):
  - `sae` R package (Molina & Marhuenda 2020), and
  - `hbsae` R package (Boonstra 2022)
- FIA’s baseline model is post-stratification
- Weights-based (post-strat or GREG) are easily applied to new domains (e.g. fires!) – but no good for tiny domains with 0 sample plots
- For unit-level models, the units are the surveyed sample plots (not individual trees) and their corresponding pixels

Some of our quality metrics

- In today’s examples:
  relative bias; ratios of mean(MSEhat) to MSE; and 95% CI coverage
- Other metrics of interest:
  - Absolute relative bias; MSE; bias in MSEhat; frequency of NAs; computational cost…
- One possible approach: take **means across reps within every domain**, and plot distribution across domains
  - Wieczorek, Nugent, and Hawala (2012),
    “A Bayesian Zero-One Inflated Beta Model for Small Area Shrinkage Estimation”

Relating *k*NN hot deck to other work

- In the terminology used by `StatMatch` R package (D'Orazio 2022):
  - We do "**random hot deck**" within groups defined by Euclidean distance and strata
  - Specifically, we use *k*NN on (transformed, centered, and scaled) continuous variables, done separately within each `tnt` stratum
  - However, unlike this and other packages, our goal is not to impute Y to get better estimates, but rather to simulate a population and samples
- We use `get.knnx()` in the `FNN` R package (Li 2019) to find the *k* NNs

Further complexities of FIA survey data

- Field crews actually record attributes **by condition class** (public vs private ownership, tree species, etc.), with a lot of nuance as to how continuous-area sample plots are divided up into discrete condition classes…
  For simplicity, today's talk only looked at totals across all condition classes (e.g. total basal area of all trees in a plot, not separated out by tree species).
- **Nonresponse**: dangerous weather conditions, private landowner does not allow field crew access to take measurements, etc. For simplicity, today's talk treated this as ignorable.

ECOMAP ecological units

- Contains regional geographic delineations for analysis of ecological relationships across ecological units. ECOMAP is the term used for a USDA Forest Service initiative to map ecological units and encourage their use in ecosystem-based approaches to forest land conservation and management. It is coordinated at the national and regional levels by USDA Forest Service staff and implemented in cooperation with State forestry agencies and others.
- ECOMAP mapping criteria are outlined in the National Hierarchical Framework of Ecological Units (https://www.ncrs.fs.fed.us/gla/reports/hierarchy.htm). The framework systematically divides the country into progressively smaller areas of land and water that have similar physical and biological characteristics and ecological processes.
    - Cleland, D.T.; Freeouf, J.A.; Keys, J.E., Jr.; Nowacki, G.J.; Carpenter, C; McNab, W.H. 2007. Ecological Subregions: Sections and Subsections of the Conterminous United States [1:3,500,000] [CD-ROM]. Sloan, A.M., cartog. Gen. Tech. Report WO-76. Washington, DC: U.S. Department of Agriculture, Forest Service.

ECOMAP ecological units

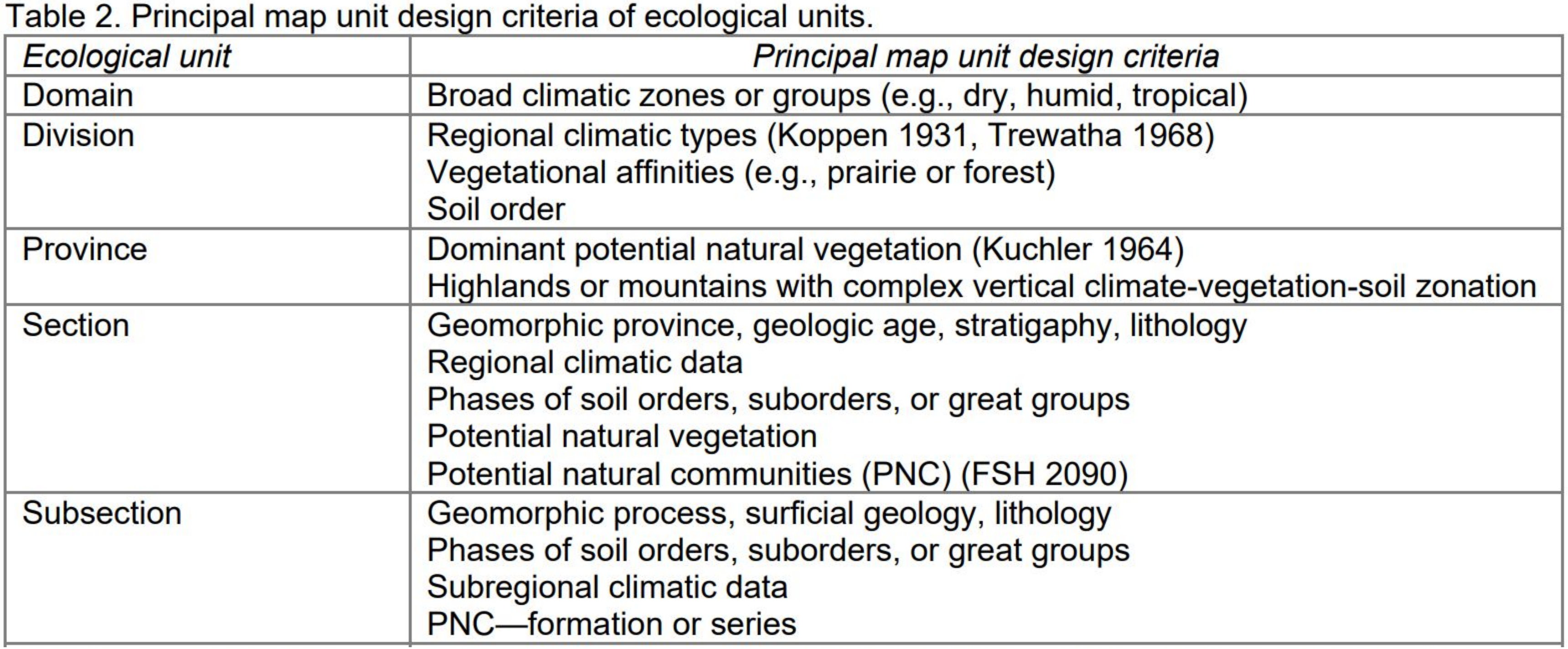
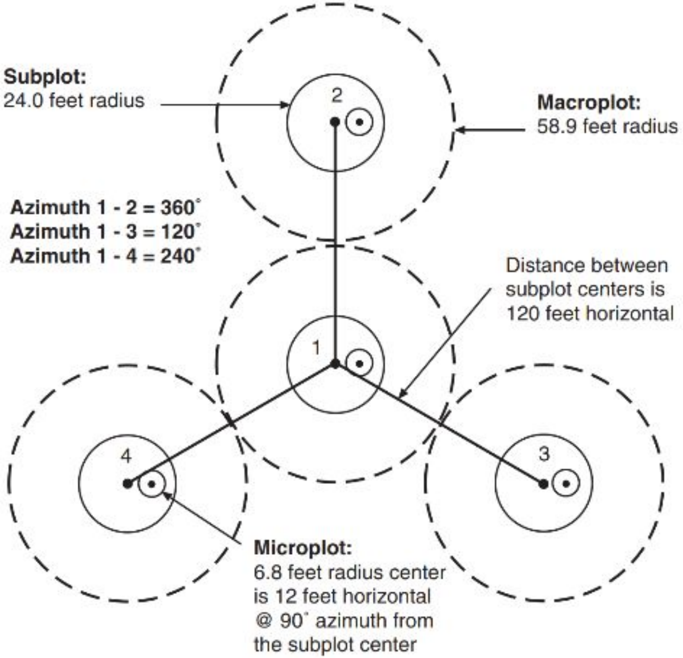


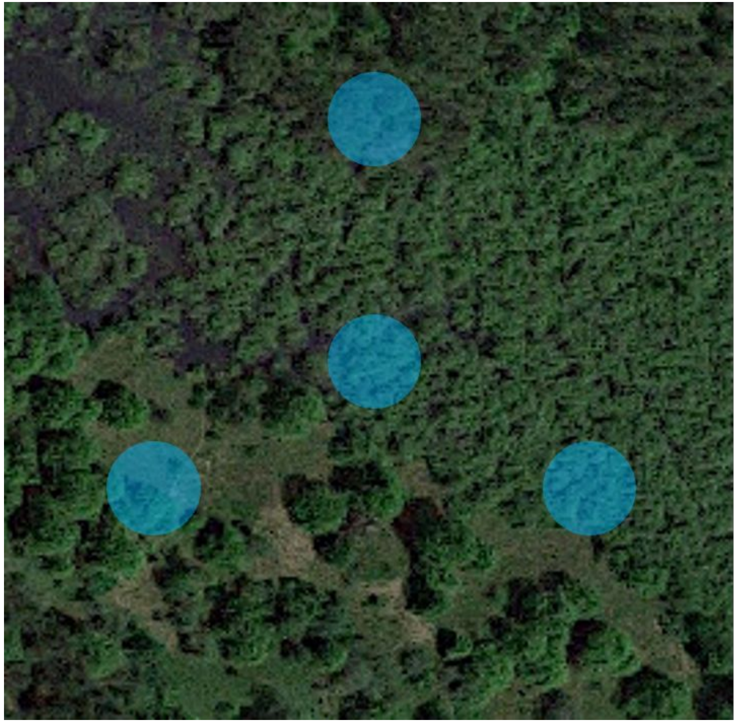
Table 2. Principal map unit design criteria of ecological units.

Ecological unit	*Principal map unit design criteria*
Domain	Broad climatic zones or groups (e.g., dry, humid, tropical)
Division	Regional climatic types (Koppen 1931, Trewatha 1968) Vegetational affinities (e.g., prairie or forest) Soil order
Province	Dominant potential natural vegetation (Kuchler 1964) Highlands or mountains with complex vertical climate-vegetation-soil zonation
Section	Geomorphic province, geologic age, stratigaphy, lithology Regional climatic data Phases of soil orders, suborders, or great groups Potential natural vegetation Potential natural communities (PNC) (FSH 2090)
Subsection	Geomorphic process, surficial geology, lithology Phases of soil orders, suborders, or great groups Subregional climatic data PNC—formation or series

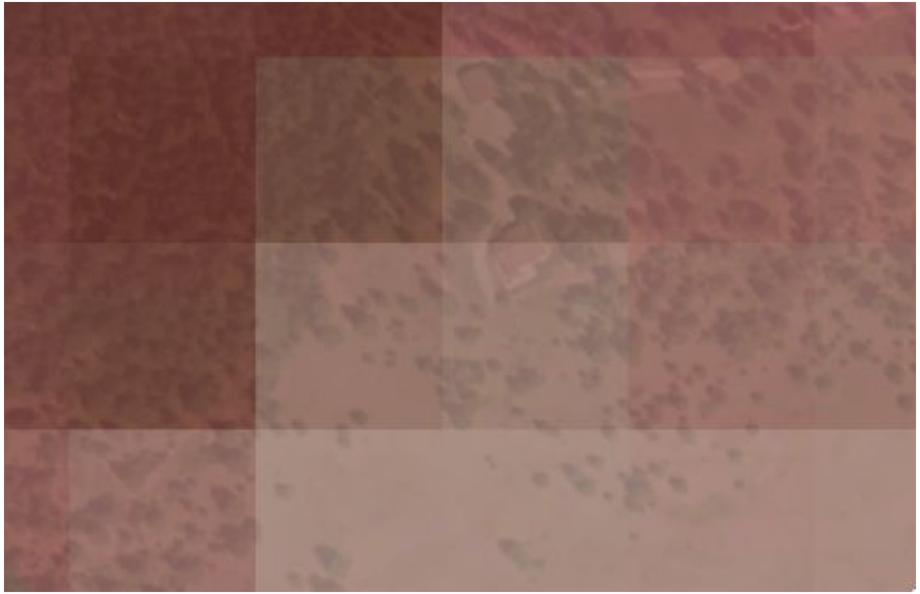
Sample plot with subplots, within a 90x90m pixel

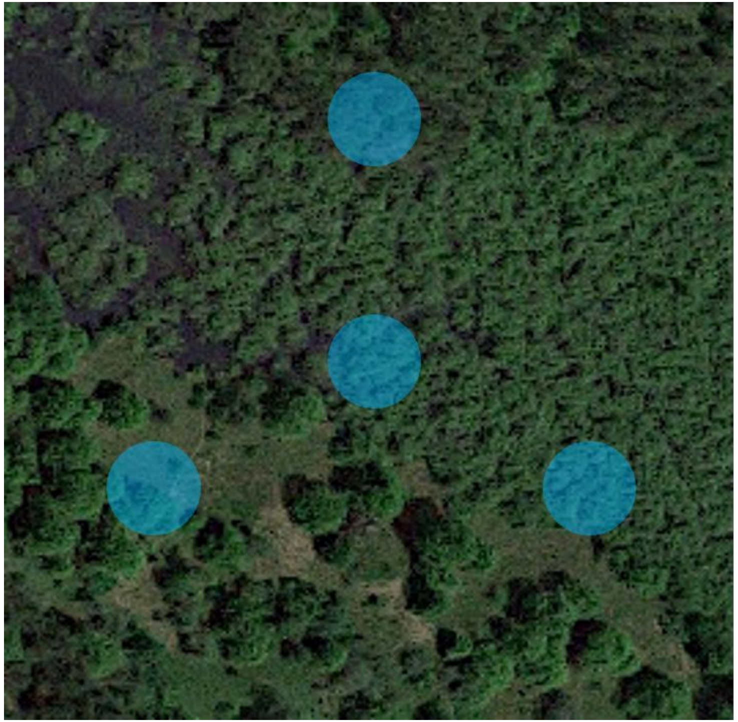


Subplot:
24.0 feet radius
2
Macroplot:
58.9 feet radius
Azimuth 1 - 2 = 360˚
Azimuth 1 - 3 = 120˚
Azimuth 1 - 4 = 240˚
Distance between subplot centers is 120 feet horizontal
1
4
3
Microplot:
6.8 feet radius center is 12 feet horizontal @ 90˚ azimuth from the subplot center

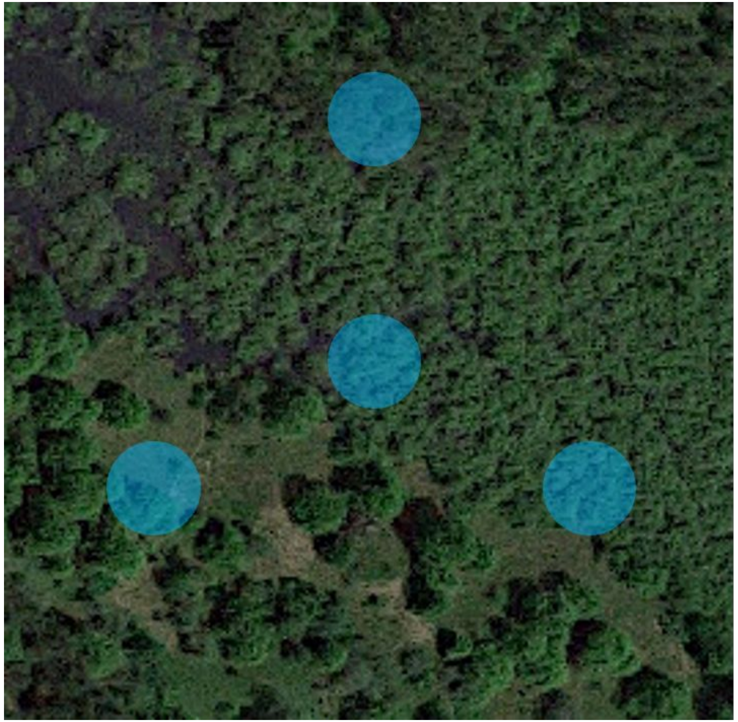


Sample plot with subplots, within a 90x90m pixel

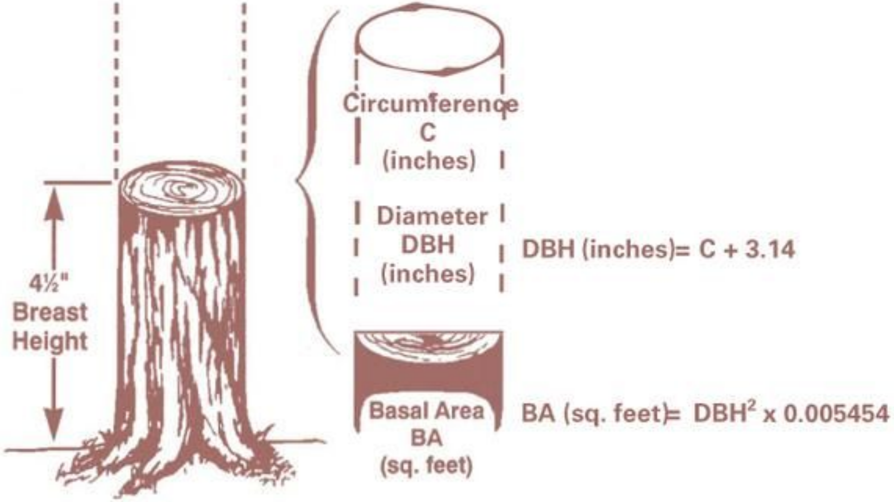


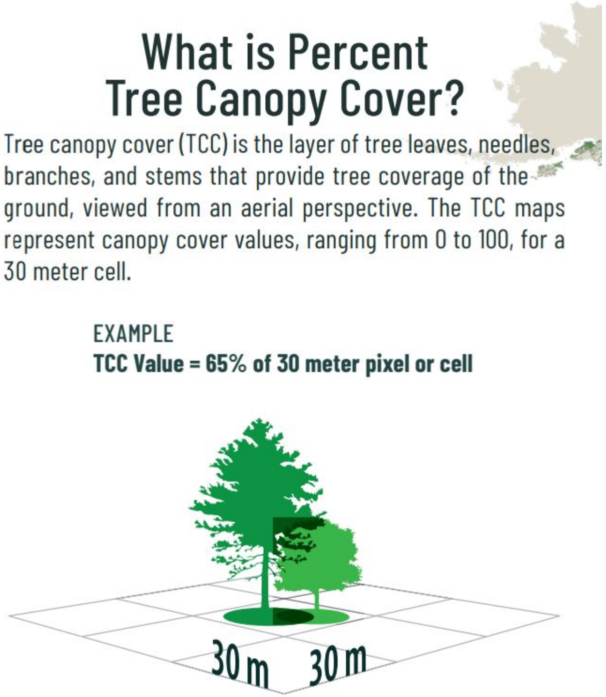


Sample plot with subplots, within a 90x90m pixel



Defining BA and TCC



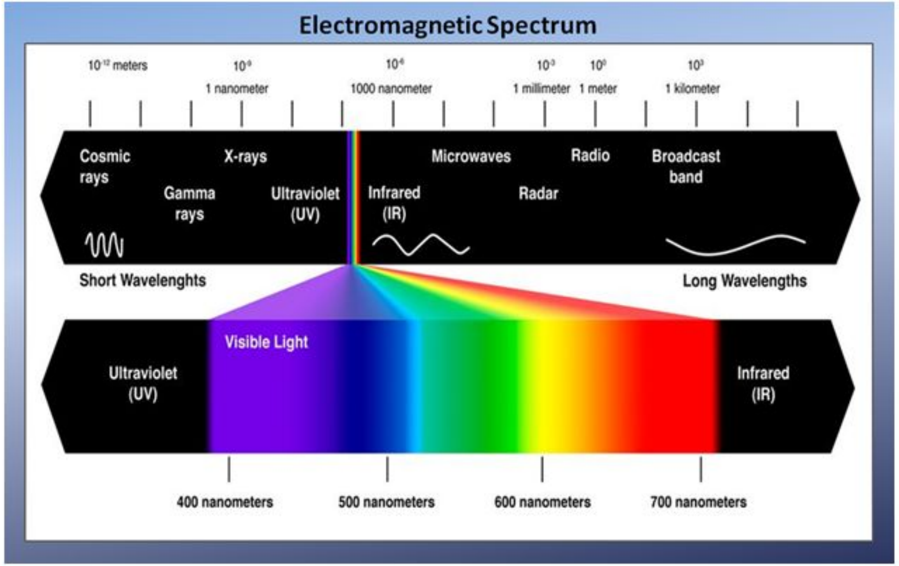


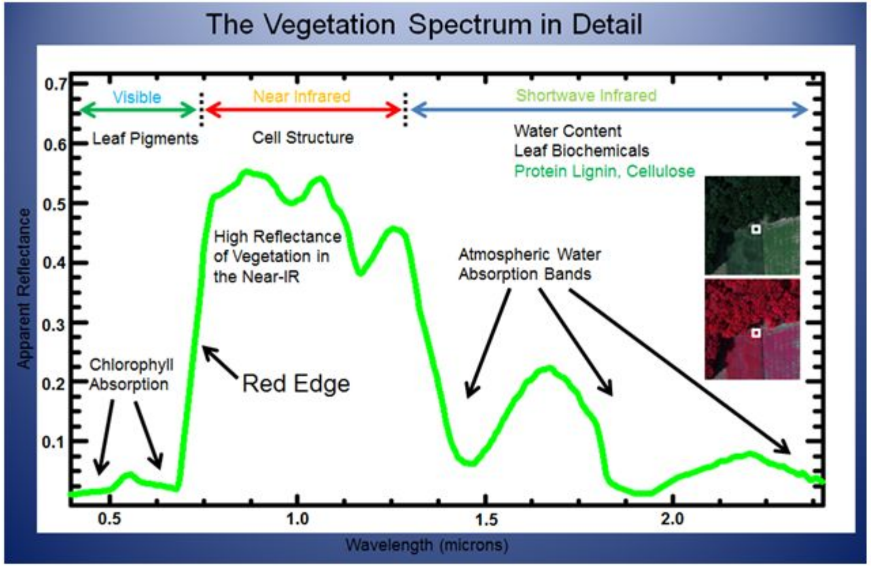
What is Percent
Tree Canopy Cover?

Tree canopy cover (TCC) is the layer of tree leaves, needles, branches, and stems that provide tree coverage of the ground, viewed from an aerial perspective. The TCC maps represent canopy cover values, ranging from 0 to 100, for a 30 meter cell.

EXAMPLE
TCC Value = 65% of 30 meter pixel or cell

EVI and related vegetation indices





EVI and related vegetation indices

Normalized Difference Vegetation Index (NDVI):

The NDVI is perhaps the most well known and often used vegetation index. The NDVI is a simple, but effective VI for quantifying green vegetation. The NDVI normalizes green leaf scattering in the near-infrared wavelength and chlorophyll absorption in the red wavelength.

$\text{NDVI} = (\text{NIR} – \text{RED}) / (\text{NIR} + \text{RED})$

The value range of an NDVI is -1 to 1 where healthy vegetation generally falls between values of 0.20 to 0.80.

Enhanced Vegetation Index (EVI):

In areas of dense canopy where the leaf area index (LAI) is high, the NDVI values can be improved by leveraging information in the blue wavelength. Information in this portion of the spectrum can help correct for soil background signals and atmospheric influences.

$\text{EVI} = 2.5[(\text{NIR} – \text{RED}) / ((\text{NIR}) + (6\text{RED}) - (7.5\text{BLUE}) + 1)]$

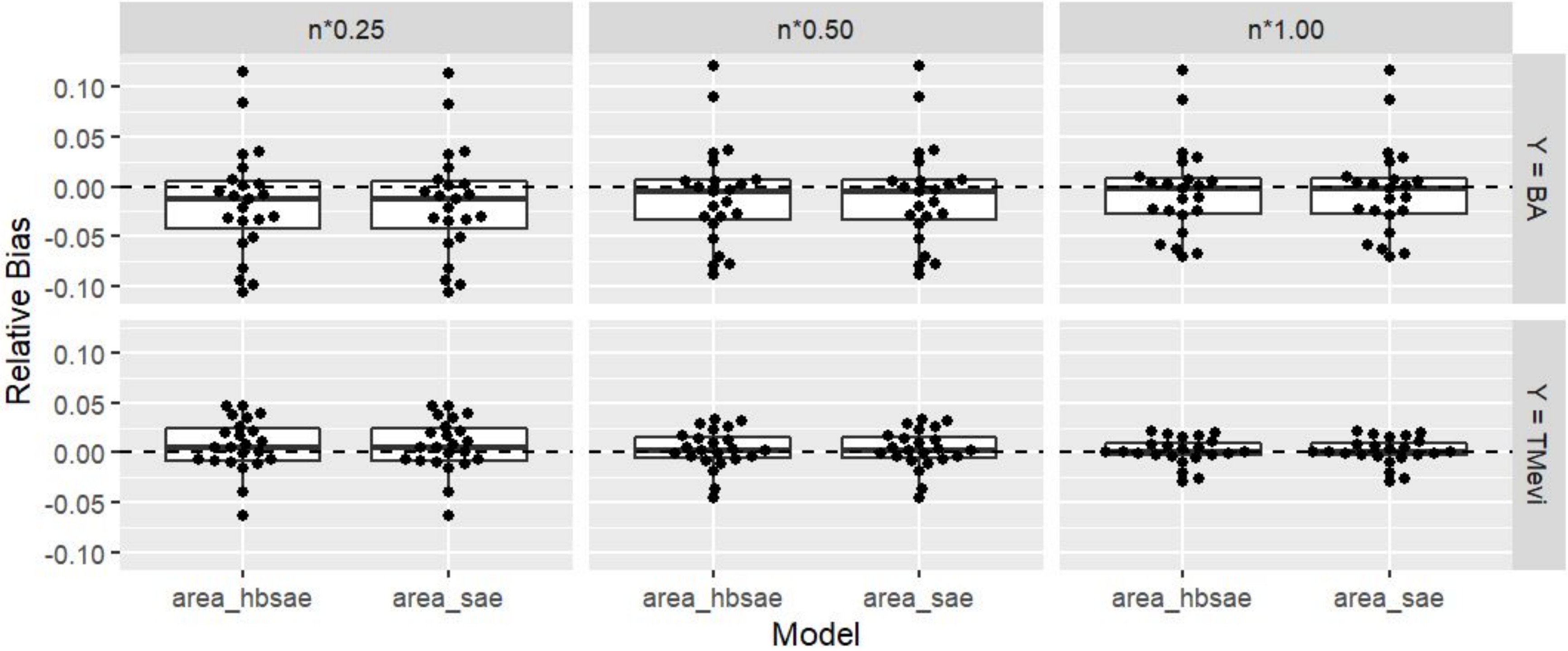
The range of values for the EVI is -1 to 1, where healthy vegetation generally falls between values of 0.20 to 0.80.

More approaches to summarizing quality metrics

- Today’s examples will use relative bias; ratios of mean(MSEhat) to MSE; and 95% CI coverage
- Other metrics of interest:
  - Absolute relative bias; MSE; bias in MSEhat; frequency of NAs; computational cost…
- **Two** approaches:
  - Wieczorek, Nugent, and Hawala (2012), “A Bayesian Zero-One Inflated Beta Model for Small Area Shrinkage Estimation” – take means across reps within every domain, and plot distribution across domains
  - Dorfman (2018), “Towards a Routine External Evaluation Protocol for SAE” – take **means across domains within every rep**, and plot distribution across reps

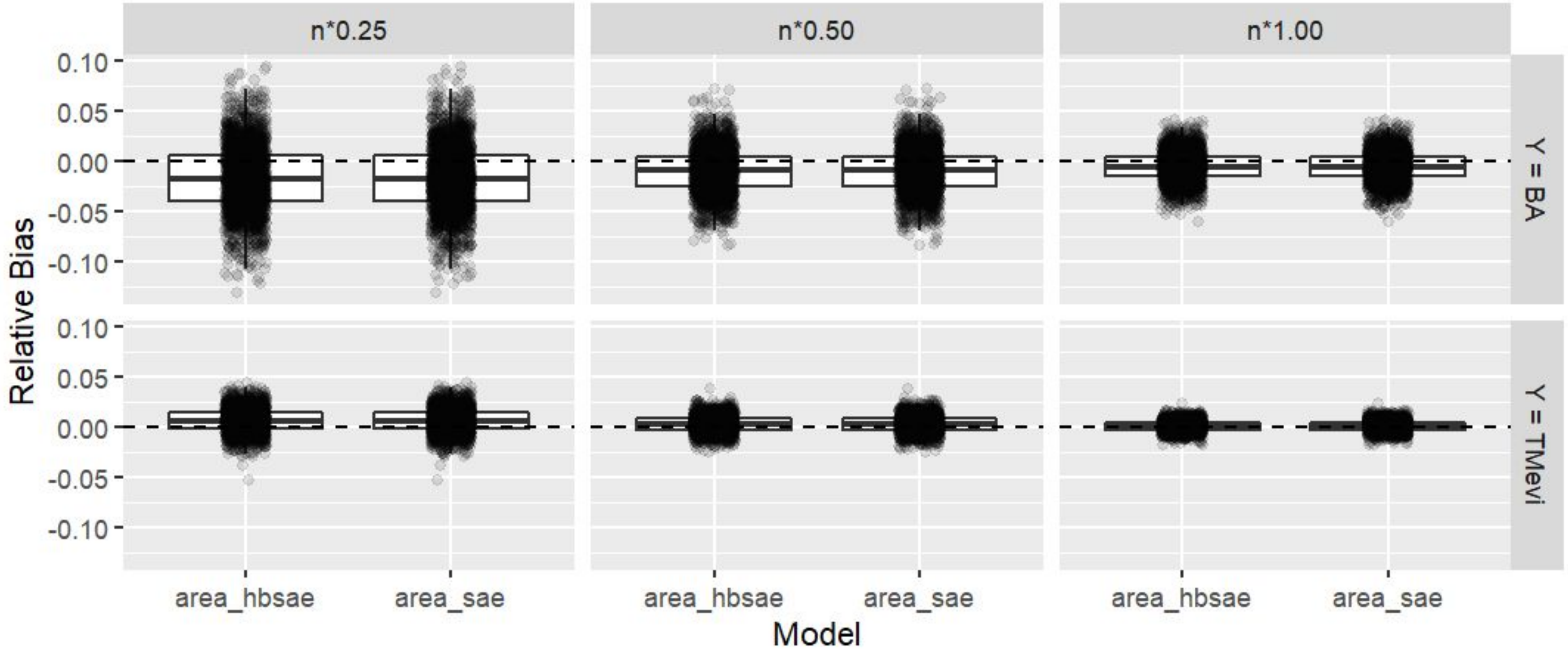
Relative bias, per domain

Relative biases, averaged across reps within each of the 23 subsections;
for different sample sizes and models

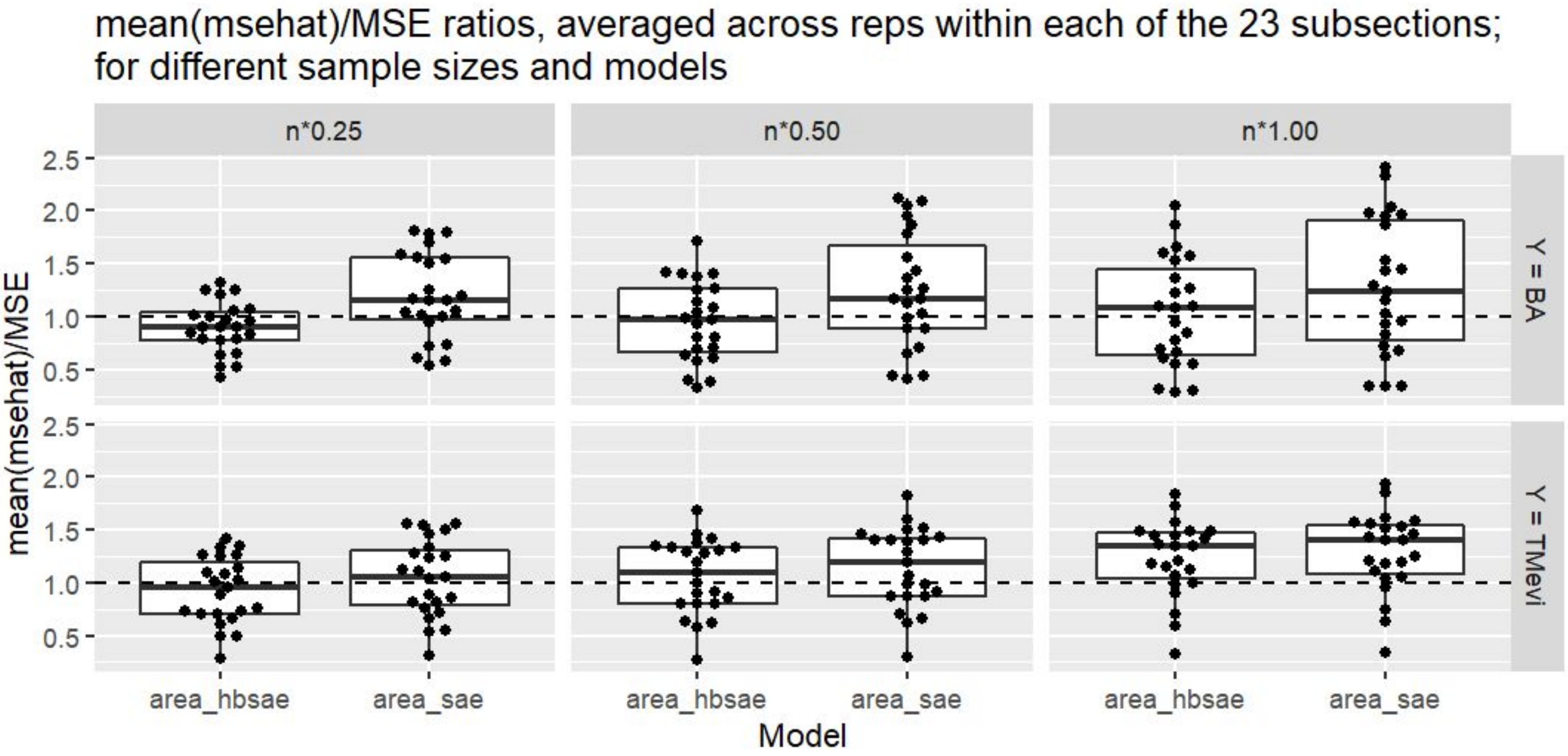


Relative bias, per simulation rep

Relative biases, averaged across subsections within each of the 2500 reps;
for different sample sizes and models

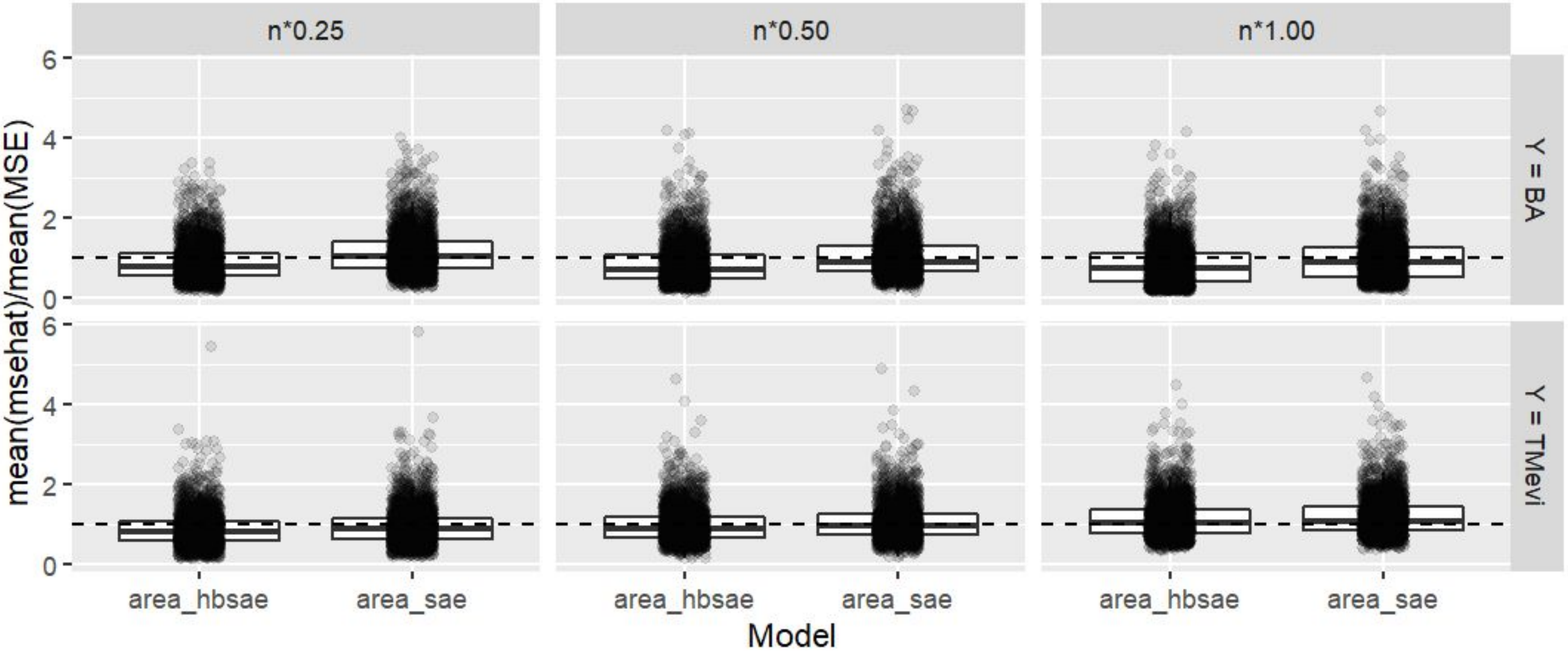


mean(MSEhat) / MSE, per domain

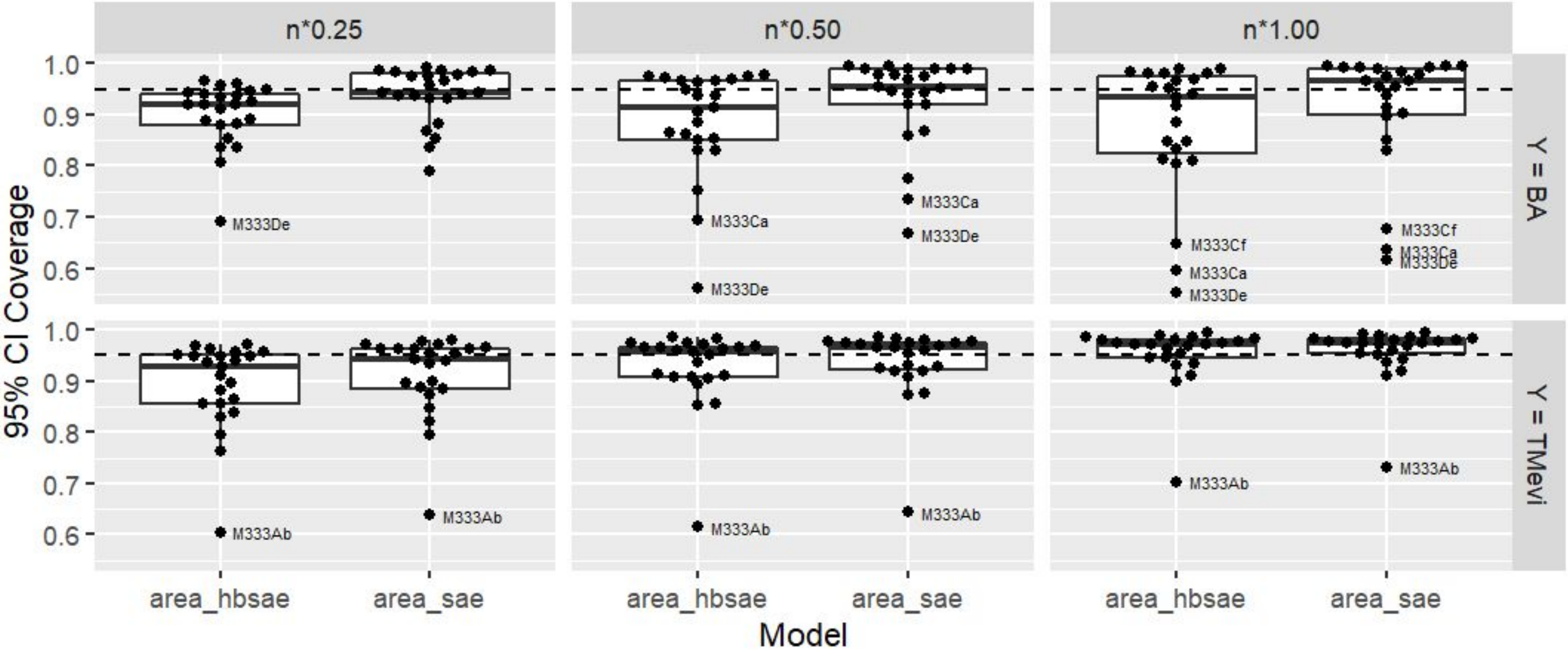


mean(MSEhat) / mean(MSE), per simulation rep

mean(msehat)/mean(MSE), averages taken across subsecs within each of the 2500 reps;
for different sample sizes and models

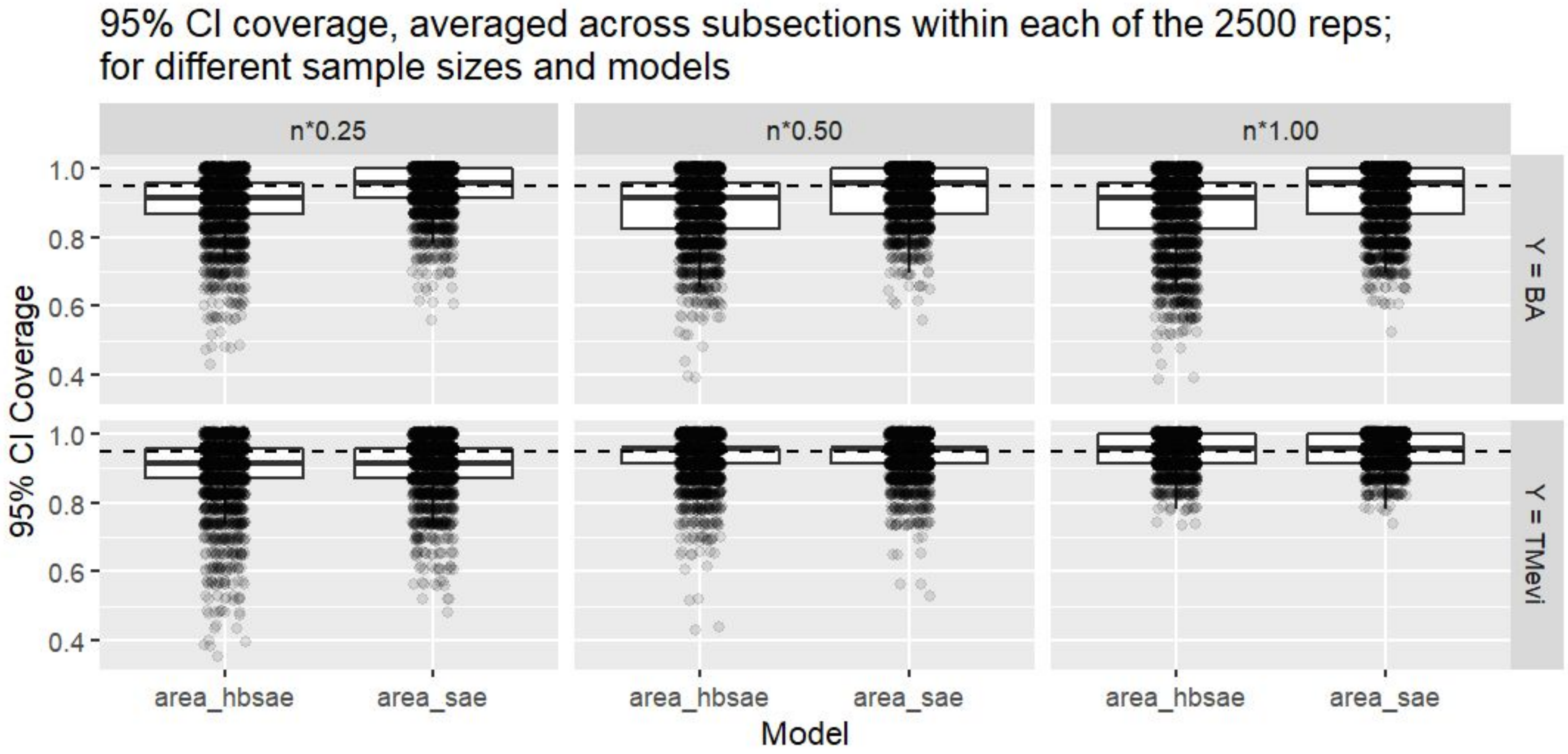


95% CI coverage, per domain



95% CI coverage, averaged across reps within each of the 23 subsections;
for different sample sizes and models

95% CI coverage, per simulation rep


95% CI coverage, averaged across subsections within each of the 2500 reps;
for different sample sizes and models
n*0.25
n*0.50
n*1.00
Y = BA
Y = TMevi
95% CI Coverage
1.0
0.8
0.6
0.4
Model
area_hbsae
area_sae