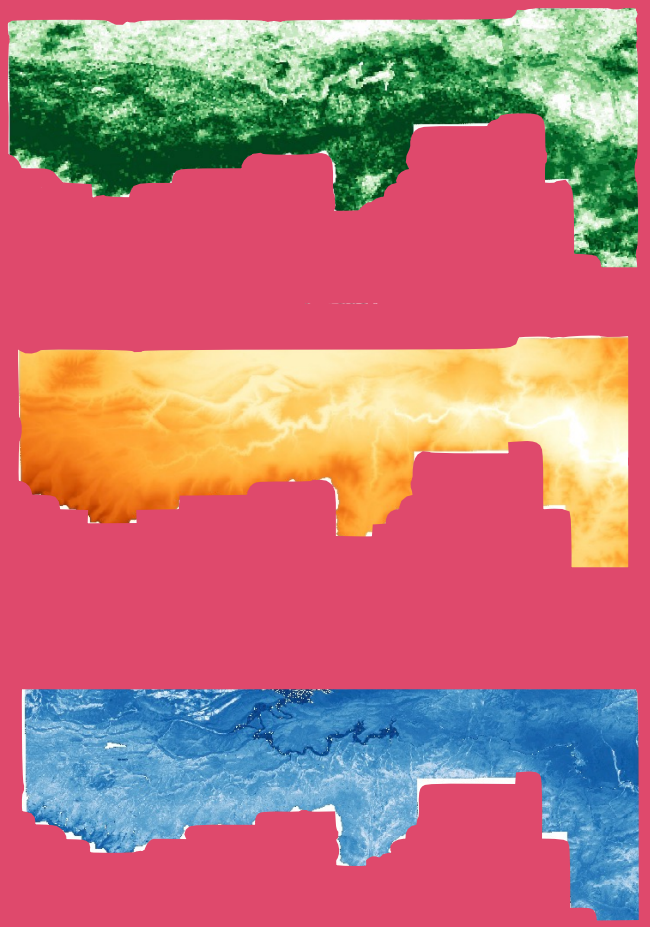
TAMURA LILLY LECTURE: A CAUTIOUS SURVEY STATISTICIAN'S APPROACH TO ESTIMATION IN THE AGE OF BIG DATA



Kelly McConville Bucknell University

Collaborators:

US Forest Inventory and Analysis Program:
Gretchen Moisen, Tracey Frescino
US Bureau of Labor Statistics:
Daniell Toth

Former students:
Josh Yamamoto, Becky Tang, George Zhu, Shirley Cheung, and Sida Li

TYPES OF SURVEY STATISTICIANS

Questionnaire Designer

How should I word the questions?

Is phone or internet better?

How do we increase response rates?

Sampler

How should we select the sample?

How can I ensure my sample is representative?

What is the cheapest sampling design?

How do we increase response rates?

Estimater

How do we account for the sampling design?

What is the best estimation method?

Can we incorporate big data?

How do we adjust for nonresponse?

SAMPLING DESIGNS

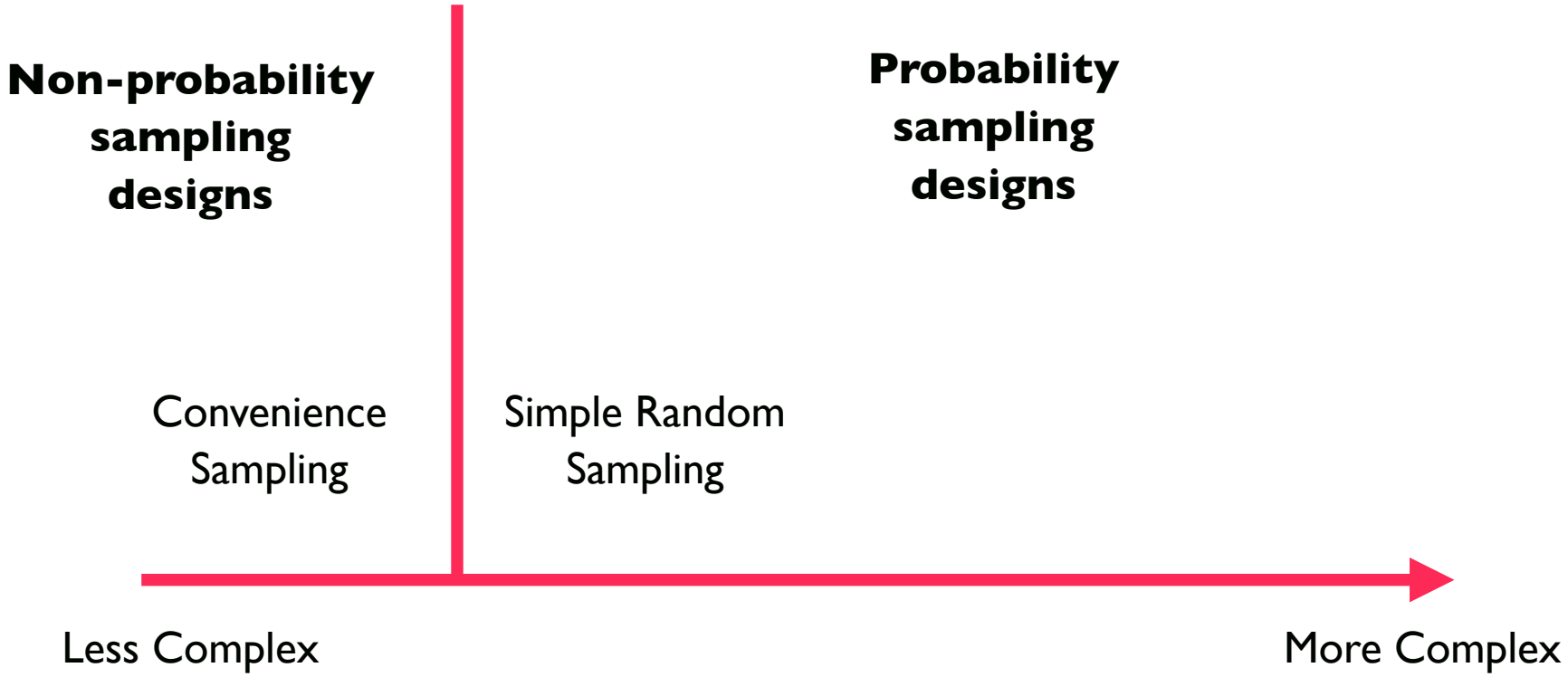
- Interested in the inclusion probabilities: $\pi_i = P(i \in s)$

CONVENIENCE SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- What are the positives and the negatives of this sampling design?

SAMPLING DESIGNS



- Interested in the inclusion probabilities: $\pi_i = P(i \in s)$

SIMPLE RANDOM SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- Sample six people at random where everyone has an equal chance of being selected.

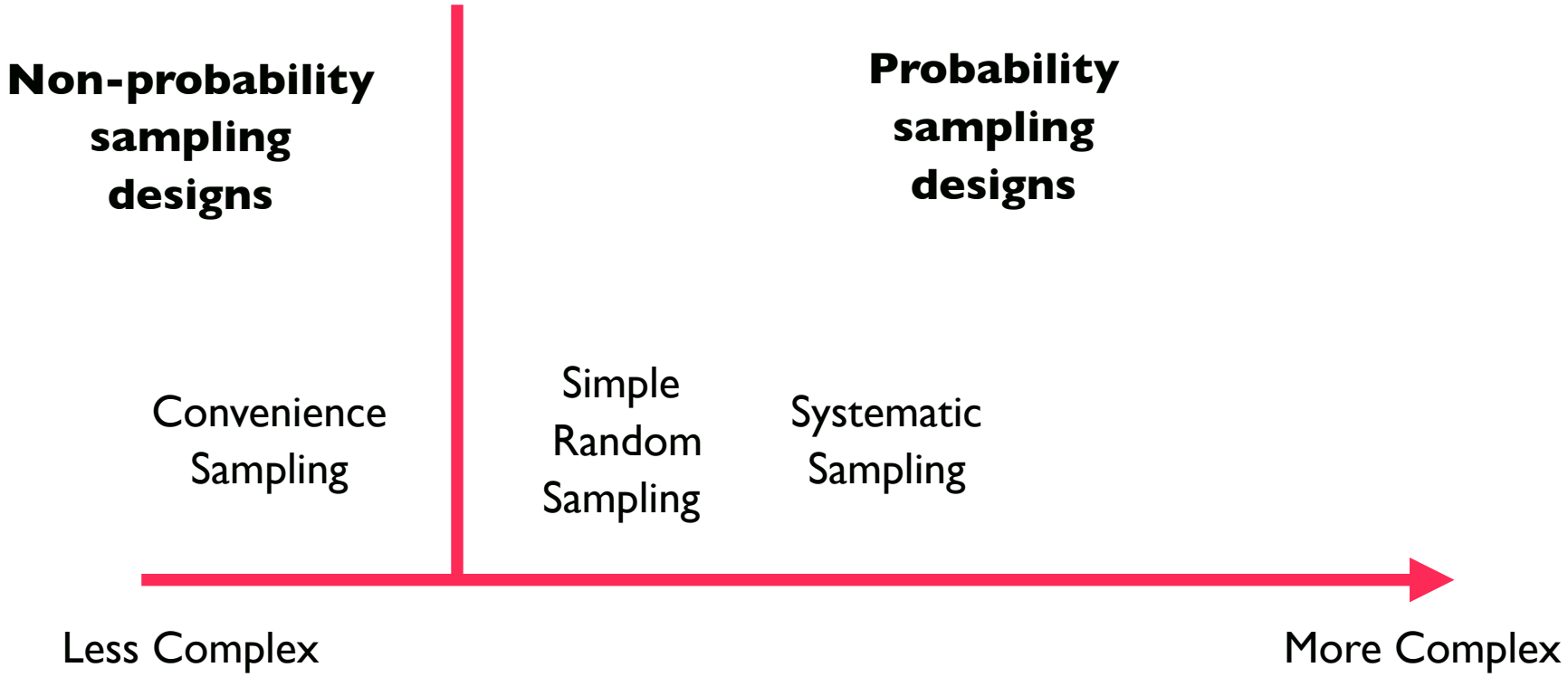
$$\pi_i = 6/40$$

SIMPLE RANDOM SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- What are the positives and the negatives of this sampling design?

SAMPLING DESIGNS



- Interested in the inclusion probabilities: $\pi_i = P(i \in s)$

SYSTEMATIC SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- Randomly select the first person and then sample every seventh person after that.

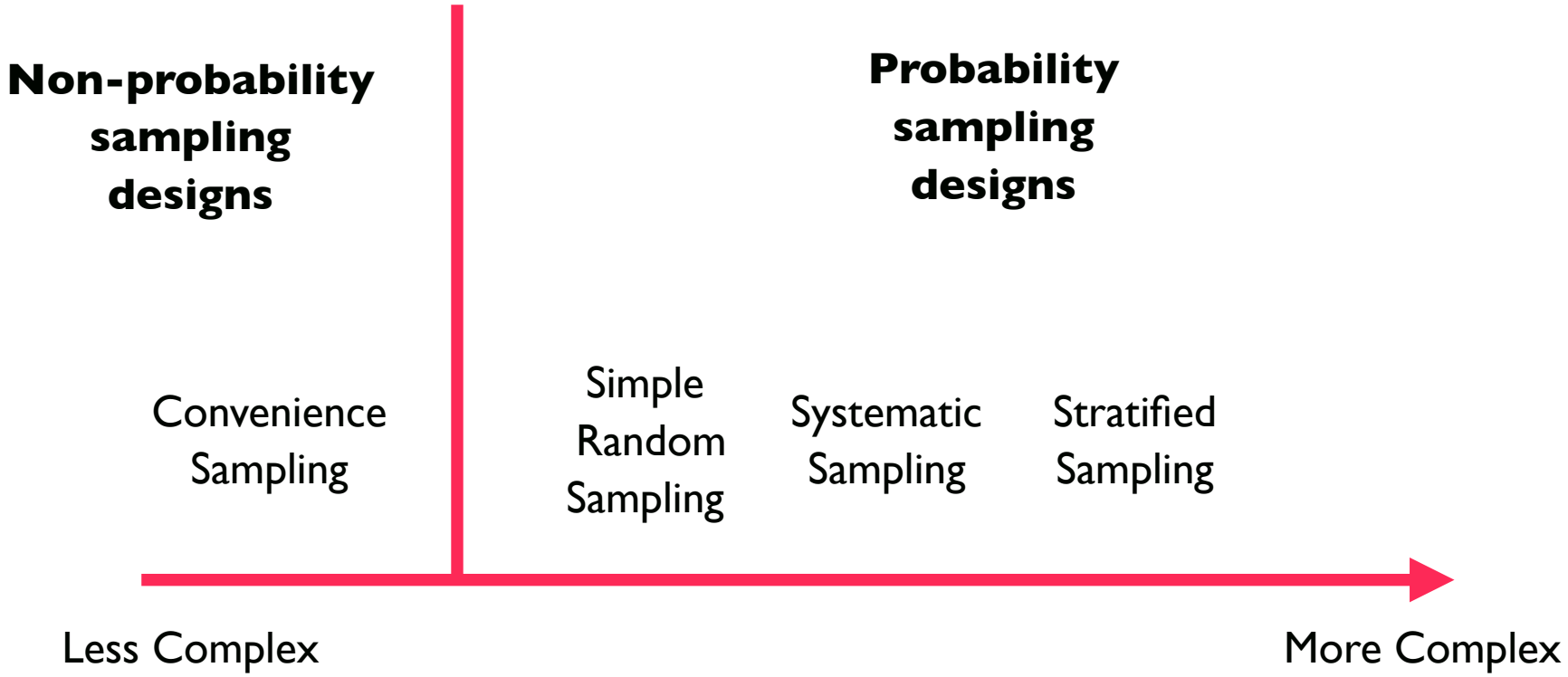
$\pi_i \approx 6/40$

SYSTEMATIC SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- What are the positives and the negatives of this sampling design?

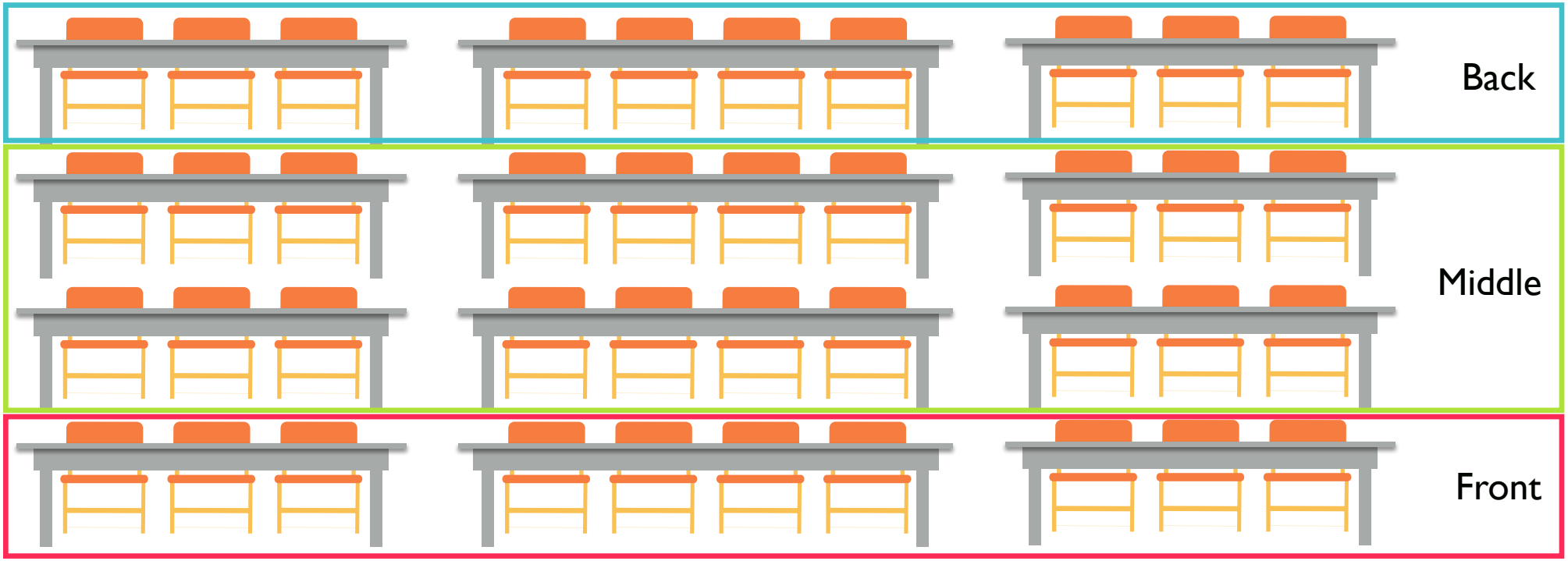
SAMPLING DESIGNS



- Interested in the inclusion probabilities: $\pi_i = P(i \in s)$

STRATIFIED SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.



- Break up the room into homogeneous strata.

STRATIFIED SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- Break up the room into homogeneous strata.
- Sample two people within each.

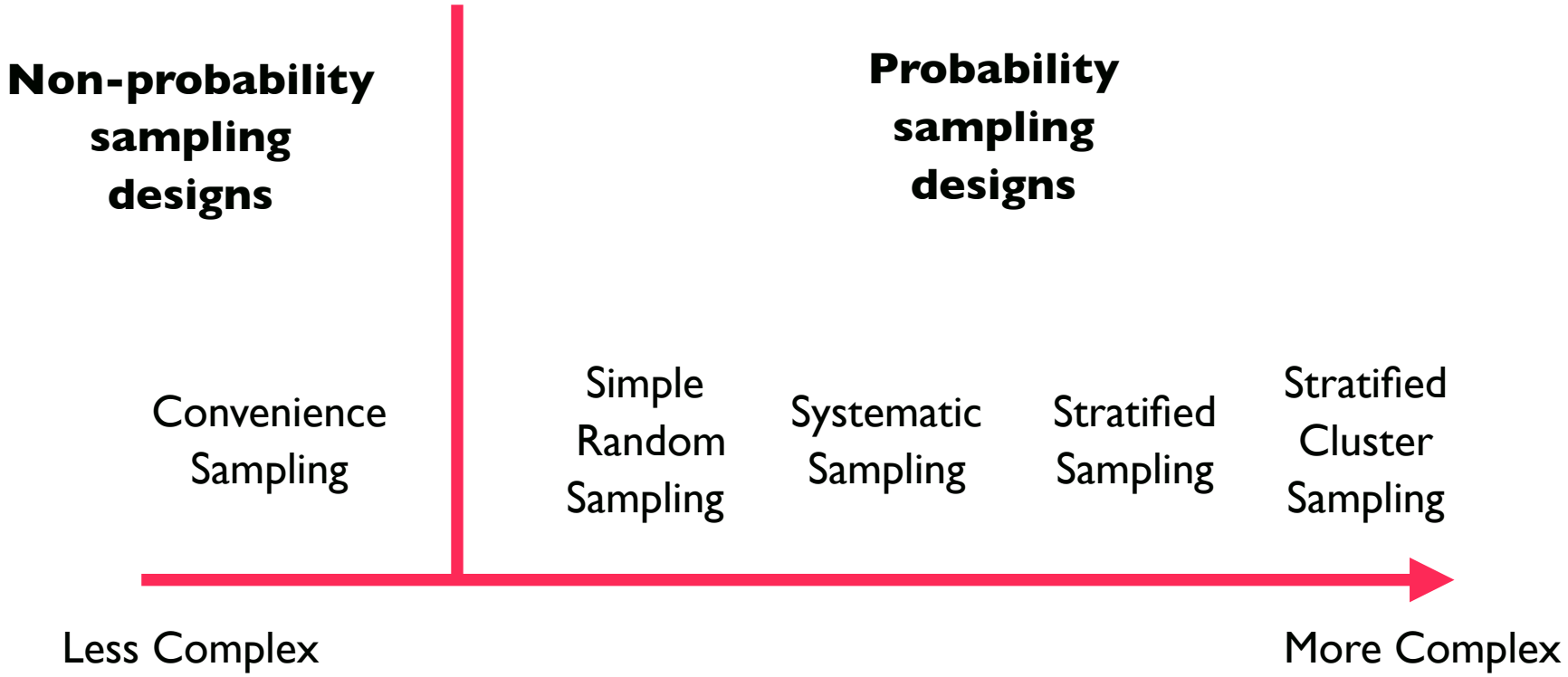
$$\pi_i = \begin{cases} 1/5 & i \text{ in Front or Back} \\ 1/10 & i \text{ in Middle} \end{cases}$$

STRATIFIED SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- What are the positives and the negatives of this sampling design?

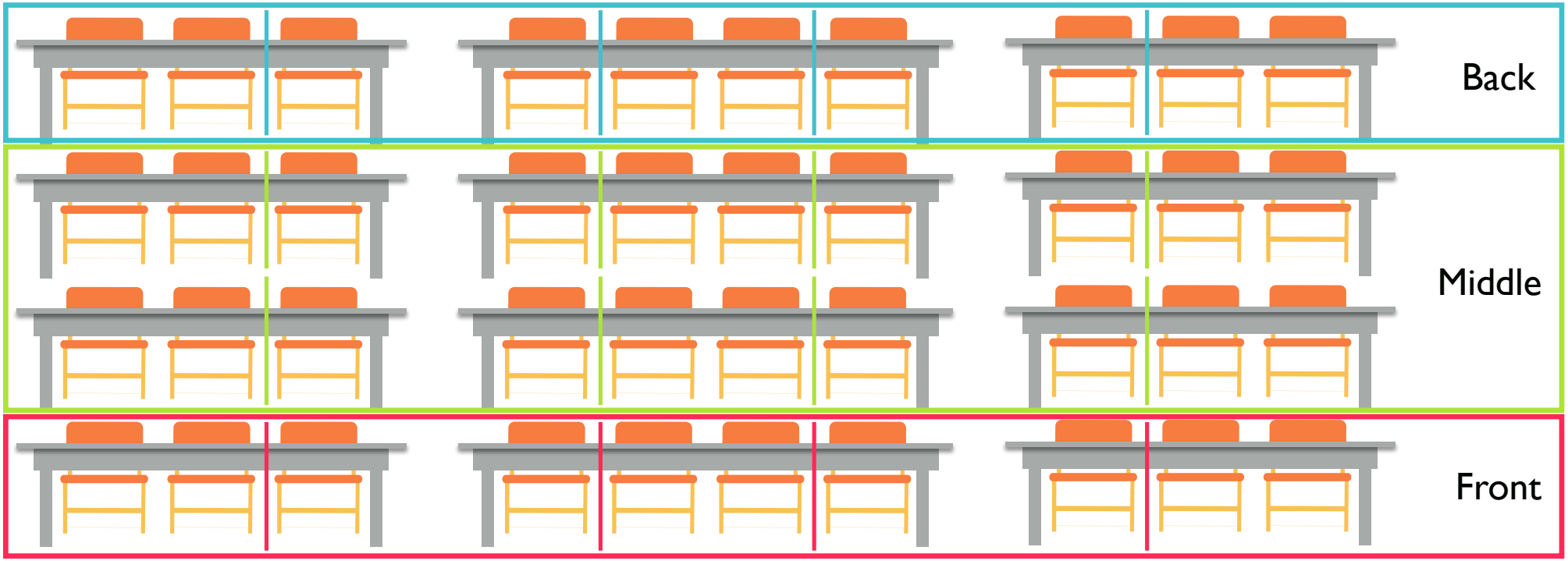
SAMPLING DESIGNS



- Interested in the inclusion probabilities: $\pi_i = P(i \in s)$

STRATIFIED CLUSTER SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.



- Break up the room into homogeneous strata.
- Within each stratum, create clusters of size 2. Sample 1 cluster per stratum.

STRATIFIED CLUSTER SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- Break up the room into homogeneous strata.
- Within each stratum, create clusters of size 2. Sample 1 cluster per stratum.

$$\pi_i = \begin{cases} 1/5 & i \text{ in Front or Back} \\ 1/10 & i \text{ in Middle} \end{cases}$$

STRATIFIED CLUSTER SAMPLING

- Want to estimate what proportion of the audience is enjoying my talk.

- What are the positives and the negatives of this sampling design?

SAMPLING DESIGNS

- Take a probability sample so that you understand how your sample was distributed.
- Consider systematic sampling to ensure sample is spread across the population.
- Consider stratification to ensuring sampling all sub-groups of interest.
- Consider clustering to reduce costs.
- Always determine the inclusion probabilities so that you can use them for estimation!

MODEL-ASSISTED

SURVEY

- Data are collected using a probability sampling design.

ESTIMATION

- What is the average number of trees per acre in Daggett County, Utah?
- How many bartenders are in the US?

- Use models to incorporate "big" data from:
  - Administrative records
  - Remote sensing

ESTIMATION SET-UP

Enumerate the finite population.

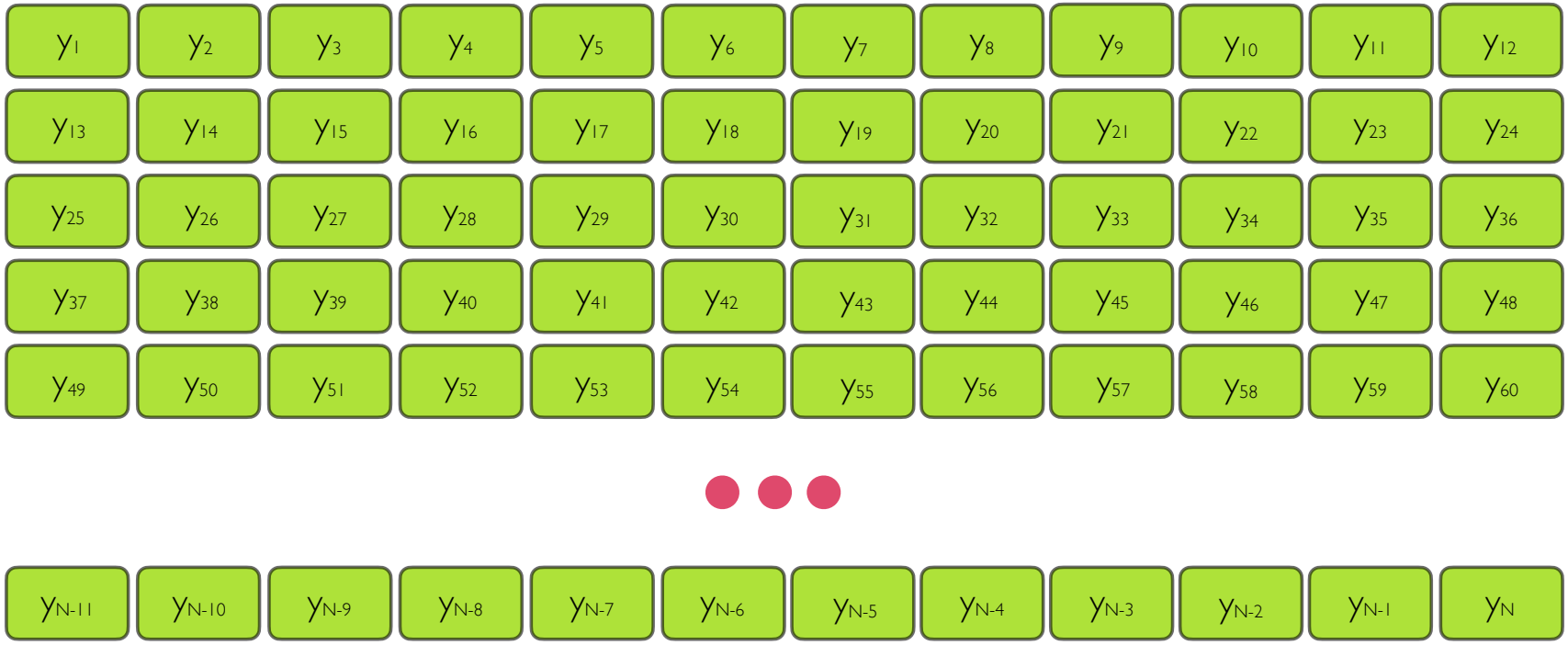
1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60

N-11	N-10	N-9	N-8	N-7	N-6	N-5	N-4	N-3	N-2	N-1	N

$$\{1, 2, \ldots, N\} = U$$

ESTIMATION SET-UP

Goal: Estimate the total of a study variable.



$$t_y = \sum_{i \in U} y_i$$

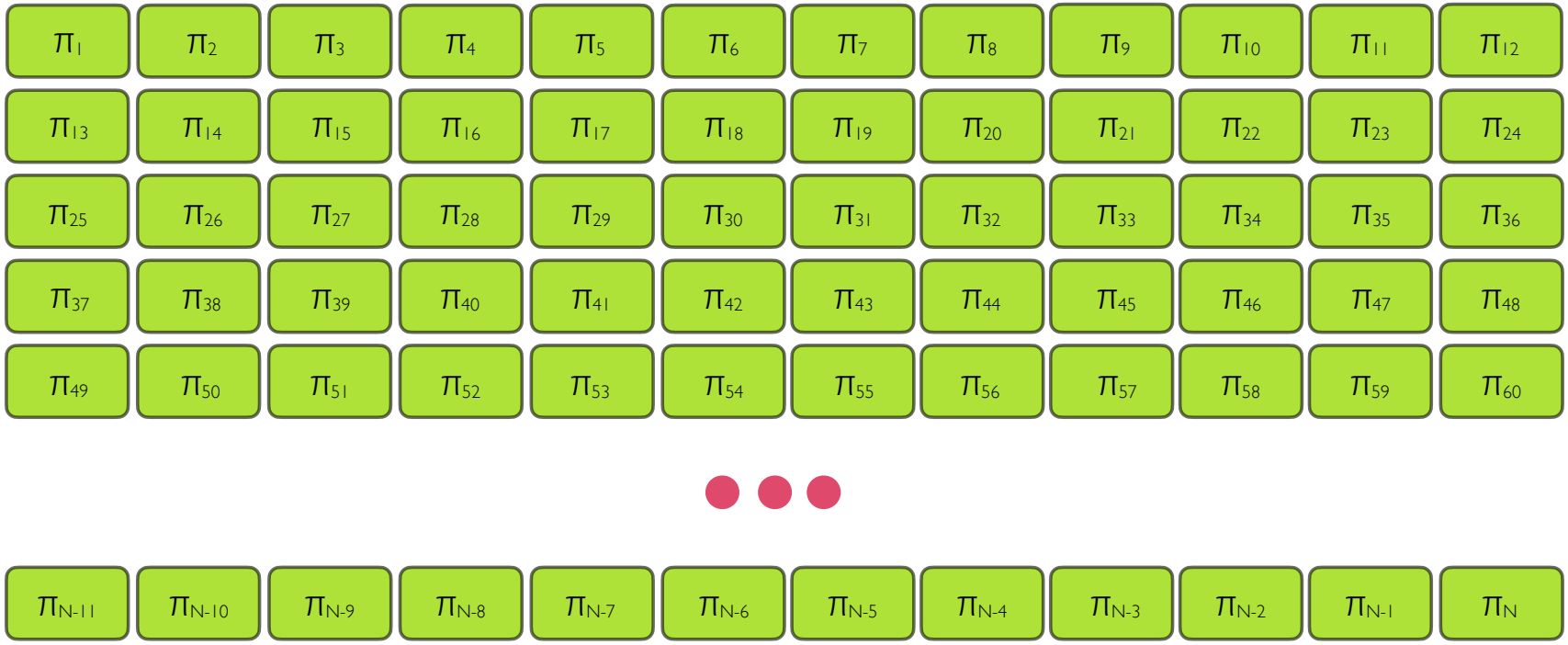
ESTIMATION SET-UP

Don't known the values of the study variable.

Assume additional data are known for every unit in the population.

ESTIMATION SET-UP

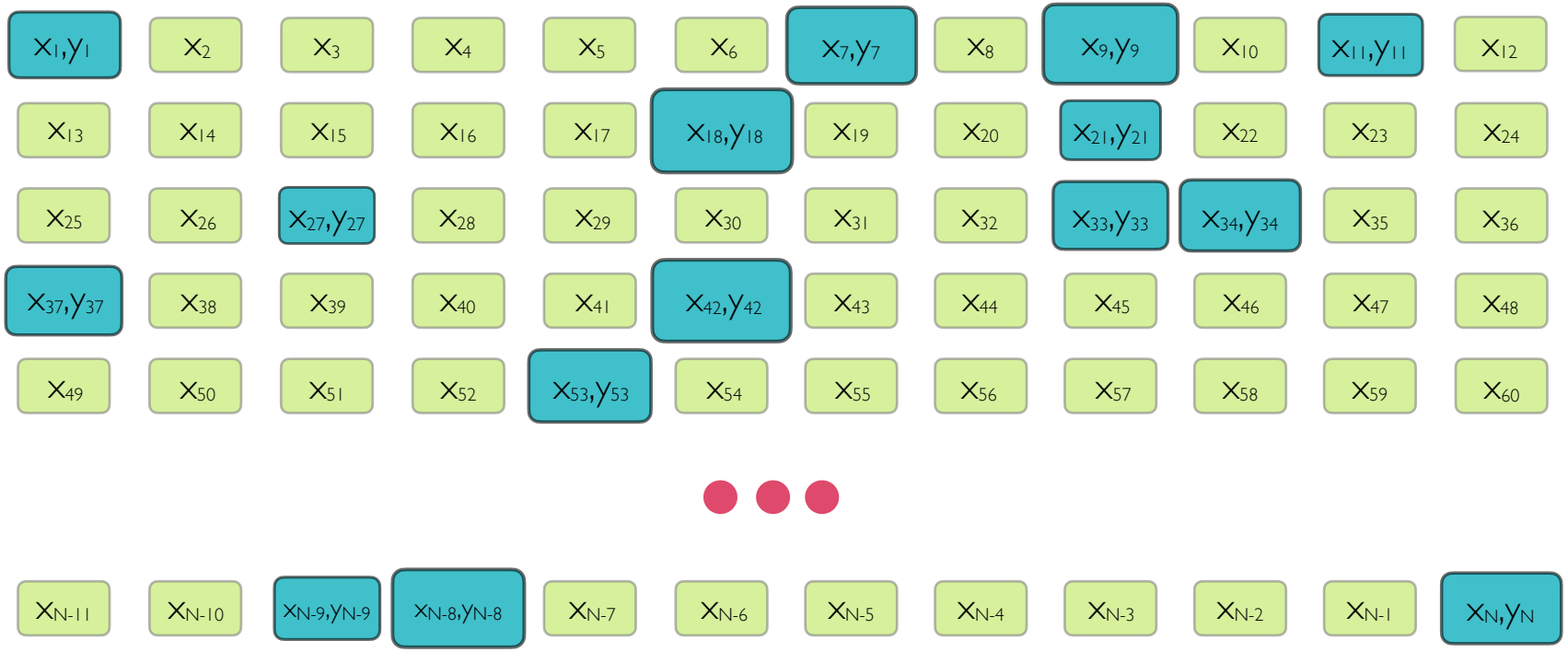
A complex sampling design is constructed.



Inclusion Probabilities: $\pi_i = P(i \in s)$ $\pi_{ij} = P(i, j \in s)$

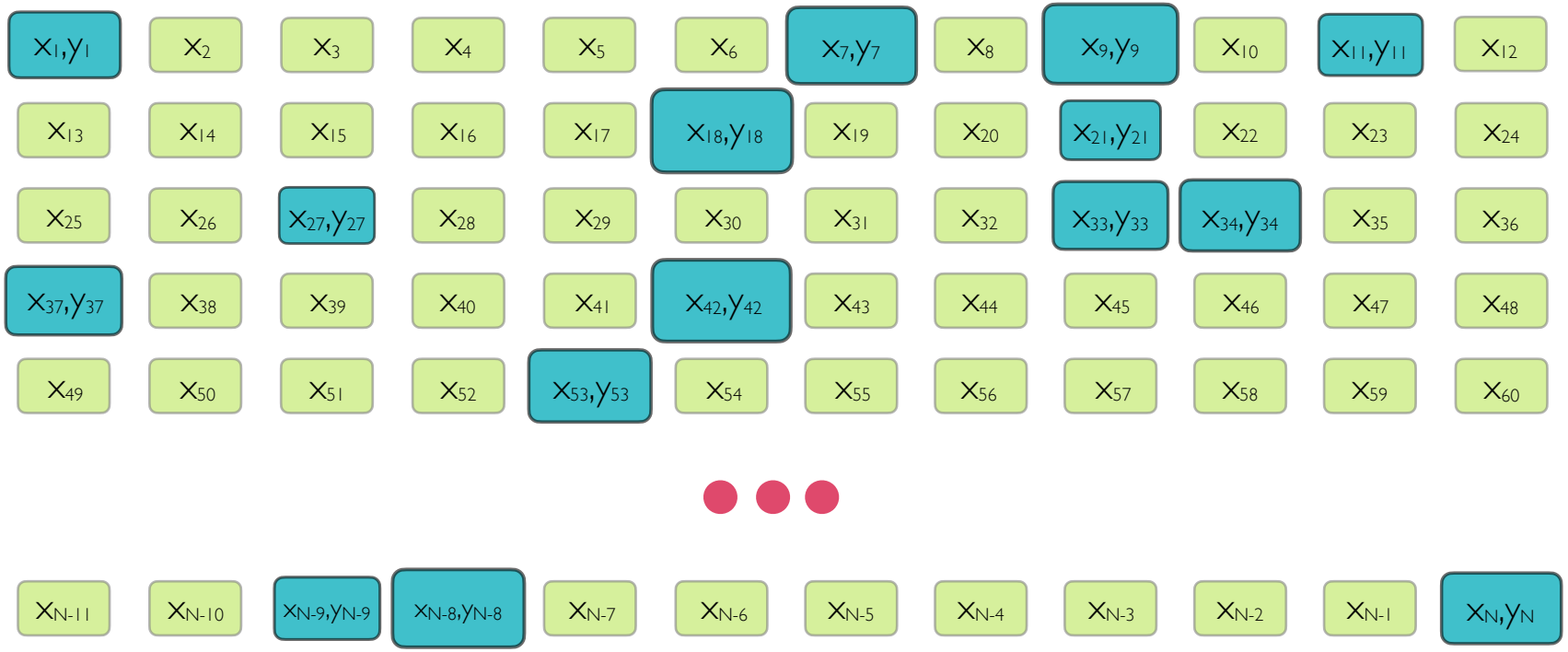
ESTIMATION

The sample is drawn. The study variable and additional data are observed on the sample.



ESTIMATION

The standard estimator uses only the sampled (i.e., blue) data.

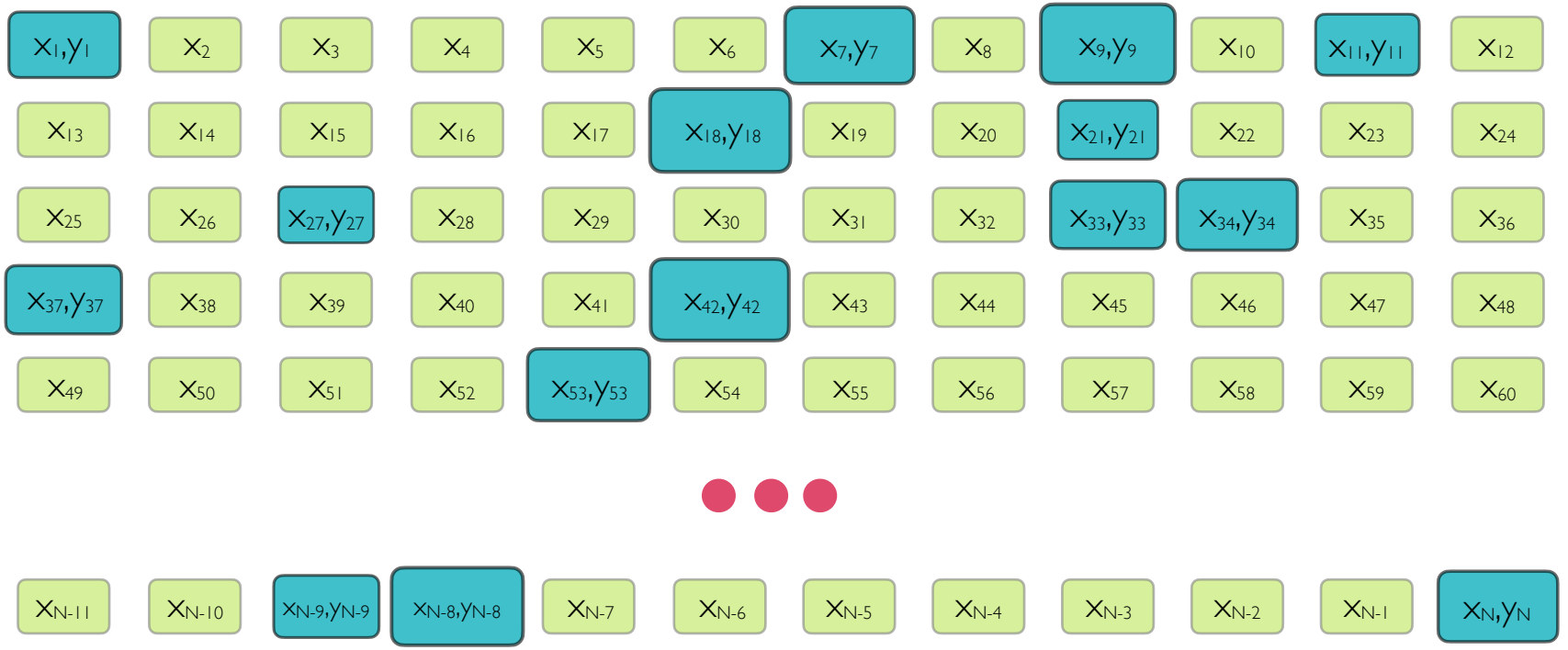


$$\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Horvitz and Thompson (1952)

ESTIMATION

The standard estimator uses only the sampled (i.e., blue) data.



It is unbiased!

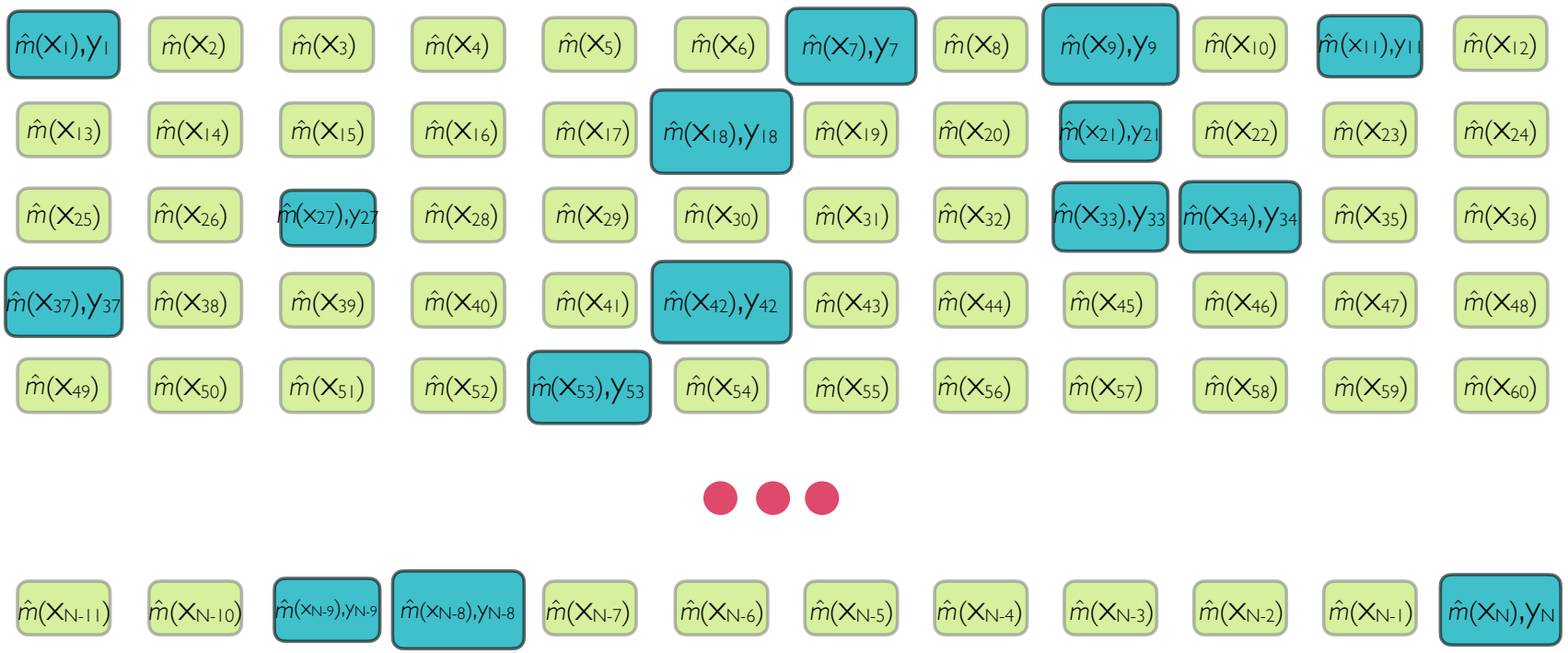
$$\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}$$

But, it is rather variable…

Horvitz and Thompson (1952)

MODEL-ASSISTED ESTIMATOR

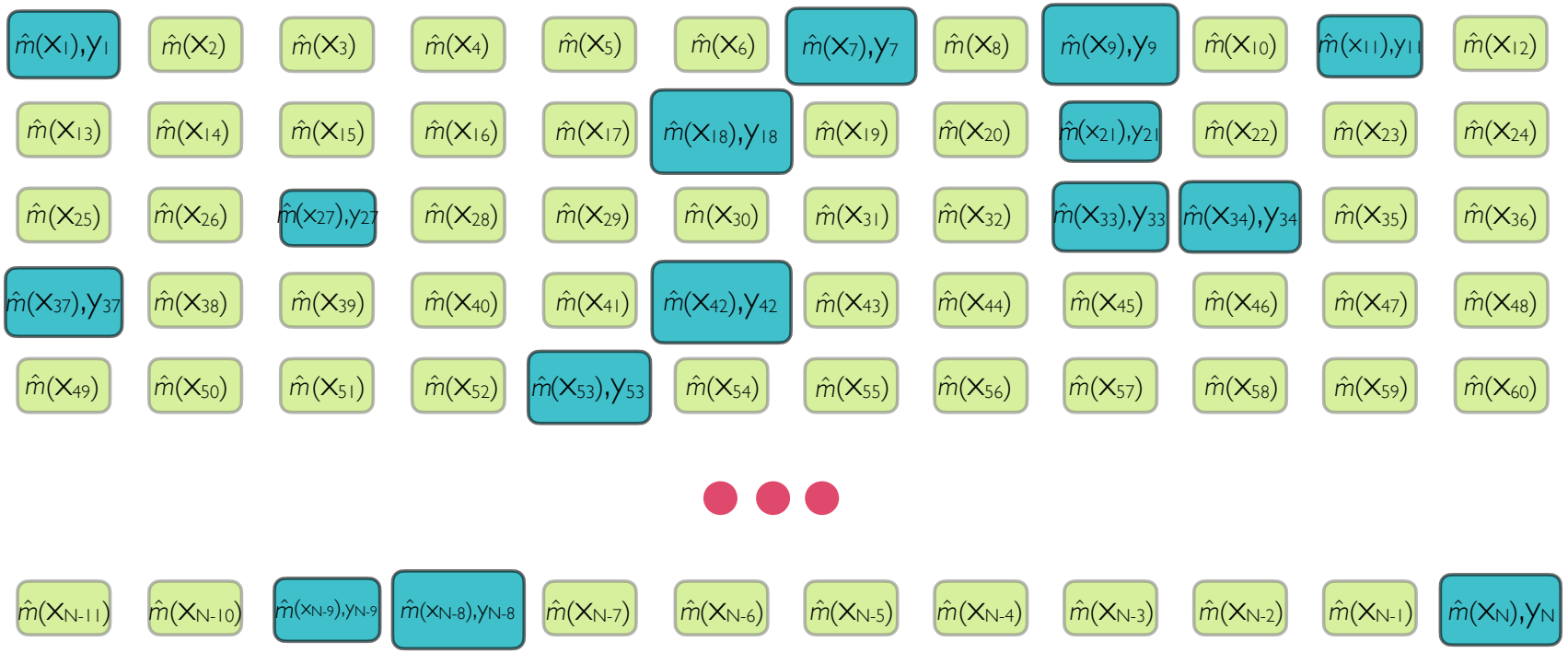
Use the sample data to estimate the study variable where it isn't observed.



$$\hat{t}_y = \sum_{i \in U} \hat{m}(x_i) + \sum_{i \in s} \frac{y_i - \hat{m}(x_i)}{\pi_i}$$

MODEL-ASSISTED ESTIMATOR

Need to determine a good assisting model to construct estimator.



$$\hat{t}_y = \sum_{i \in U} \hat{m}(x_i) + \sum_{i \in s} \frac{y_i - \hat{m}(x_i)}{\pi_i}$$

MODEL-ASSISTED ESTIMATOR

- Model assisted estimator for t_y:

$$\hat{t}_y = \sum_{i \in U} \hat{m}(x_i) + \sum_{i \in s} \frac{y_i - \hat{m}(x_i)}{\pi_i}$$

  - For many assisting models, the estimator has nice properties:
    - Asymptotically Unbiased
    - Small variance
  - But, the **size of the variance** depends on how well the assisting model captures the relationship between the study variable and the additional data.
  - Standard variance estimator:

$$\widehat{\mathrm{Var}}\left(\hat{t}_y\right) = \sum_{i,j \in s} \sum \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} (y_i - \hat{m}(x_i))(y_j - \hat{m}(x_j))$$

WHICH ASSISTING MODEL SHOULD ONE USE?

- Answer depends on…
  - What additional data are available.
  - Appropriately modeling the relationship between the study variable and additional data.
  - Active area of research!

WHICH ASSISTING MODEL SHOULD ONE USE?

PENALIZED REGRESSION

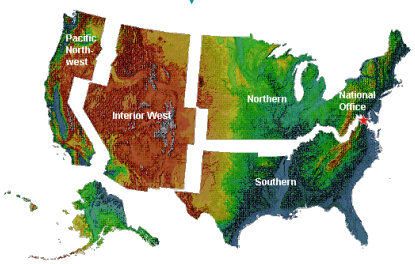
- Estimate forest attributes in Daggett County, Utah.

REGRESSION TREES

- Estimate employment counts for U.S. establishments.

U.S. FOREST INVENTORY AND ANALYSIS (FIA)

FIA



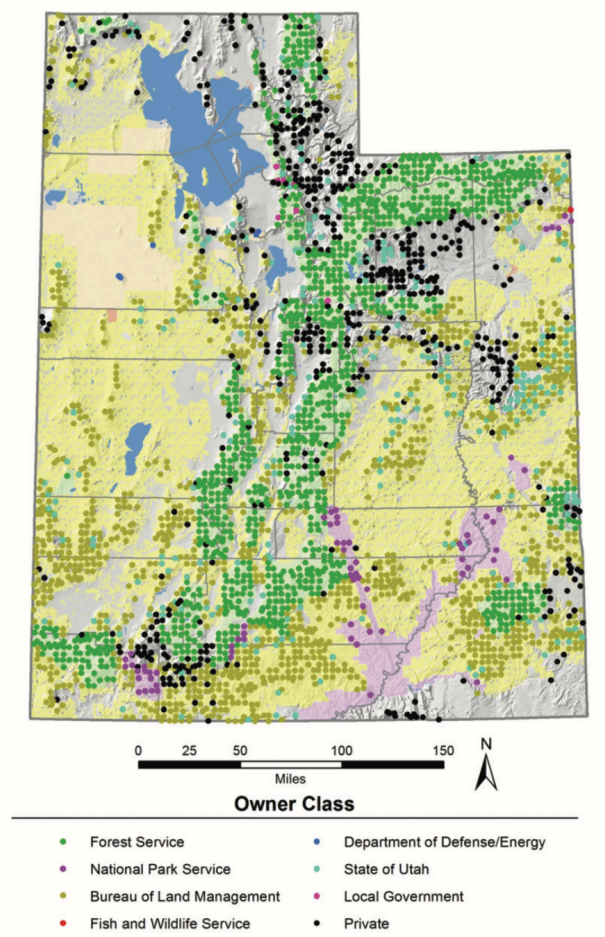


Figure 3—Distribution of inventory plots on forest land by owner class, Utah, 2003–2012. (Note: plot locations are approximate; some plots on private land were randomly swapped.)

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

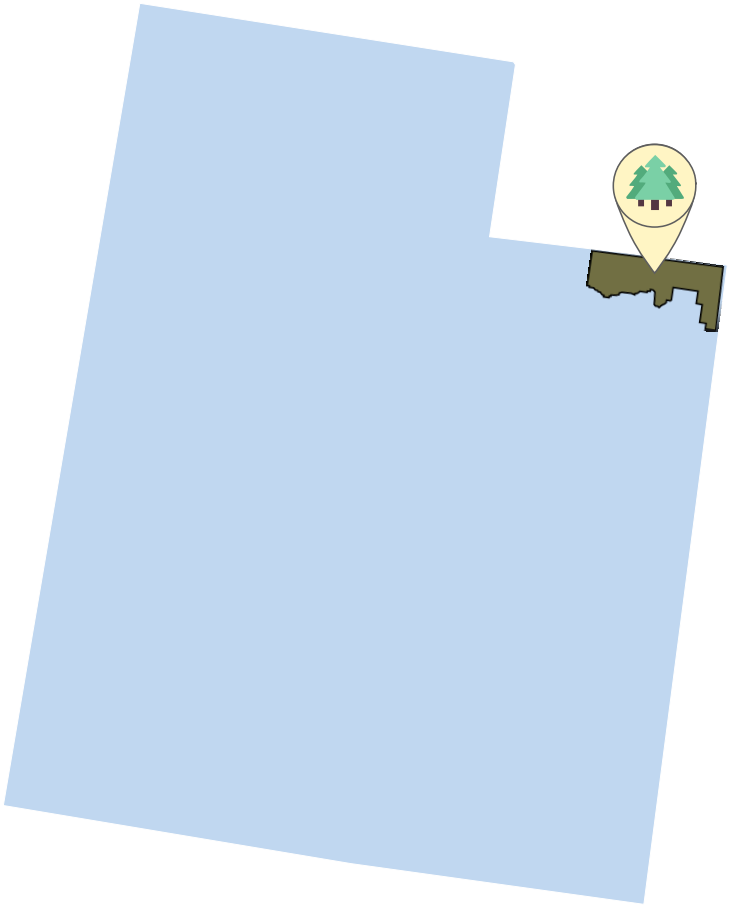
Table B16—Net volume of live trees (at least 5.0 inches d.b.h./d.r.c.), in million cubic feet, on forest land by forest-type group and stand origin, Utah, 2003-2012.

Forest-type group	Stand origin		All forest land
	Natural stands	Artificial regeneration	
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.

DAGGETT COUNTY, UT

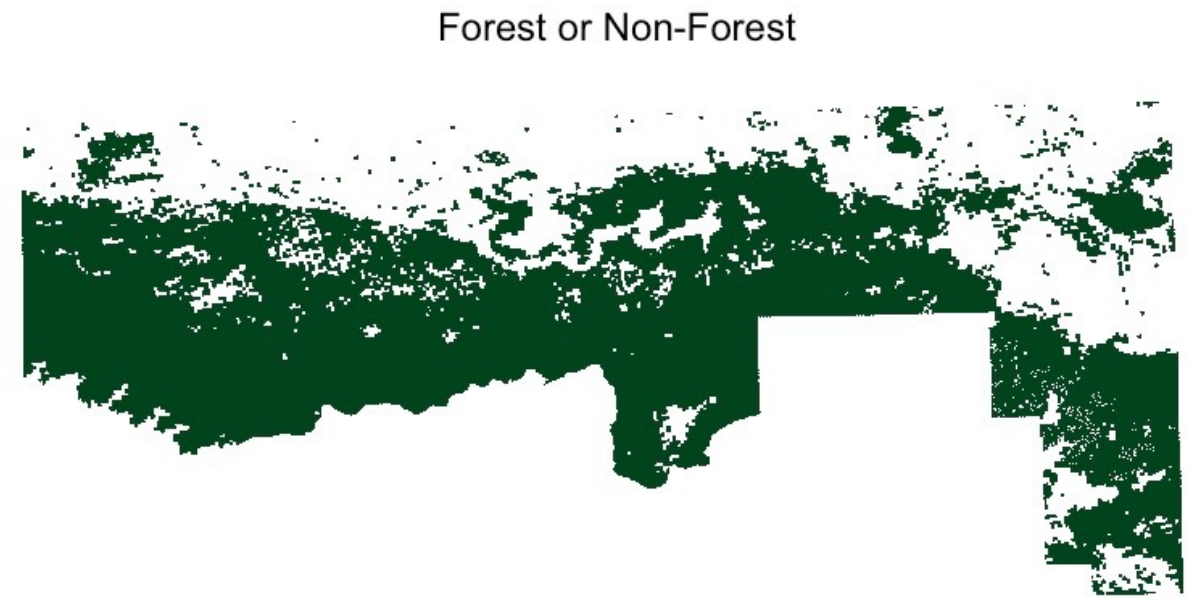
- County is the smallest estimation unit for FIA.
  - Many Forest attributes are estimated.
    - EX: average trees per acre
- Over a 10 year period using a randomized systematic sample, the FIA field crews visit 80 ground plots and collect data.
  - On each map, we have over 2 MILLON pixels of data!



What assisting model should we use?

ESTIMATING FOREST ATTRIBUTES

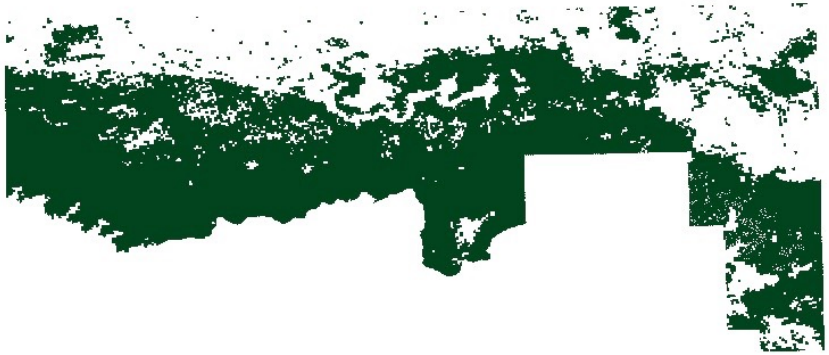
FIA currently uses only one auxiliary variable.



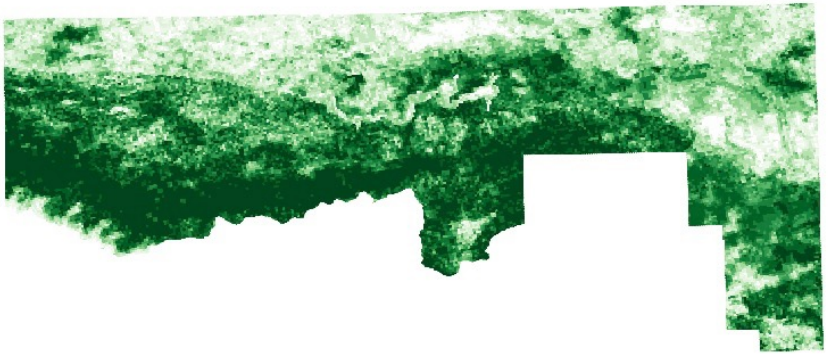
ESTIMATING FOREST ATTRIBUTES

But, FIA has access to many additional variables.

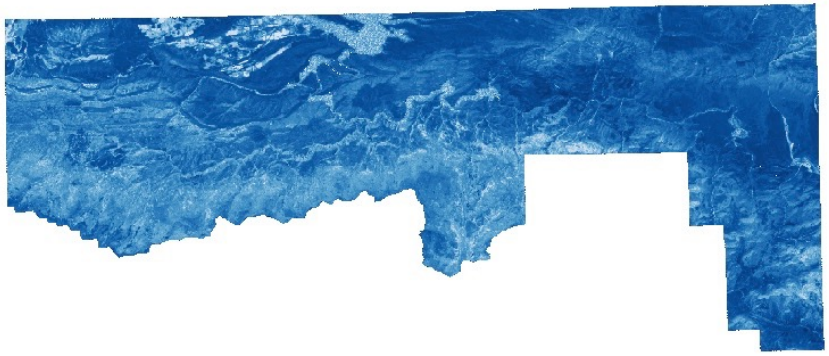
Forest or Non-Forest



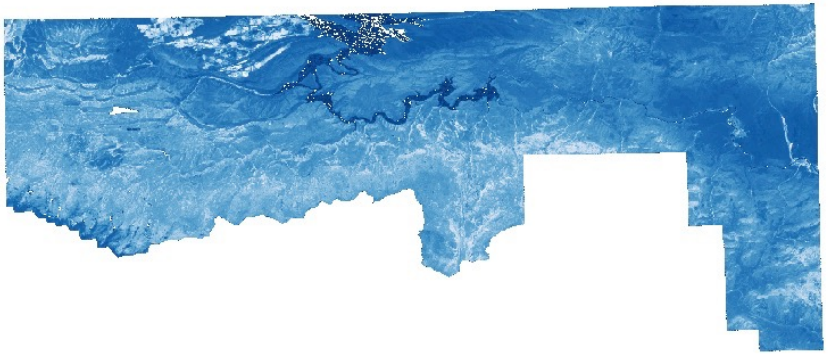
Probabilty of Forest



Normalized Burn Ratio



Normalized Difference Vegetation Index



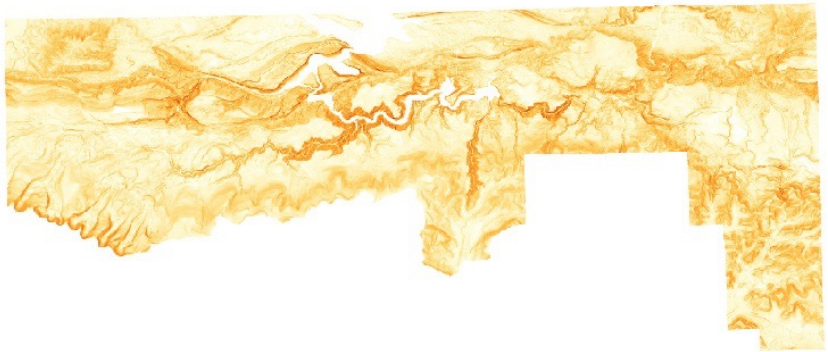
ESTIMATING FOREST ATTRIBUTES

But, FIA has access to many additional variables.

Elevation

Eastness

Slope



ESTIMATING FOREST ATTRIBUTES

What assisting model should we use?

- FIA has access to many additional variables but some might provide redundant or not useful information about the study variable.
- FIA needs to estimate hundreds of forest attributes.
- Want a simple model that can be applied to all attributes.
- Use linear regression with model selection!

LINEAR REGRESSION

- **Model:**

$$y_i = m(\mathbf{x}_i) + \epsilon_i$$
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$
$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

where ϵ_i are independent random variables with variance σ^2.

- Survey-weighted least squares coefficient estimates:

$$\hat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i \in s} \pi_i^{-1} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 \right\}$$

SURVEY WEIGHTED LASSO

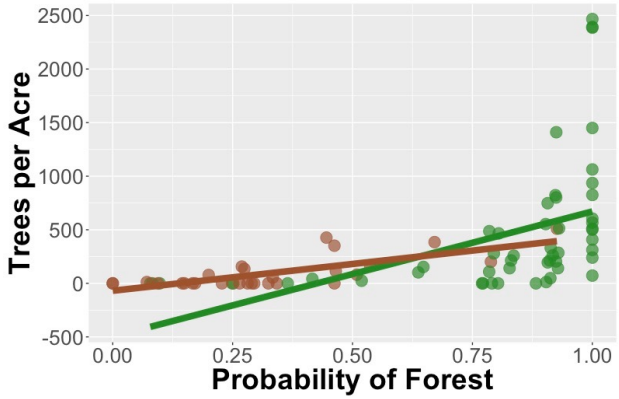
- Get new coefficient estimates by adding Tibshirani's (1996) LASSO penalty:

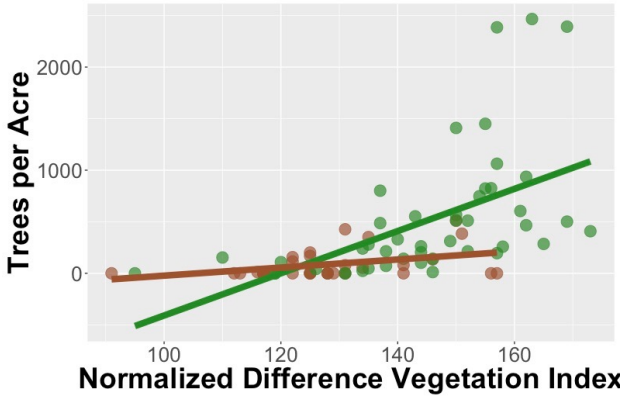
$$\hat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i \in s} \pi_i^{-1} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

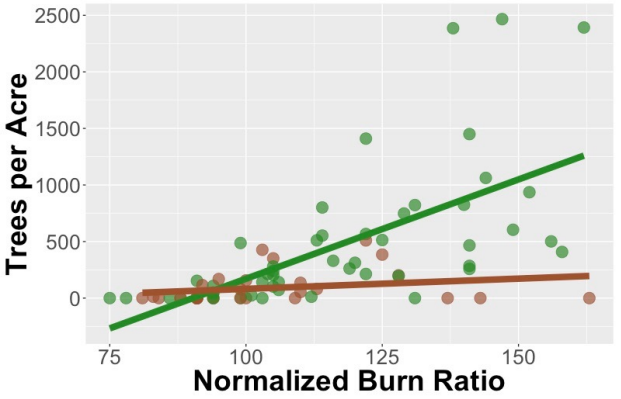
- Plug the LASSO coefficient estimates into the model assisted estimator:

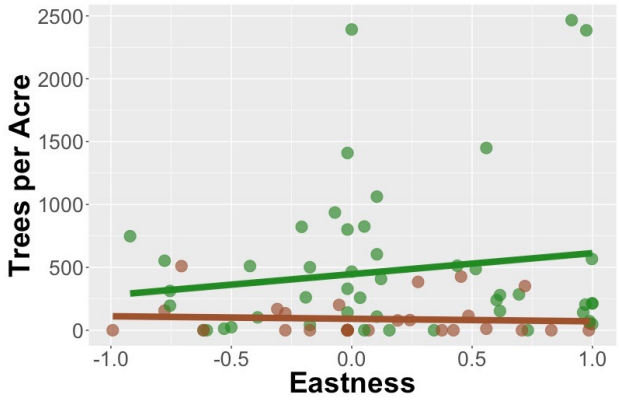
$$\hat{t}_y = \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s + \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s}{\pi_i}$$

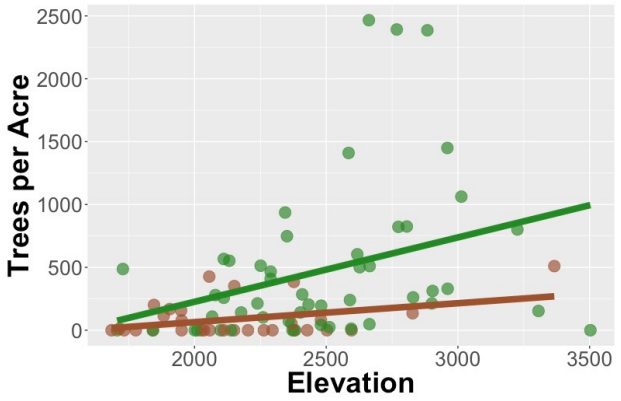
BACK TO DAGGETT COUNTY, UTAH

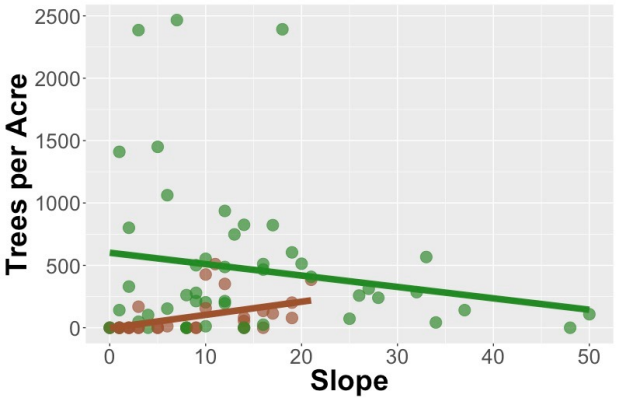










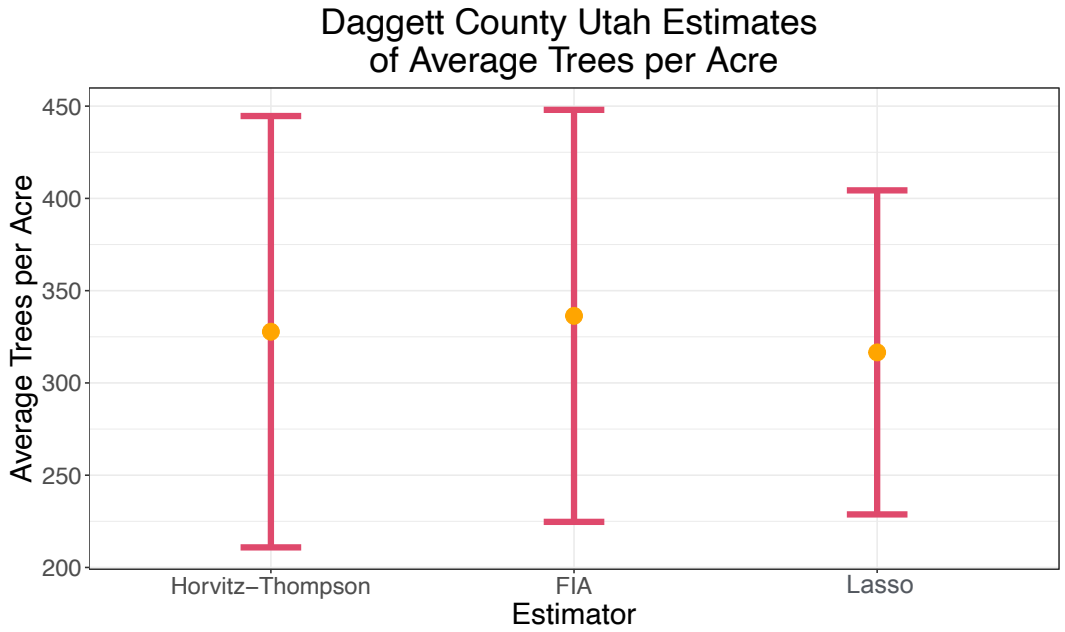


Non-Forest Forest

ESTIMATING FOREST ATTRIBUTES VIA THE LASSO

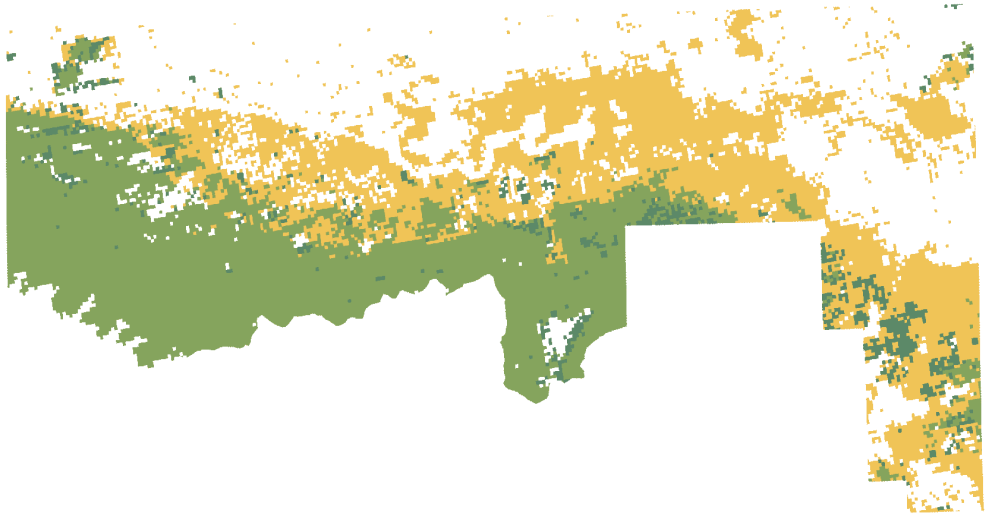
- Coefficients selected:

Coefficients	Values
(Intercept)	406.70
NDVI	118.78
Slope	-4.13
NBR	156.24
Elevation	41.58
Slope:Forest	-8.83

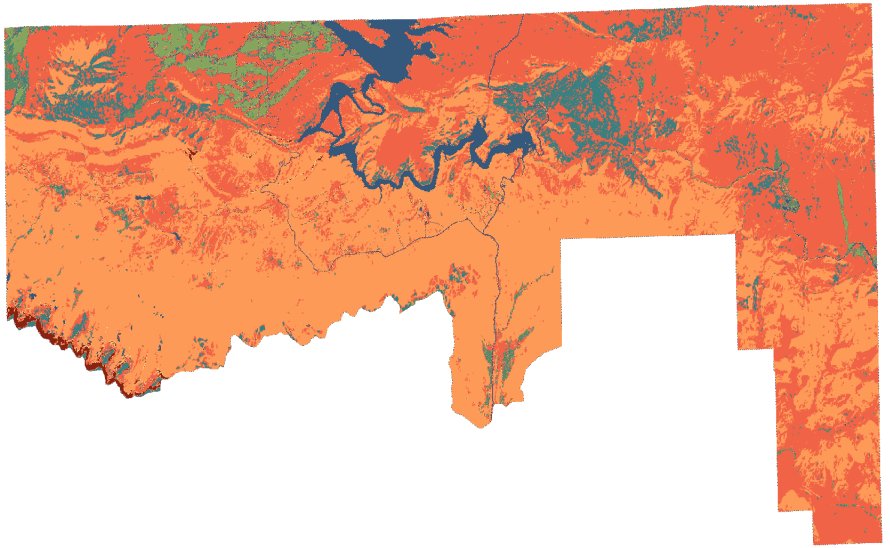


WHAT ABOUT WHEN THE AUXILIARY DATA ARE CATEGORICAL?

Forest Groups



Land Cover Classes



ESTIMATING OCCUPATIONAL STATISTICS

- The US Bureau of Labor Statistics produces statistics related to labor economics.
- For many of their surveys, the population of interest is establishments in the U.S.
  - Occupational Employment Statistics (OES) estimates employment and wage data.
  - EX: Total number of bartenders

ESTIMATING OCCUPATIONAL STATISTICS

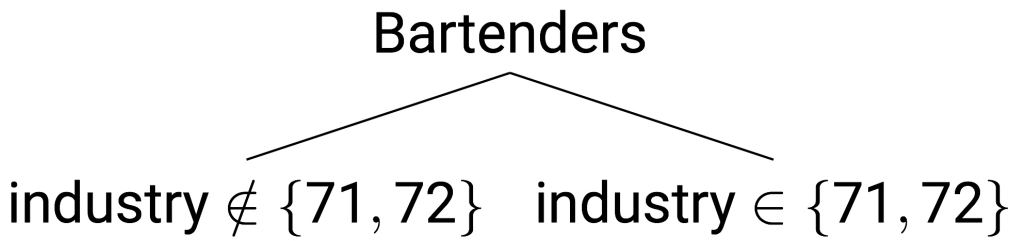
- BLS has access to the Quarterly Census of Employment and Wages (QCEW) for all US establishments!
- QCEW includes useful information on:
  - Size class
  - Geographic information
  - Industry type
  - Whether or not it is a multi-establishment firm

What assisting model should we use?

Why should we not use linear regression?

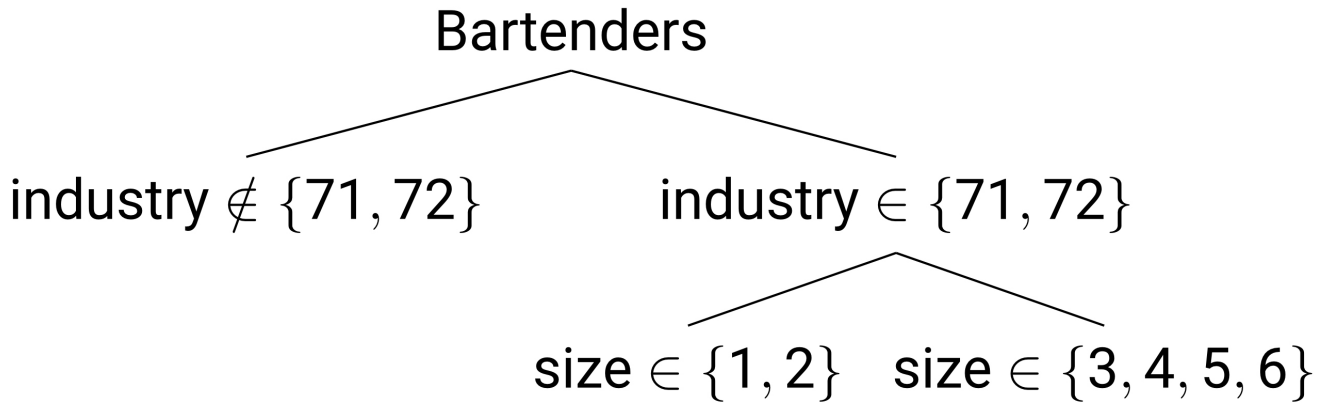
REGRESSION TREES

- Recursively splits the sample into two disjoint groups based on a predictor variable.



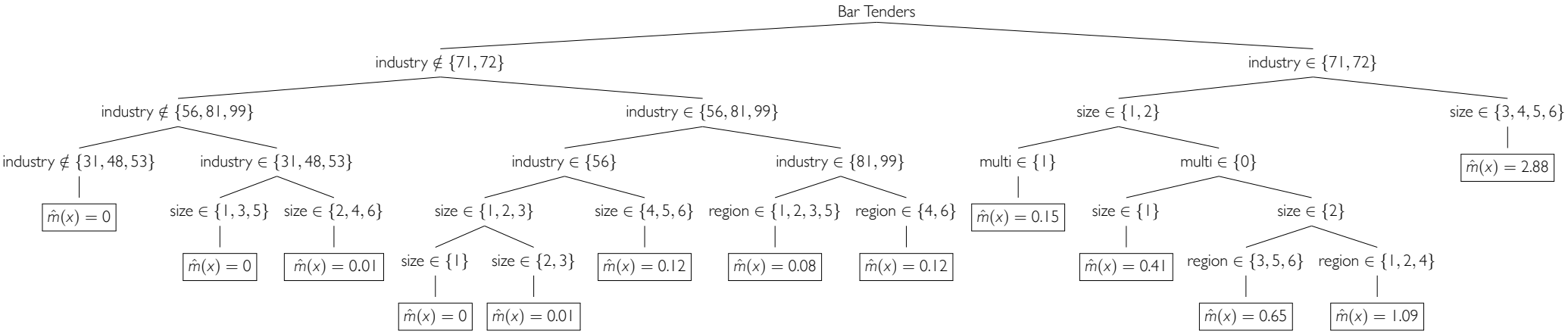
REGRESSION TREES

- Recursively splits the sample into two disjoint groups based on a predictor variable.



REGRESSION TREES

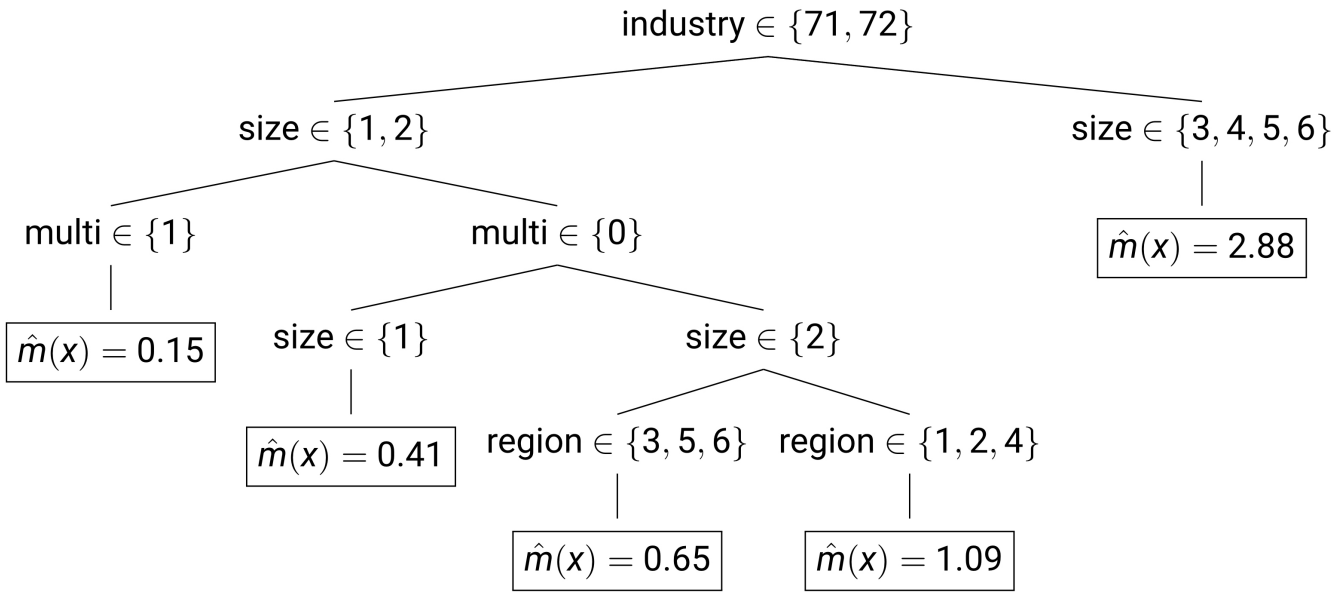
- Recursively splits the sample into two disjoint groups based on a predictor variable.



- Stops when additional splits are no longer that much more predictive.

REGRESSION TREES

- Recursively splits the sample into two disjoint groups based on a predictor variable.



- At each end node, estimate study variable with

$$\hat{m}(x) = \frac{1}{\hat{N}_k} \sum_{i \in s} \frac{y_i I(i \in \text{ node } k)}{\pi_i} \qquad \text{where} \quad \hat{N}_k = \sum_{i \in s} \frac{I(i \in \text{ node } k)}{\pi_i}$$

REGRESSION TREES

- Plug the regression tree estimates into the model assisted esimator:

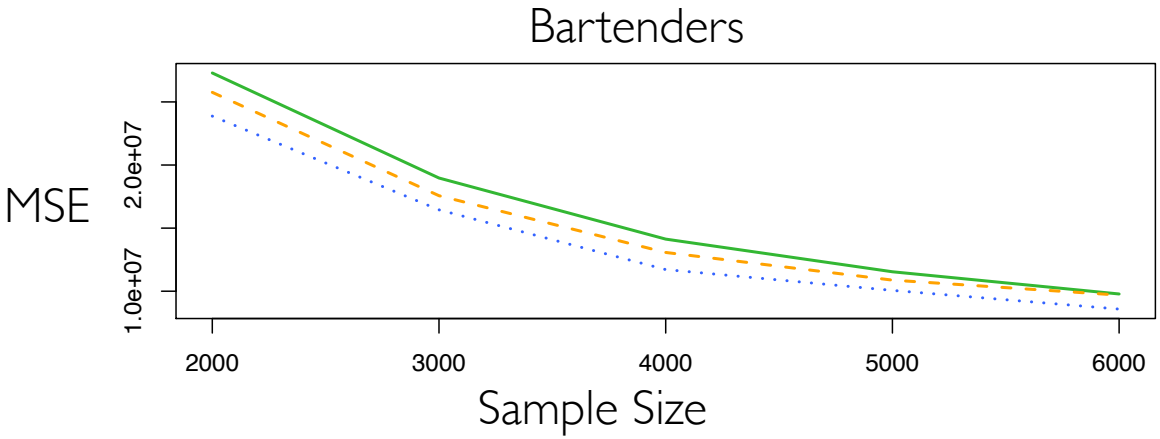
$$\hat{t}_y = \sum_{i \in U} \hat{m}(x_i) + \sum_{i \in s} \frac{y_i - \hat{m}(x_i)}{\pi_i}$$

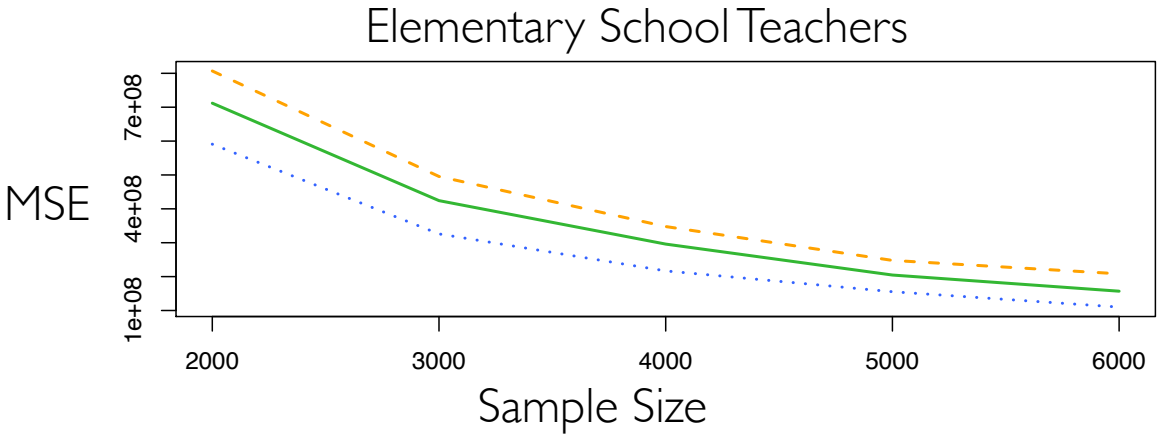
SIMULATION STUDY

- Took 1000 repeated samples from a set of 187,115 establishments
- Compared the empirical mean squared error (MSE)

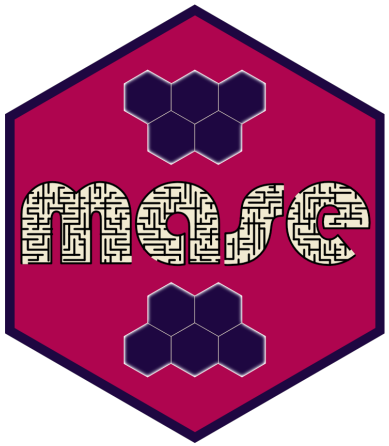
$$\frac{1}{1000}\sum_{b=1}^{1000}(\hat{t}_y^{(b)} - t_y)^2$$

- Horvitz-Thompson estimator
- Step-wise Linear regression estimator
- Regression tree estimator





R PACKAGE



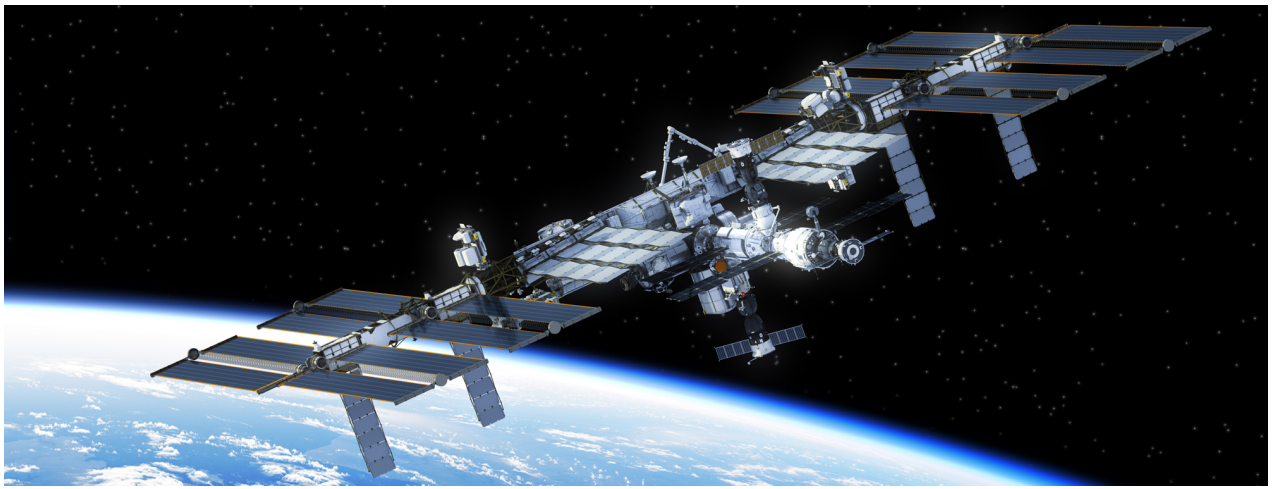
- mase: Model-Assisted Survey Estimation
- Contains the estimators discussed in today's talk (and more!)
- Co-wrote and updated with several students
  - Josh Yamamoto, Becky Tang, George Zhu, Shirley Cheung, and Sida Li

MODEL-ASSISTED ESTIMATION WITH MACHINE LEARNING TOOLS

- Adapting machine learning tools to survey estimation is a ripe area of research.
  - Interesting new models to explore!

MODEL-ASSISTED ESTIMATION WITH MACHINE LEARNING TOOLS

- Adapting machine learning tools to survey estimation is a ripe area of research.
  - Interesting new models to explore!
  - Cool data sources!



MODEL-ASSISTED ESTIMATION WITH MACHINE LEARNING TOOLS

- Adapting machine learning tools to survey estimation is a ripe area of research.
  - Interesting new models to explore!
  - Cool data sources!
  - Engaged undergraduate collaborators!

REFERENCES

- Basil, M. R. K., Huque, S., McConville, K. S., Moisen, G. G., and T. S. Frescino. Creating Homogeneous Landfire Vegetation Classes for Forest Inventory Applications in the Interior West. Gen. Tech. Rep. U. S. Department of Agriculture, Forest Service, Southern Research Station, In Press.
- Breidt, F. J., G. Claeskens, and J. D. Opsomer. Model-assisted estimation for complex surveys using penalised splines. Biometrika, 92:831–846, 2005.
- Breidt, F. J. and J. D. Opsomer. Model-assisted survey estimation with modern prediction techniques. Statistical Science, To Appear, 2017.
- Breidt, F. J. and J. D. Opsomer. Local polynomial regression estimators in survey sampling. Annals of Statistics, 28:1026–1053, 2000.
- Goga, C. Variance reduction in surveys with auxiliary information: a nonparametric approach involving regression splines. The Canadian Journal of Statistics/La revue canadienne de statistique, 33(2), 163-180, 2005.
- Horvitz, D. G., and Thompson, D. J. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association, 47(260), 663-685, 1952.
- Lehtonen, R. and Veijanen, A. Logistic generalized regression estimators. Survey Methodology 24, 51-55, 1998.
- McConville, Moisen, G. G, Frescino, T. S. A Tutorial in Model-Assisted Estimation with Application to Forest Inventory. Forests. 11:2, 244.
- McConville, K. S., Breidt, F. J., Lee, T., & Moisen, G. G. Model-assisted survey regression estimation with the lasso. Journal of Survey Statistics and Methodology, 5(2), 131-158, 2017.
- McConville, K. S. and F. J. Breidt. Survey design asymptotics for the model-assisted penalised spline regression estimator. Journal of Nonparametric Statistics, 25:745–763, 2013.
- McConville, K. S. and Toth, D. Automated selection of post-strata using a model-assisted regression tree estimator. Scandinavian Journal of Statistics, 42, 2: 389–413, 2019.
- Montanari, G. E. and M. G. Ranalli. Nonparametric model calibration estimation in survey sampling. Journal of the American Statistical Association, 100(472):1429–1442, 2005.
- Rintoul, M. A., Maebius, S, Alvarado, E, Lloyd-Damnjanovic, A., Toyohara, M., McConville, K. S., Moisen, G. G., and T. S. Frescino. An Alternative Post-Stratification Scheme to Decrease Variance of Forest Attributes in the Interior West. Gen. Tech. Rep. U. S. Department of Agriculture, Forest Service, Southern Research Station, In Press.
- Sarndal, C. E., B. Swensson, and J. Wretman. Model-Assisted Survey Sampling. Springer-Verlag, New York, 1992.
- Tibshirani, R. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 267-288, 1996.
- Toth, D. and J. Eltinge. Building consistent regression trees from complex sample data. Journal of the American Statistical Association, 106:1626–1636, 2011.

THANKS FOR LISTENING!

QUESTIONS?