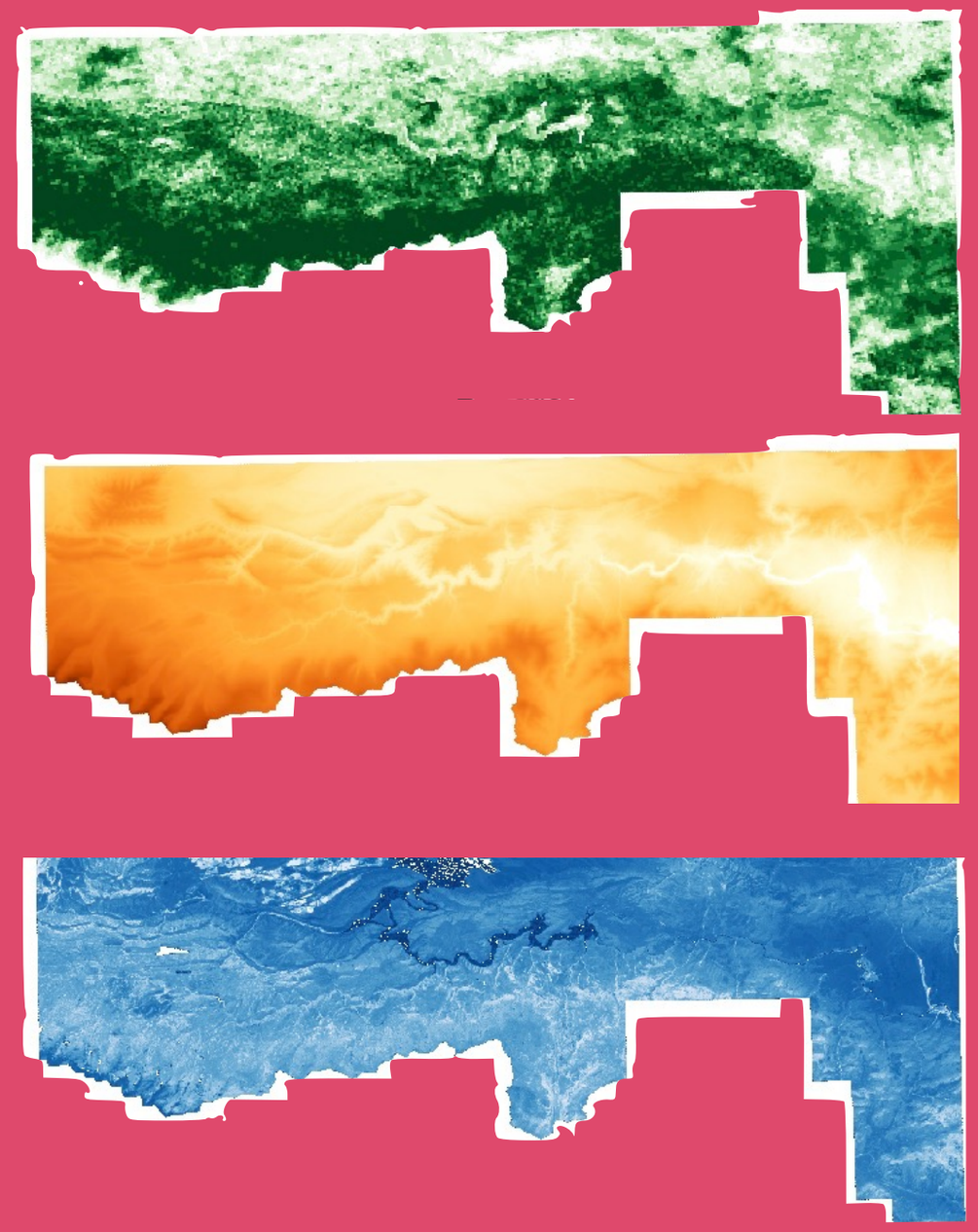
AI MODELS IN SURVEY ESTIMATION



Kelly McConville Reed College

Collaborators:

US Forest Inventory and Analysis Program: Gretchen Moisen, Tracey Frescino

Colorado State University: F. Jay Breidt and UC, Davis: Thomas Lee

US Bureau of Labor Statistics: Daniell Toth

SURVEY ESTIMATION SETUP

- Goal: Construct estimates of finite population quantities.
- Use data collected under a complex sampling design.
- Also use other available data (e.g., big data).
- Combining these data sources can increase the efficiency of our estimates!

ESTIMATION SET-UP

Enumerate the region (e.g., county).

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.

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{18}	y_{19}	y_{20}	y_{21}	y_{22}	y_{23}	y_{24}
y_{25}	y_{26}	y_{27}	y_{28}	y_{29}	y_{30}	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}	y_{36}
y_{37}	y_{38}	y_{39}	y_{40}	y_{41}	y_{42}	y_{43}	y_{44}	y_{45}	y_{46}	y_{47}	y_{48}
y_{49}	y_{50}	y_{51}	y_{52}	y_{53}	y_{54}	y_{55}	y_{56}	y_{57}	y_{58}	y_{59}	y_{60}

● ● ●

y_{N-11}	y_{N-10}	y_{N-9}	y_{N-8}	y_{N-7}	y_{N-6}	y_{N-5}	y_{N-4}	y_{N-3}	y_{N-2}	y_{N-1}	y_N

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

ESTIMATION SET-UP

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

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}
X_{25}	X_{26}	X_{27}	X_{28}	X_{29}	X_{30}	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}
X_{37}	X_{38}	X_{39}	X_{40}	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	X_{46}	X_{47}	X_{48}
X_{49}	X_{50}	X_{51}	X_{52}	X_{53}	X_{54}	X_{55}	X_{56}	X_{57}	X_{58}	X_{59}	X_{60}

● ● ●

X_{N-11}	X_{N-10}	X_{N-9}	X_{N-8}	X_{N-7}	X_{N-6}	X_{N-5}	X_{N-4}	X_{N-3}	X_{N-2}	X_{N-1}	X_N

ESTIMATION SET-UP

A complex sampling design is constructed.

π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}	π_{12}
π_{13}	π_{14}	π_{15}	π_{16}	π_{17}	π_{18}	π_{19}	π_{20}	π_{21}	π_{22}	π_{23}	π_{24}
π_{25}	π_{26}	π_{27}	π_{28}	π_{29}	π_{30}	π_{31}	π_{32}	π_{33}	π_{34}	π_{35}	π_{36}
π_{37}	π_{38}	π_{39}	π_{40}	π_{41}	π_{42}	π_{43}	π_{44}	π_{45}	π_{46}	π_{47}	π_{48}
π_{49}	π_{50}	π_{51}	π_{52}	π_{53}	π_{54}	π_{55}	π_{56}	π_{57}	π_{58}	π_{59}	π_{60}

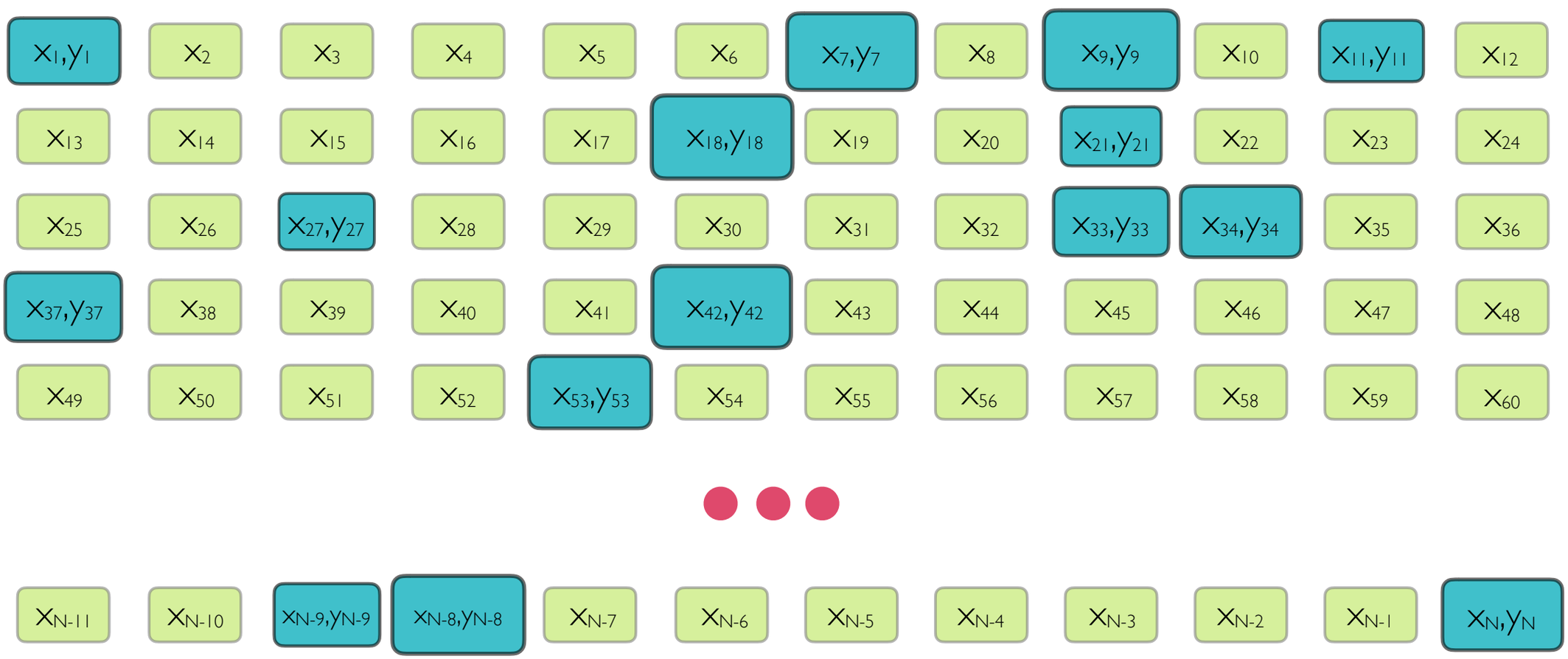
● ● ●

π_{N-11}	π_{N-10}	π_{N-9}	π_{N-8}	π_{N-7}	π_{N-6}	π_{N-5}	π_{N-4}	π_{N-3}	π_{N-2}	π_{N-1}	π_N

$$\pi_i = P(i \in s)$$

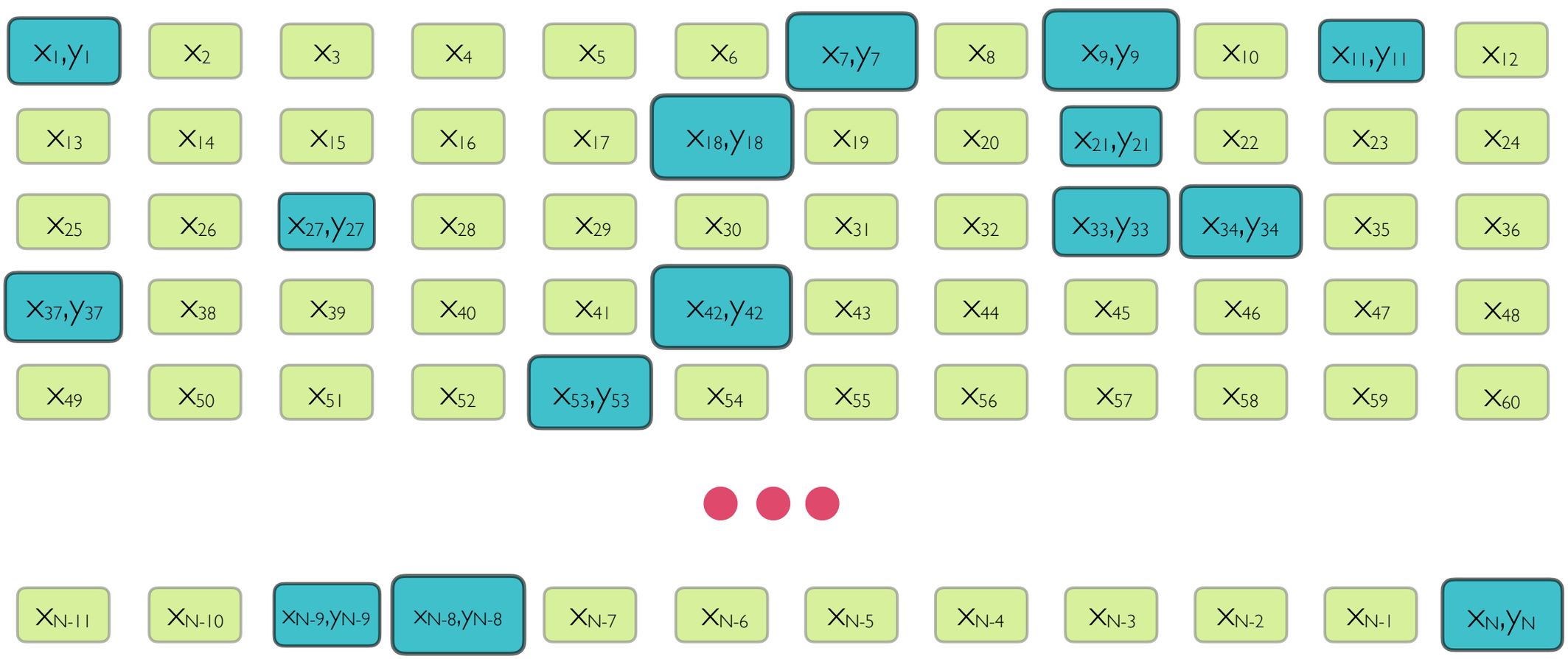
ESTIMATION

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.

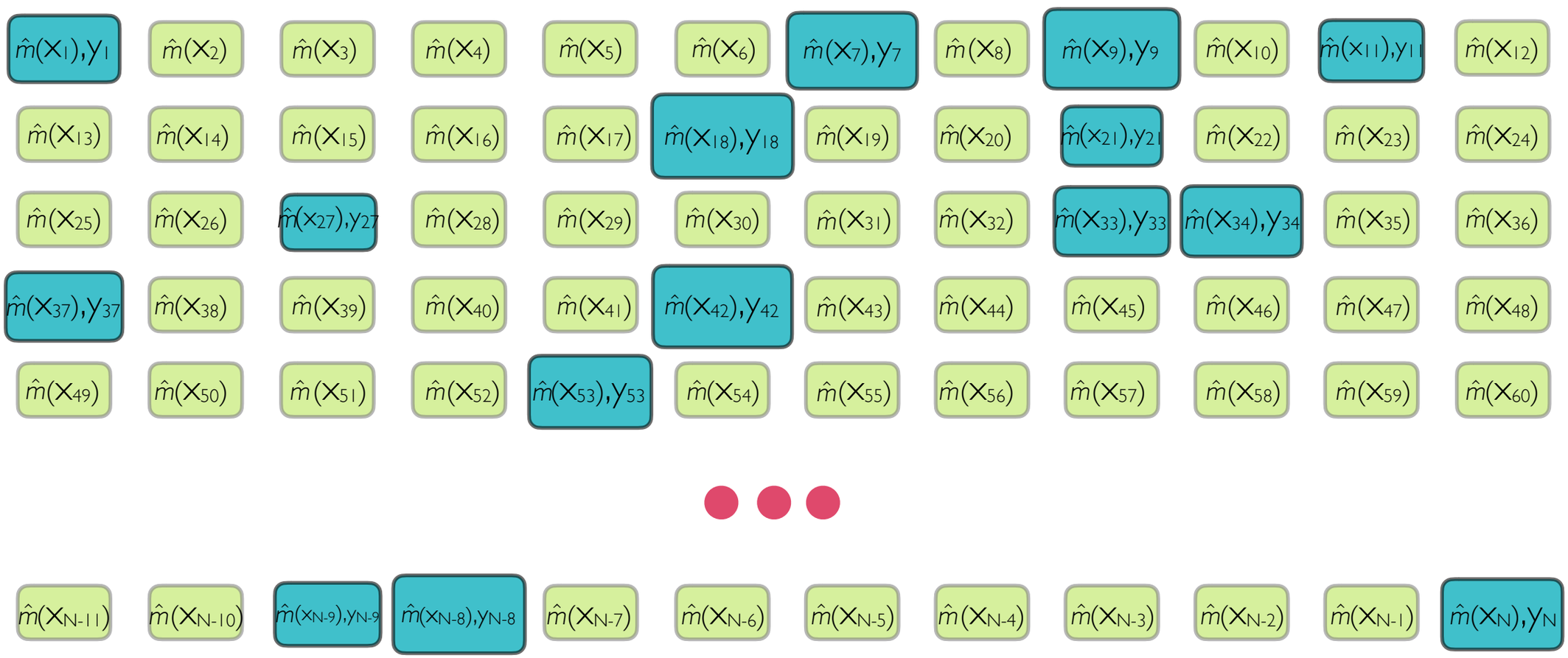


$$\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Horvitz and Thompson (1952)

ESTIMATION

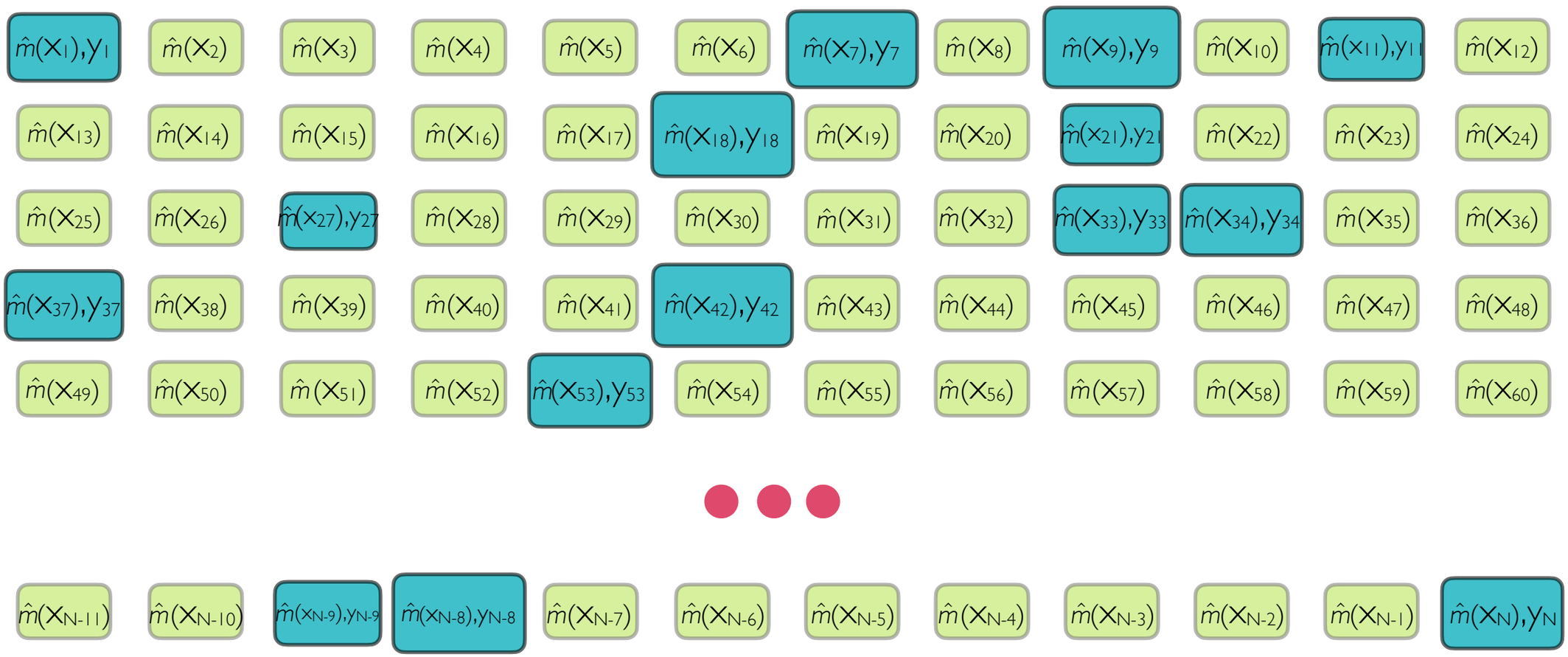
Can use the additional data to predict the study variable.



$\hat{m}(x_i)$ = predicted value for y_i

MODEL-ASSISTED SURVEY ESTIMATION

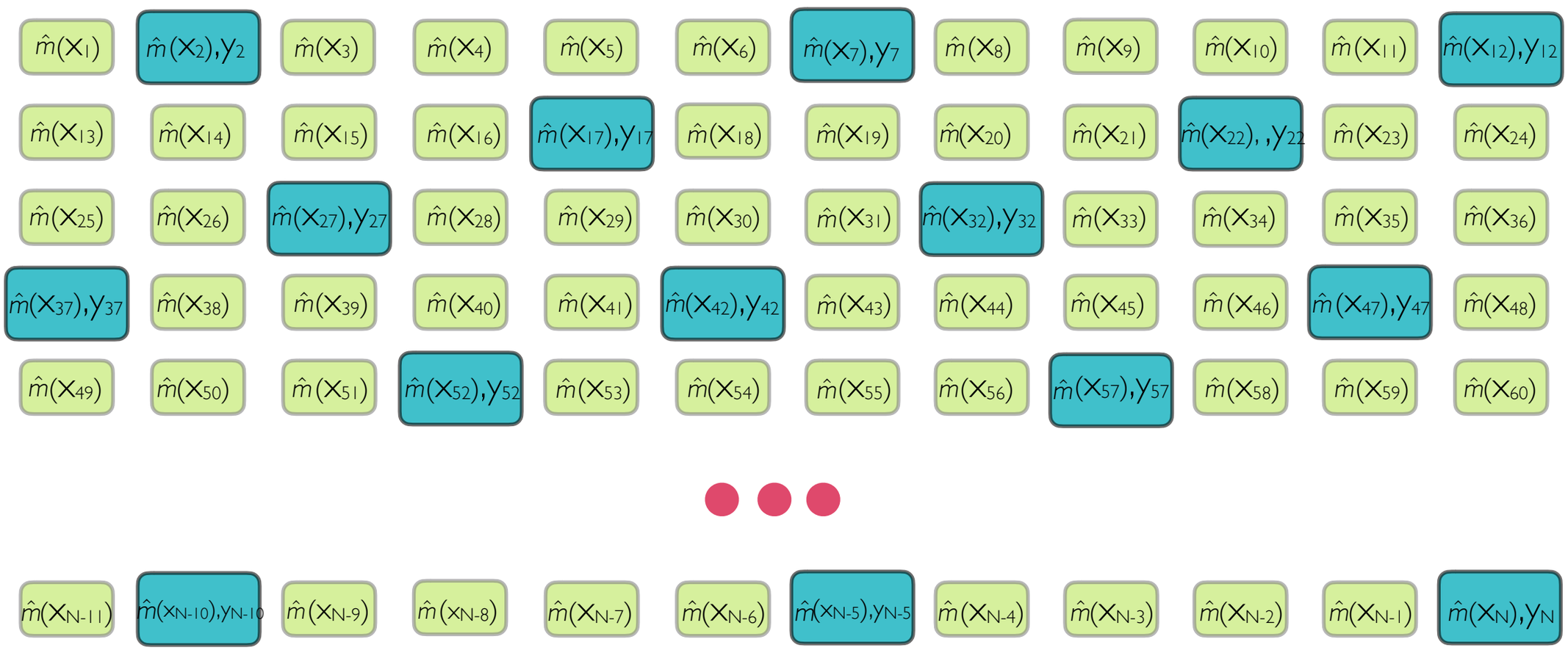
Construct an estimator that is robust to model mis-specification.



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

MODEL-ASSISTED SURVEY ESTIMATION

Need to determine a good assisting model to construct estimator.

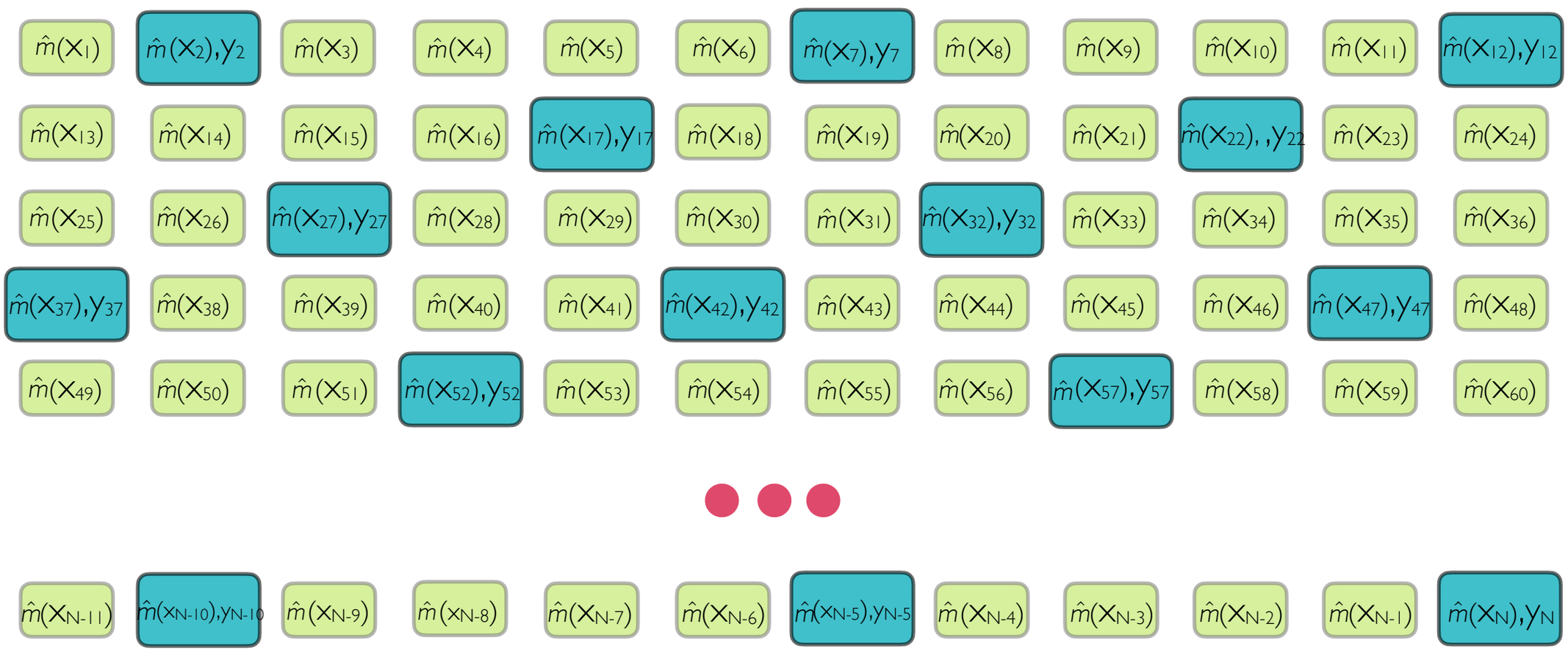


$$\hat{t}_y = \sum_{i \in U} \hat{m}(x_i) + \sum_{i \in s} \frac{y_i - \hat{m}(x_i)}{\pi_i}$$

MODEL-ASSISTED SURVEY ESTIMATION

Need to determine a good assisting model to construct estimator.

What additional data are available?



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

MODEL-ASSISTED ESTIMATOR

- Generalized regression 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 (and therefore narrow confidence intervals)
- But, the **size of the variance** depends on how well the assisting model captures the relationship between the study variable and the additional data.

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.

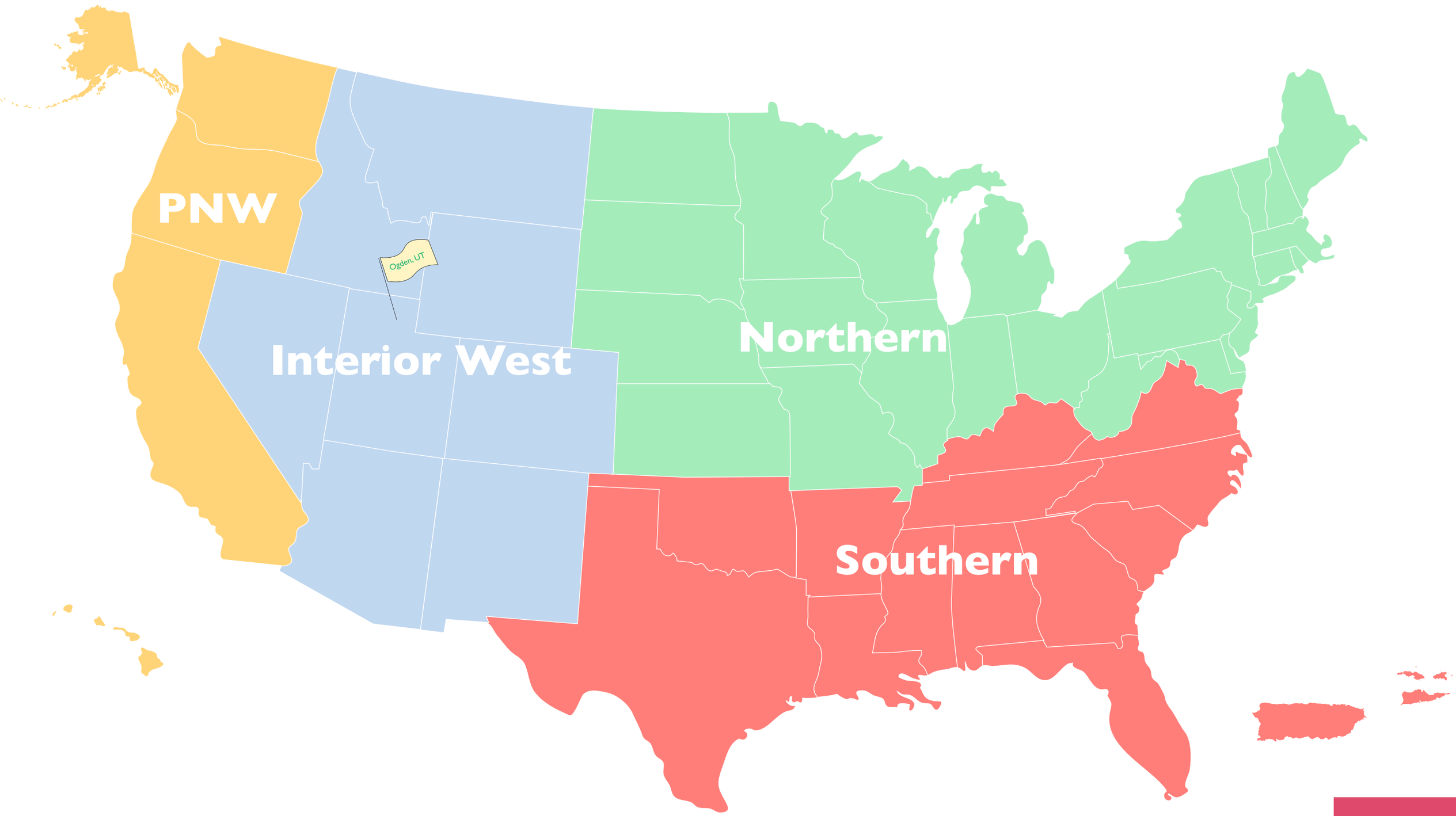
WHICH ASSISTING MODEL SHOULD ONE USE?

- Active research question!
- **Incomplete** list of references:
  - Linear Regression (Cassel, Sarndal, and Wretman 1976)
  - Logistic Regression (Lehtonen and Veijanen 1998)
  - Local polynomial regression (Breidt and Opsomer 2000; Montanari and Ranalli 2005)
  - Penalized Splines (Breidt, Claeskens, and Opsomer 2005; McConville and Breidt 2013)
  - Regression Splines (Goga 2005)
  - Neural Networks (Montanari and Ranalli 2005)

WHICH ASSISTING MODEL SHOULD ONE USE?

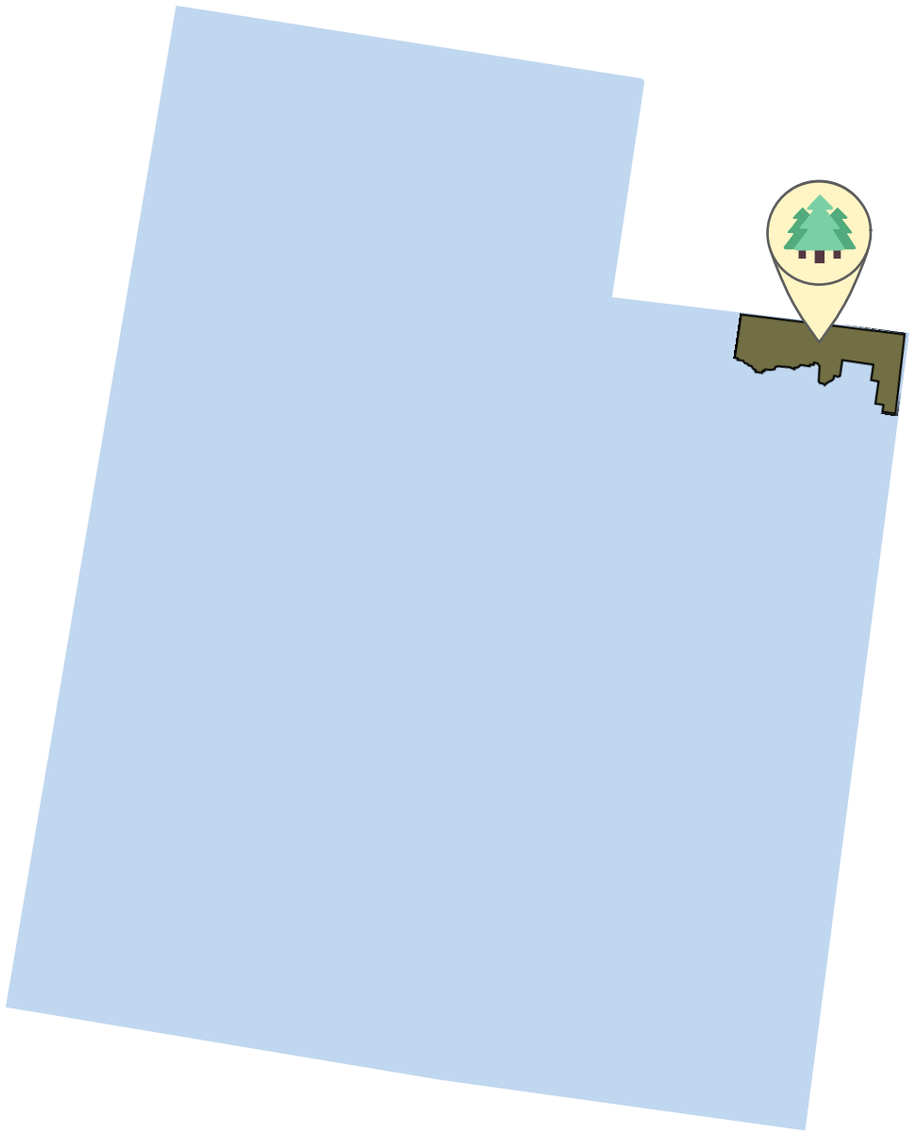
- Look at two examples from my work
  - **Penalized Regression:** Forest attributes in Daggett County, Utah
    - Average number of trees per acre
    - McConville et al (2020)
  - **Regression Trees:** Employment counts for U.S. establishments
    - Total number of bartenders
    - McConville and Toth (2019)

U.S. FOREST INVENTORY AND ANALYSIS (FIA)



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!

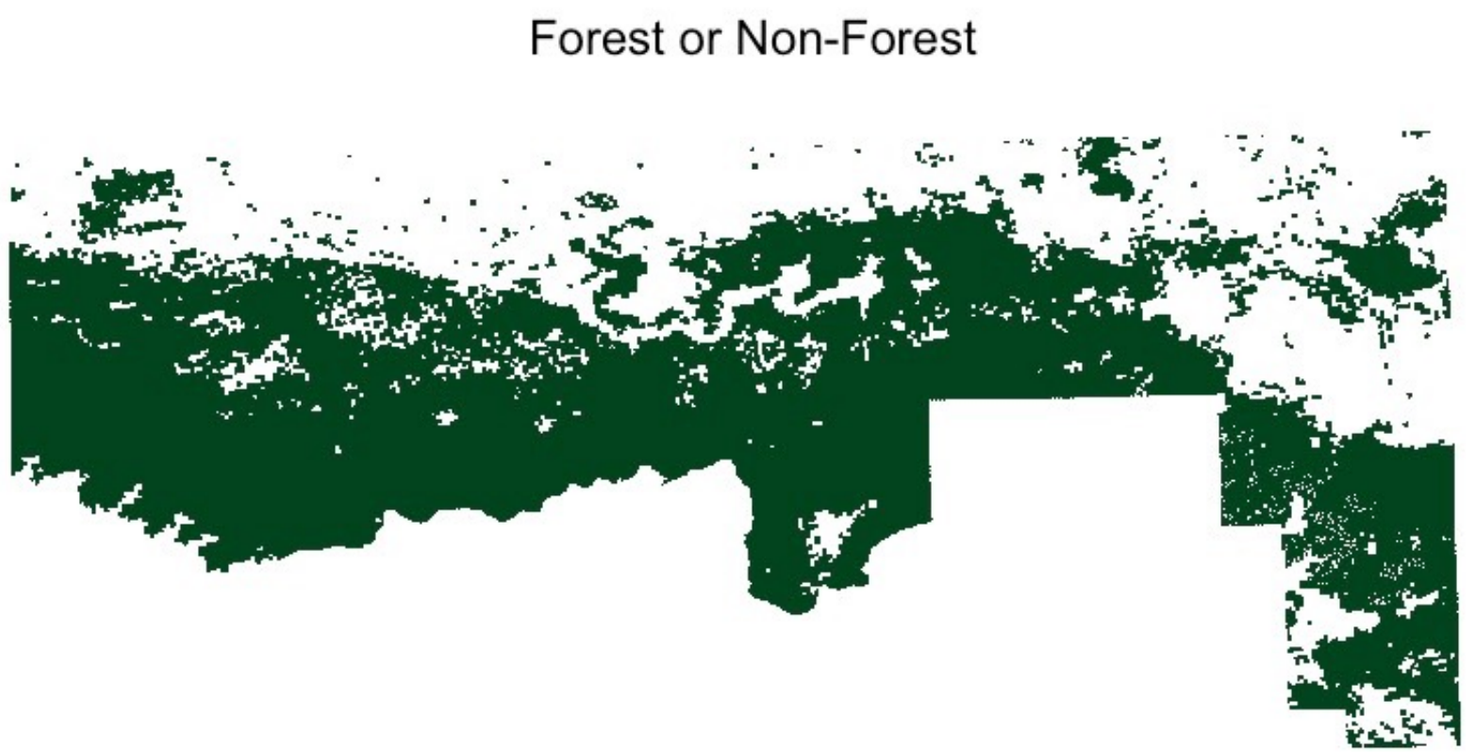


ESTIMATING FOREST ATTRIBUTES

- What assisting model should we use?

ESTIMATING FOREST ATTRIBUTES

FIA currently uses only one auxiliary variable.

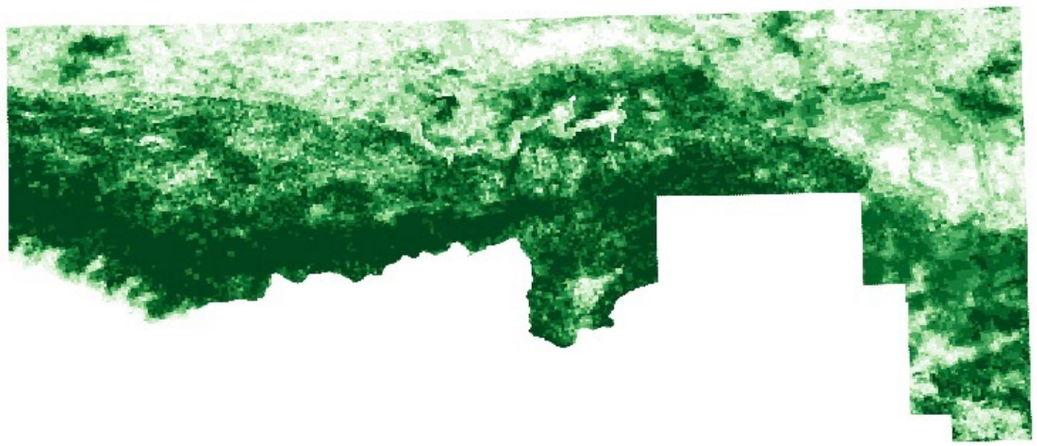


ESTIMATING FOREST ATTRIBUTES

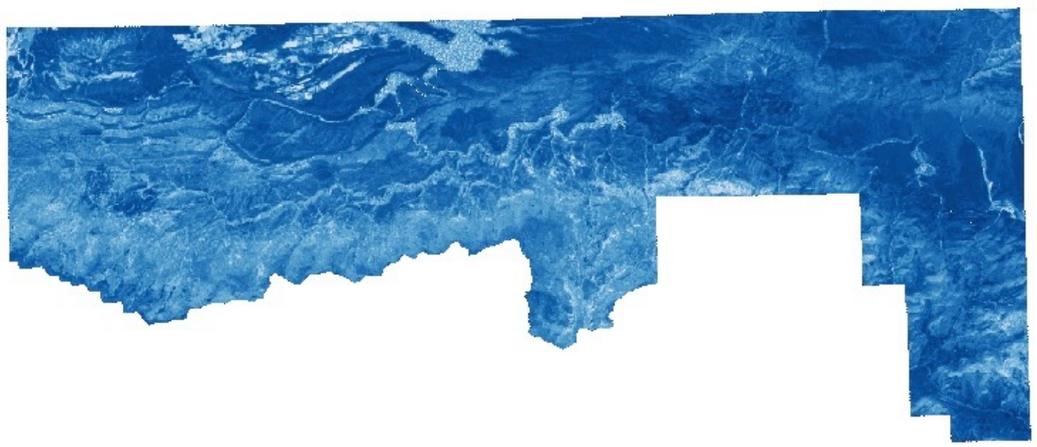
But, FIA has access to many additional variables.

Forest or Non-Forest

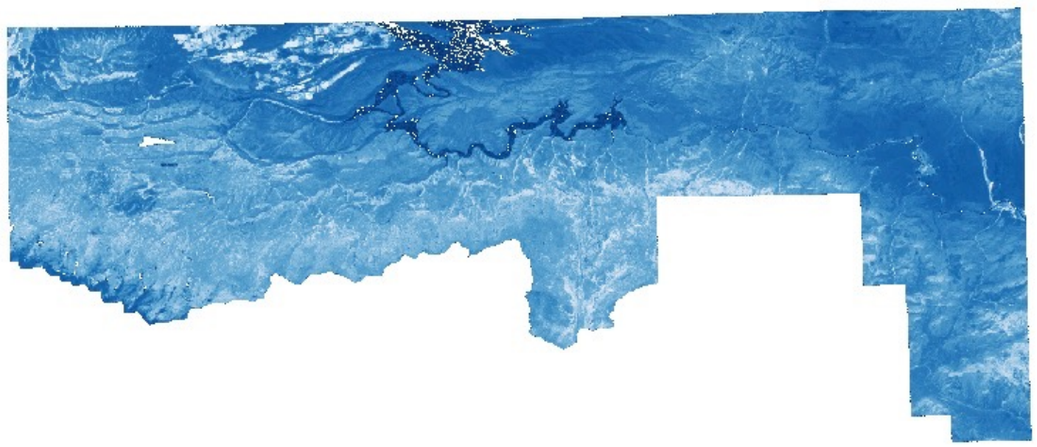
Probabilty of Forest



Normalized Burn Ratio



Normalized Difference Vegetation Index



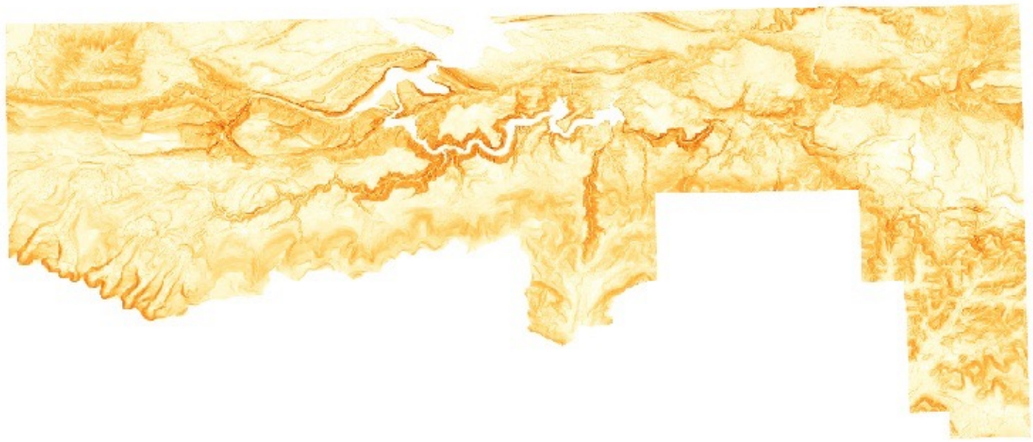
ESTIMATING FOREST ATTRIBUTES

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.
  - Some variables may be extraneous.
- FIA has to estimate hundreds of forest attributes.
  - Want a simple model that can be applied to all attributes.
- **Use linear regression with model selection!**

THE LASSO

- **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.

- Adjust Tibshirani's (1996) LASSO coefficient estimation criterion to account for the sampling design:

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

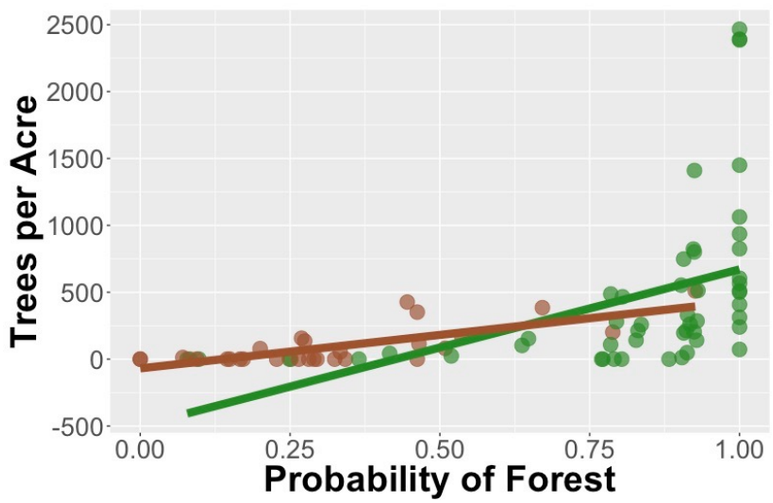
SURVEY WEIGHTED LASSO

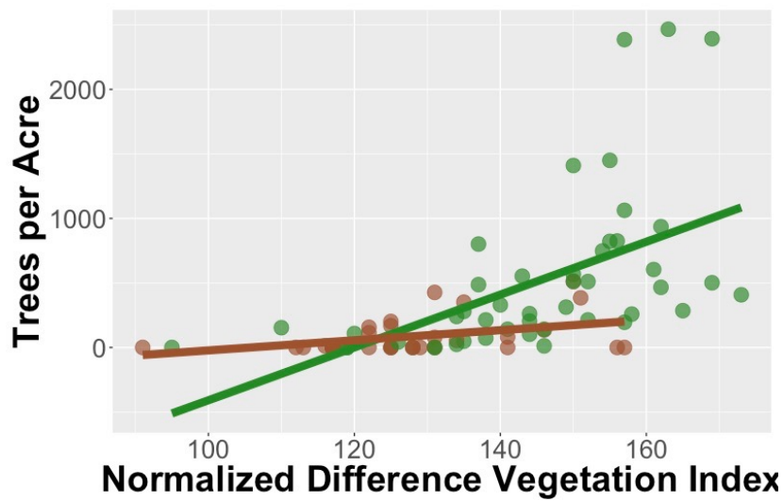
- For Daggett County
  - Six numerical maps and one categorical map

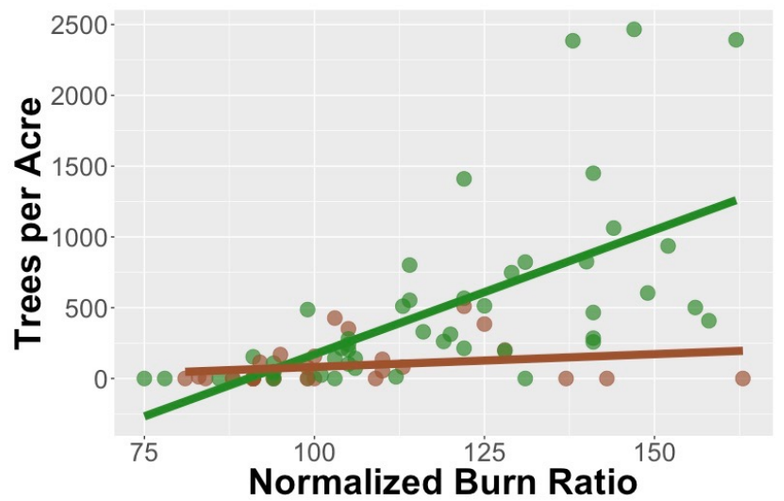
$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_7 x_{i7} + \beta_8 x_{i1} x_{i7} + \ldots + \beta_{13} x_{i6} x_{i7} + \epsilon_i$$

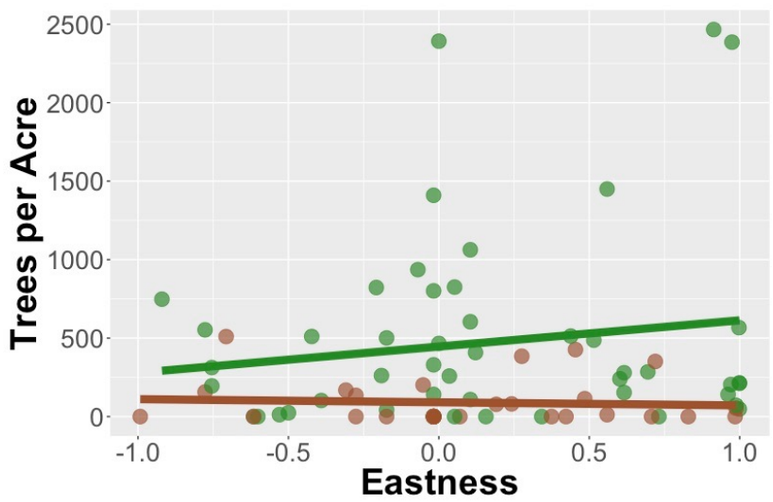
- **Lasso Estimator:** $\hat{t}_y = \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s + \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s}{\pi_i}$

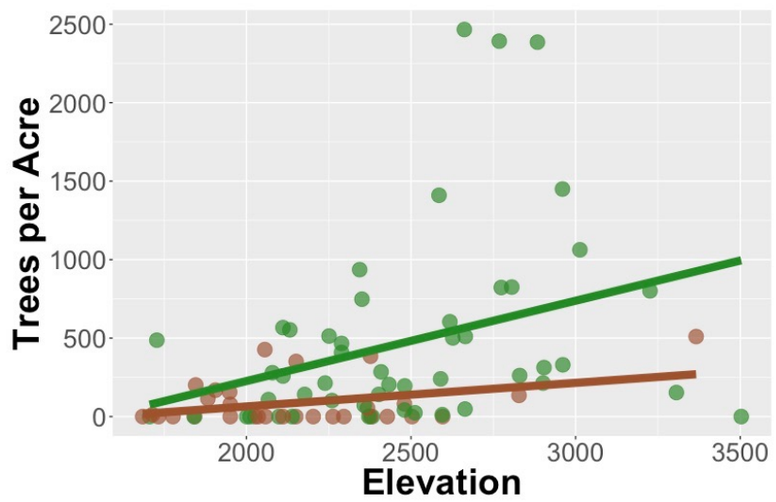
RETURN TO DAGGETT COUNTY

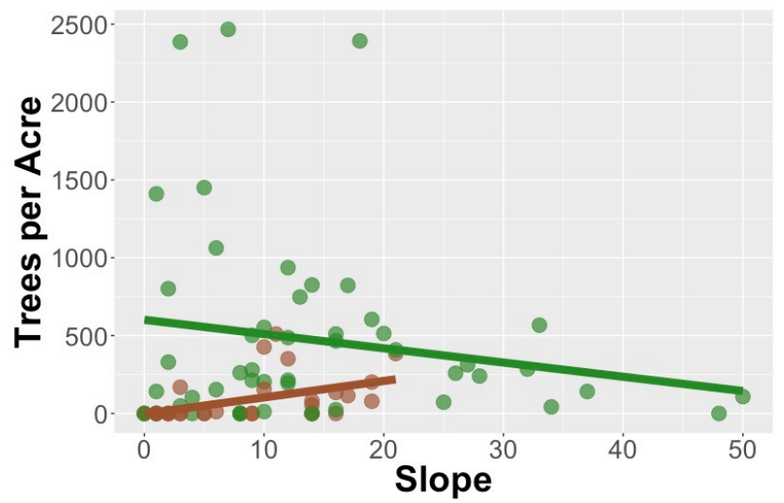












ESTIMATING FOREST ATTRIBUTES VIA THE LASSO

- Coefficients selected:

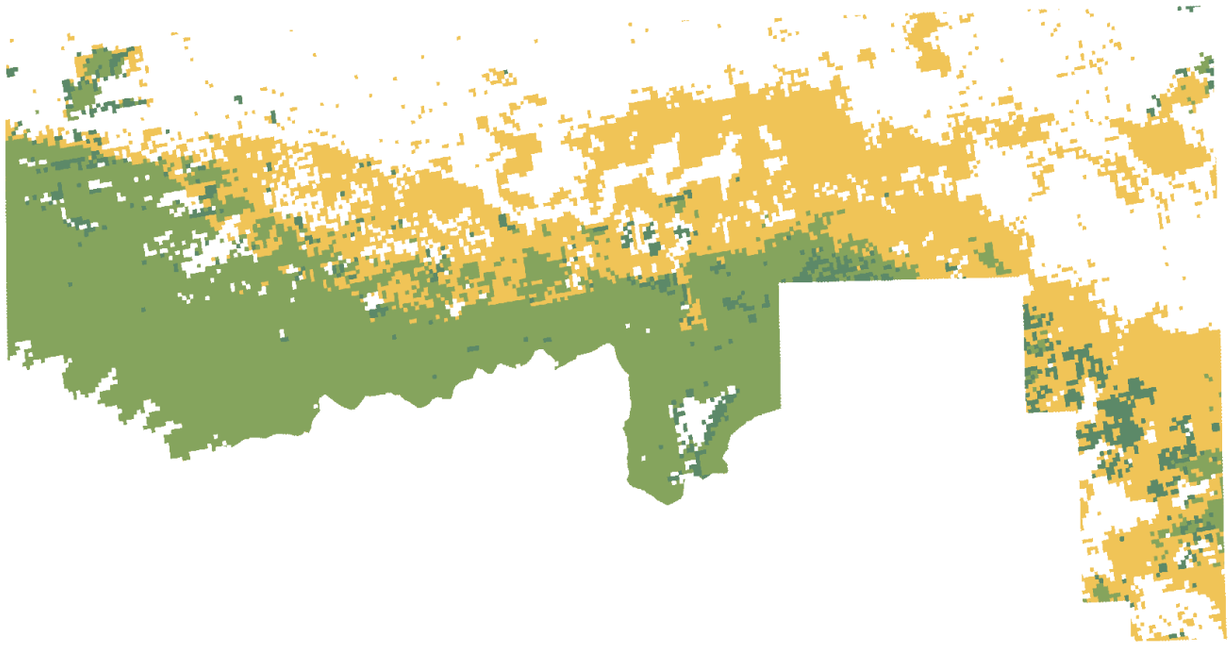
Coefficients	Values
(Intercept)	406.70
NDVI	118.78
Slope	-4.13
NBR	156.24
Elevation	41.58
Slope:Forest	-8.83

- Results:

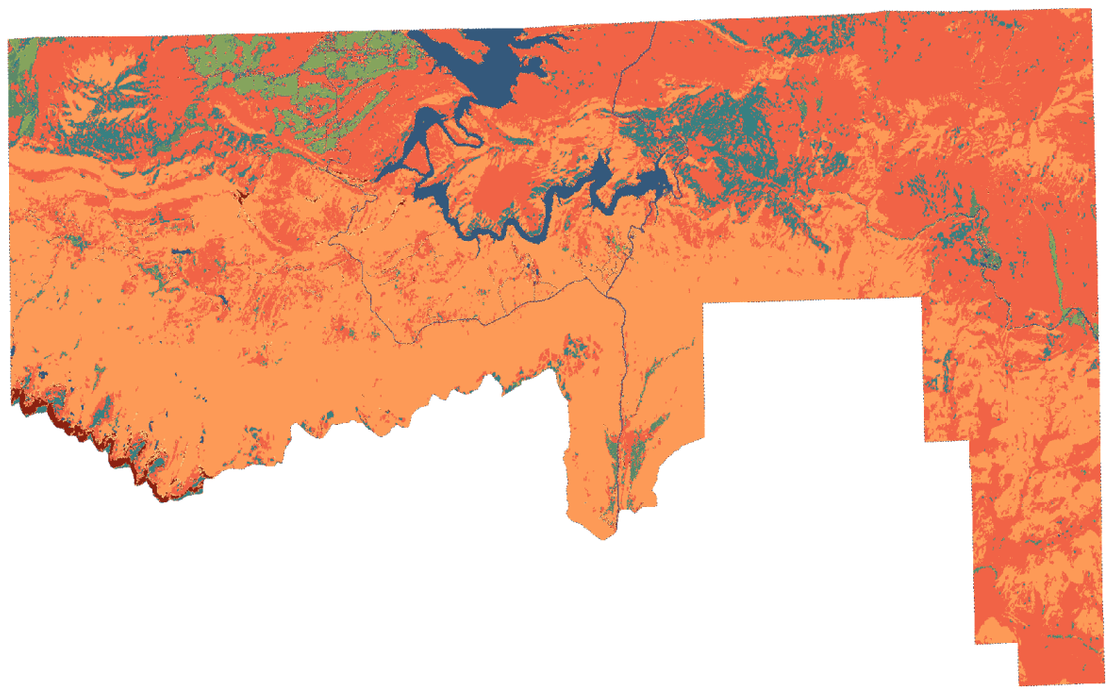
	Trees per Acre	
	Estimate	Standard Error
HT	327.76	58.44
FIA	336.35	55.82
LASSO	316.55	43.90

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

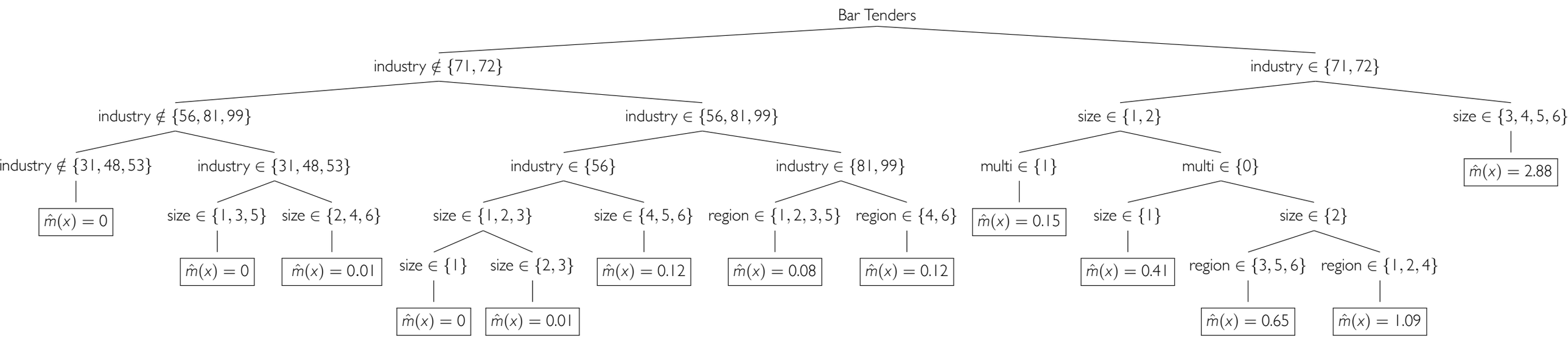
- BLS has access to the Quarterly Census of Employment and Wages (QCEW) **for all US establishments!**
- QCEW includes useful information on:
  - Size class
  - Geographic information
  - Industry type
  - Whether or not it is a multi-establishment firm

ESTIMATING OCCUPATIONAL STATISTICS

- What is the structure of the QCEW data?
    - Mostly categorical auxiliary variables.
    - Some variables have many categories.
        - The relationship between the study variable and the auxiliary data may include complex interactive effects.
- What model should we consider?
    - Survey-weighted regression trees!

REGRESSION TREES

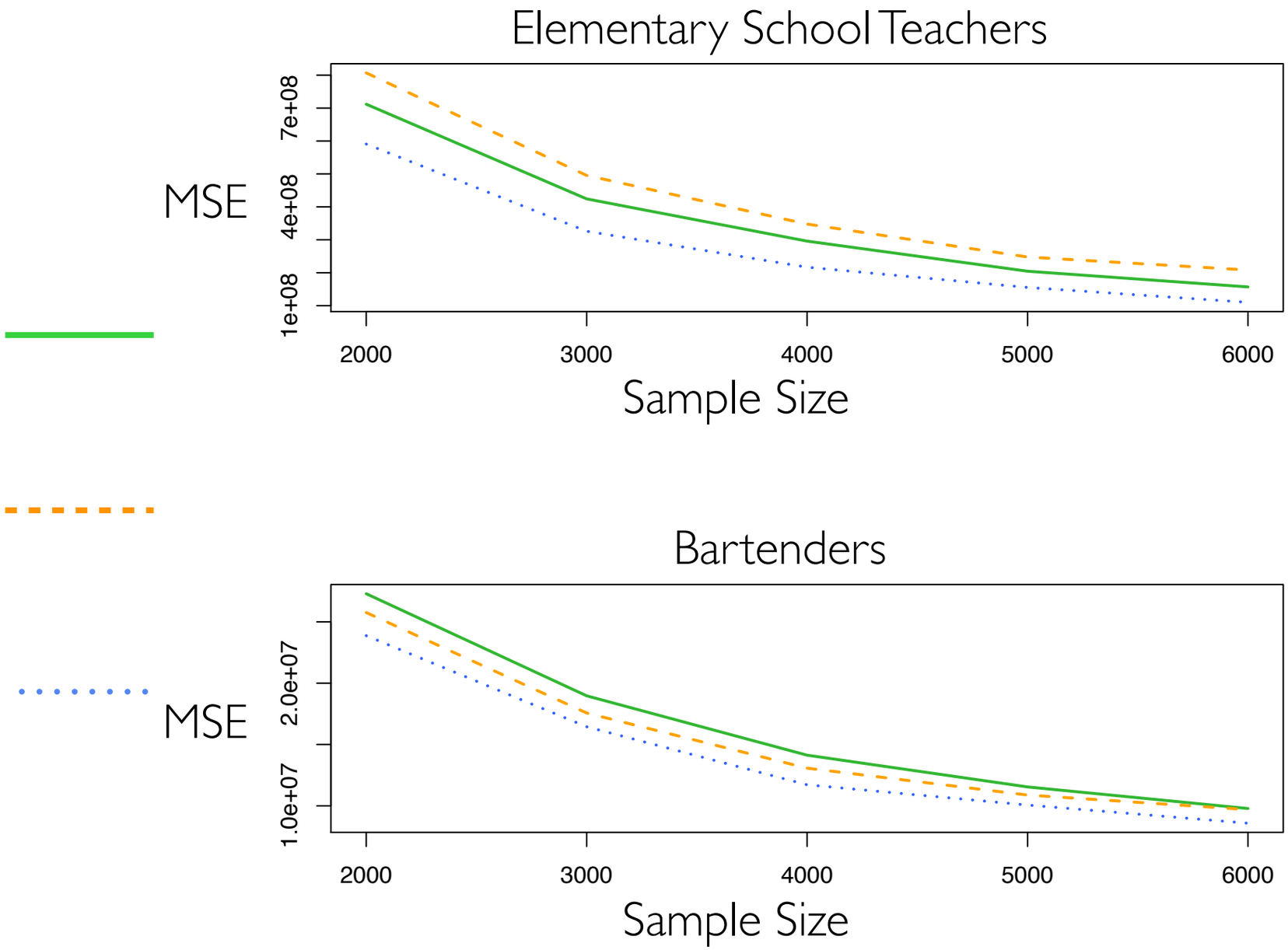
- **Recursively** splits the sample into two disjoint groups based on an auxiliary variable.



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

SIMULATION STUDY

- MSE of estimators for two occupations
  - Horvitz-Thompson estimator
  - Linear regression estimator
  - Regression tree estimator



SURVEY ESTIMATION WITH MACHINE LEARNING TOOLS

- Adapting machine learning tools to survey estimation is a ripe area of research.
- Take your favorite tool and modify it to work for a complex sampling design!
- As the quantity and types of auxiliary data increase, there will be continued interest in using these tools to increase the precision of survey estimators.

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.

QUESTIONS?