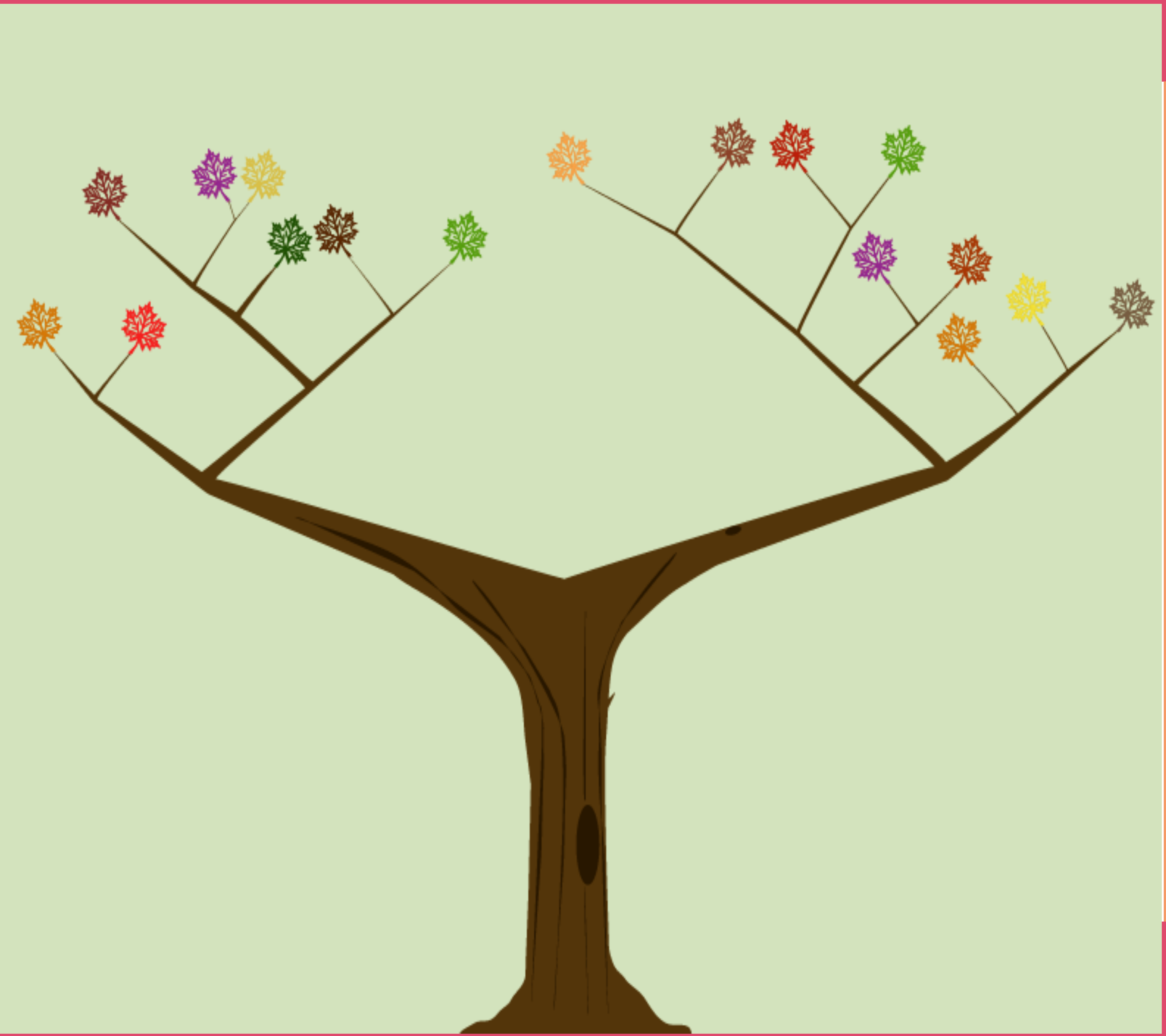
SURVEY ESTIMATION WITH ELASTIC NET REGRESSION



Dr. Kelly McConville

Collaborators:

- Swarthmore: Becky Tang, George Zhu
- FIA: Gretchen Moisen, Tracey Frescino
- Colorado State: Jay Breidt
- UC, Davis: Thomas Lee

COMBINING DATA SOURCES TO IMPROVE ESTIMATOR EFFICIENCY

MODEL-ASSISTED SURVEY ESTIMATION

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

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

MODEL-ASSISTED SURVEY ESTIMATION

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

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

MODEL-ASSISTED SURVEY ESTIMATION

Assume auxiliary 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

MODEL-ASSISTED SURVEY ESTIMATION

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

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

MODEL-ASSISTED SURVEY ESTIMATION

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

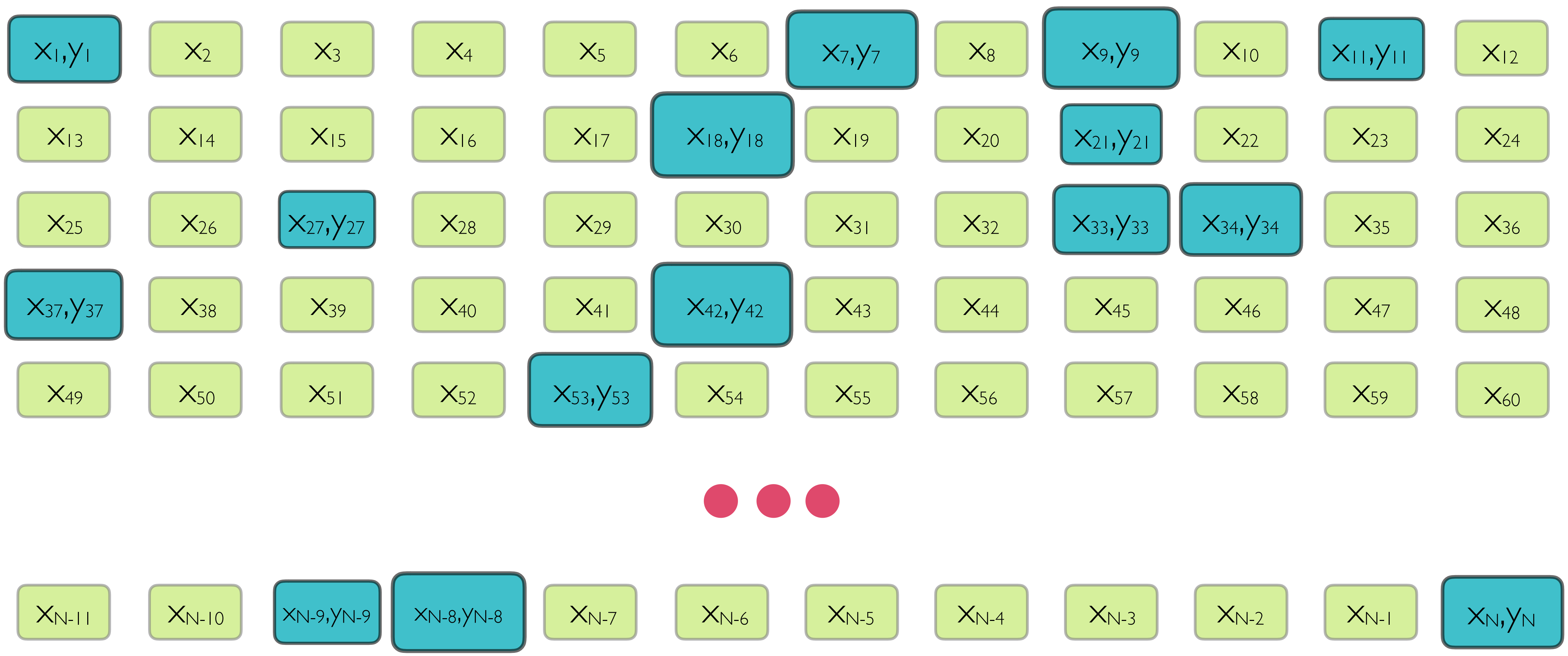
x_1, y_1	x_2	x_3	x_4	x_5	x_6	x_7, y_7	x_8	x_9, y_9	x_{10}	x_{11}, y_{11}	x_{12}
x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}, y_{18}	x_{19}	x_{20}	x_{21}, y_{21}	x_{22}	x_{23}	x_{24}
x_{25}	x_{26}	x_{27}, y_{27}	x_{28}	x_{29}	x_{30}	x_{31}	x_{32}	x_{33}, y_{33}	x_{34}, y_{34}	x_{35}	x_{36}
x_{37}, y_{37}	x_{38}	x_{39}	x_{40}	x_{41}	x_{42}, y_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}
x_{49}	x_{50}	x_{51}	x_{52}	x_{53}, y_{53}	x_{54}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}	x_{60}

...

x_{N-11}	x_{N-10}	x_{N-9}, y_{N-9}	x_{N-8}, y_{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, y_N

MODEL-ASSISTED SURVEY ESTIMATION

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

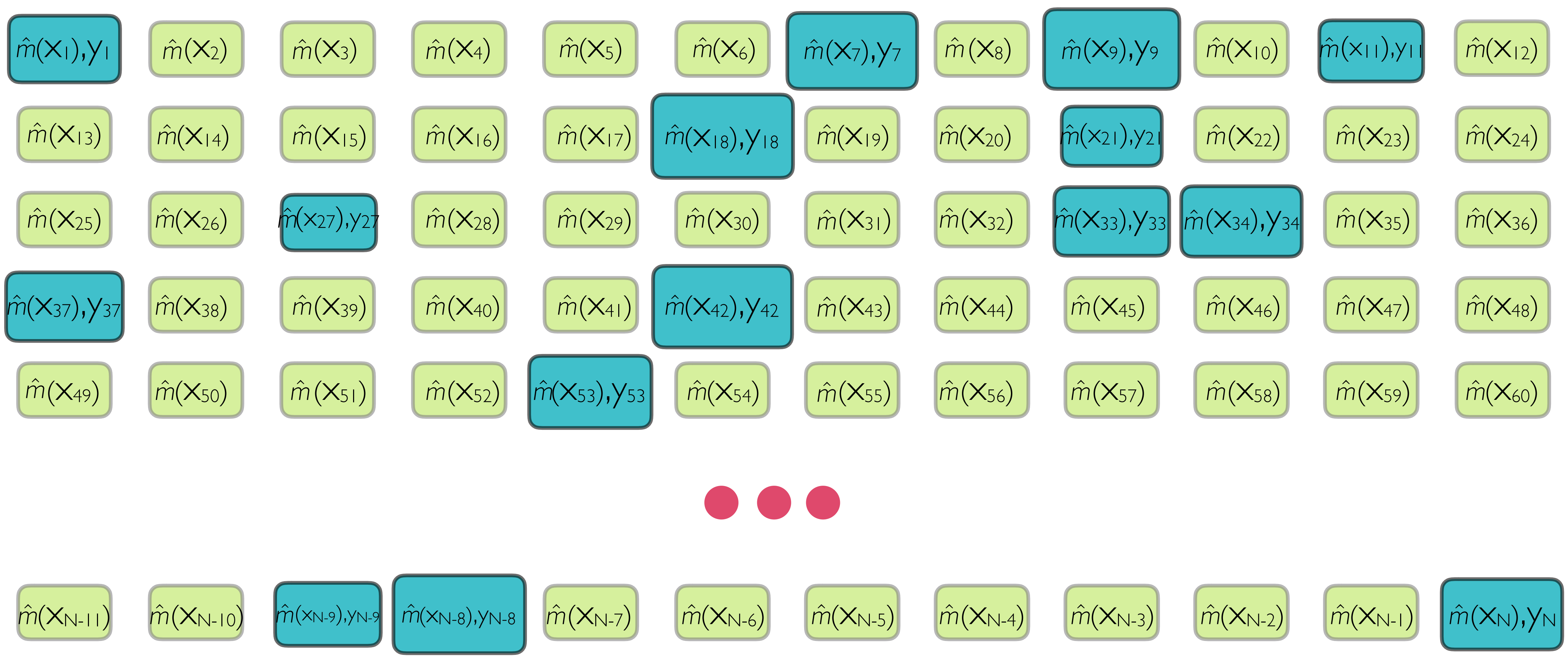


$$\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Horvitz and Thompson (1952)

MODEL-ASSISTED SURVEY ESTIMATION

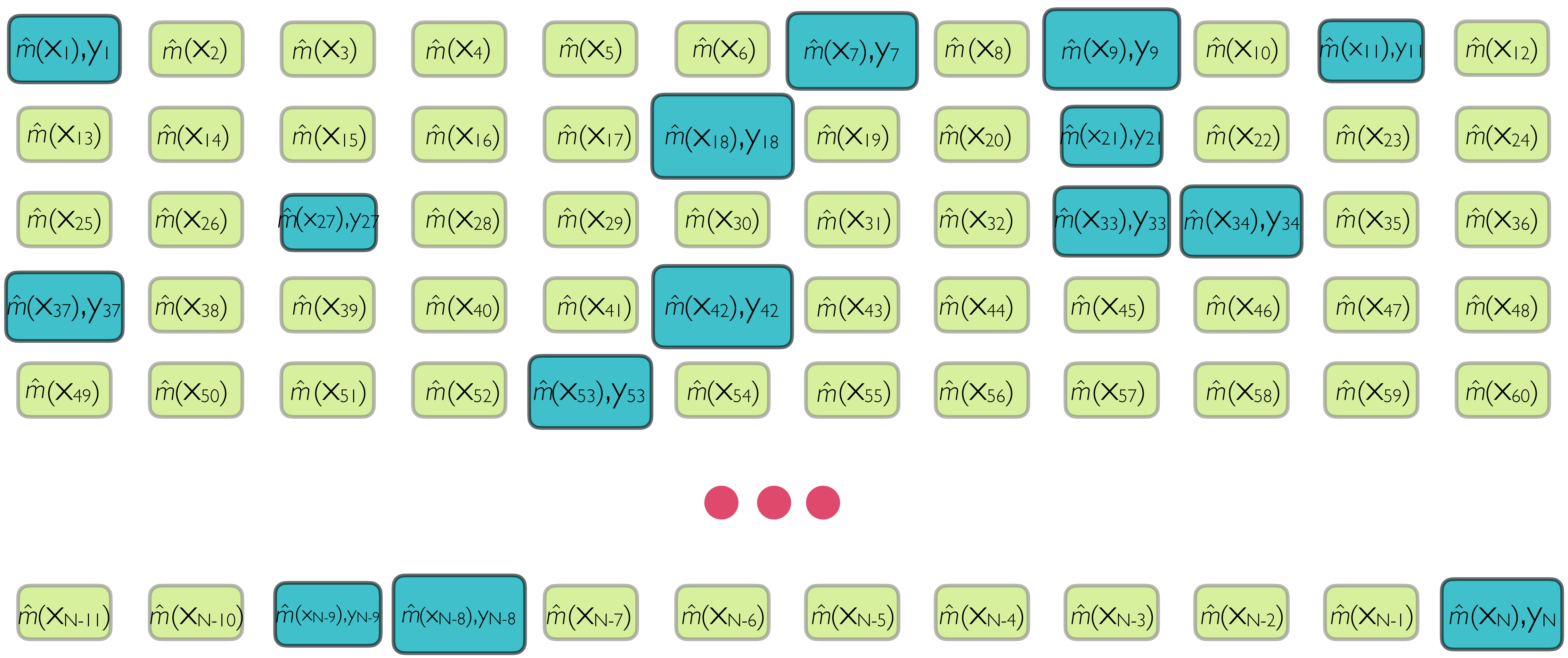
Can use the auxiliary data to predict the study variable.



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

MODEL-ASSISTED SURVEY ESTIMATION

Construct the estimator.



$$\hat{t}_y = \sum_{i \in U} \hat{m}(x_i)$$

MODEL-ASSISTED SURVEY ESTIMATION

Adjust the estimator for model mis-specification.

$\hat{m}(x_1), y_1$	$\hat{m}(x_2)$	$\hat{m}(x_3)$	$\hat{m}(x_4)$	$\hat{m}(x_5)$	$\hat{m}(x_6)$	$\hat{m}(x_7), y_7$	$\hat{m}(x_8)$	$\hat{m}(x_9), y_9$	$\hat{m}(x_{10})$	$\hat{m}(x_{11}), y_{11}$	$\hat{m}(x_{12})$
$\hat{m}(x_{13})$	$\hat{m}(x_{14})$	$\hat{m}(x_{15})$	$\hat{m}(x_{16})$	$\hat{m}(x_{17})$	$\hat{m}(x_{18}), y_{18}$	$\hat{m}(x_{19})$	$\hat{m}(x_{20})$	$\hat{m}(x_{21}), y_{21}$	$\hat{m}(x_{22})$	$\hat{m}(x_{23})$	$\hat{m}(x_{24})$
$\hat{m}(x_{25})$	$\hat{m}(x_{26})$	$\hat{m}(x_{27}), y_{27}$	$\hat{m}(x_{28})$	$\hat{m}(x_{29})$	$\hat{m}(x_{30})$	$\hat{m}(x_{31})$	$\hat{m}(x_{32})$	$\hat{m}(x_{33}), y_{33}$	$\hat{m}(x_{34}), y_{34}$	$\hat{m}(x_{35})$	$\hat{m}(x_{36})$
$\hat{m}(x_{37}), y_{37}$	$\hat{m}(x_{38})$	$\hat{m}(x_{39})$	$\hat{m}(x_{40})$	$\hat{m}(x_{41})$	$\hat{m}(x_{42}), y_{42}$	$\hat{m}(x_{43})$	$\hat{m}(x_{44})$	$\hat{m}(x_{45})$	$\hat{m}(x_{46})$	$\hat{m}(x_{47})$	$\hat{m}(x_{48})$
$\hat{m}(x_{49})$	$\hat{m}(x_{50})$	$\hat{m}(x_{51})$	$\hat{m}(x_{52})$	$\hat{m}(x_{53}), y_{53}$	$\hat{m}(x_{54})$	$\hat{m}(x_{55})$	$\hat{m}(x_{56})$	$\hat{m}(x_{57})$	$\hat{m}(x_{58})$	$\hat{m}(x_{59})$	$\hat{m}(x_{60})$
$\hat{m}(x_{N-11})$	$\hat{m}(x_{N-10})$	$\hat{m}(x_{N-9}), y_{N-9}$	$\hat{m}(x_{N-8}), y_{N-8}$	$\hat{m}(x_{N-7})$	$\hat{m}(x_{N-6})$	$\hat{m}(x_{N-5})$	$\hat{m}(x_{N-4})$	$\hat{m}(x_{N-3})$	$\hat{m}(x_{N-2})$	$\hat{m}(x_{N-1})$	$\hat{m}(x_N), y_N$

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

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: $\lim_{N \to \infty} \mathrm{E}\left[\frac{\hat{t}_y - t_y}{N}\right] = 0$
    - Small variance
  - But, the **size of the variance** depends on how well the assisting model captures the relationship between the study variable and the auxiliary data.

WHICH ASSISTING MODEL SHOULD ONE USE?

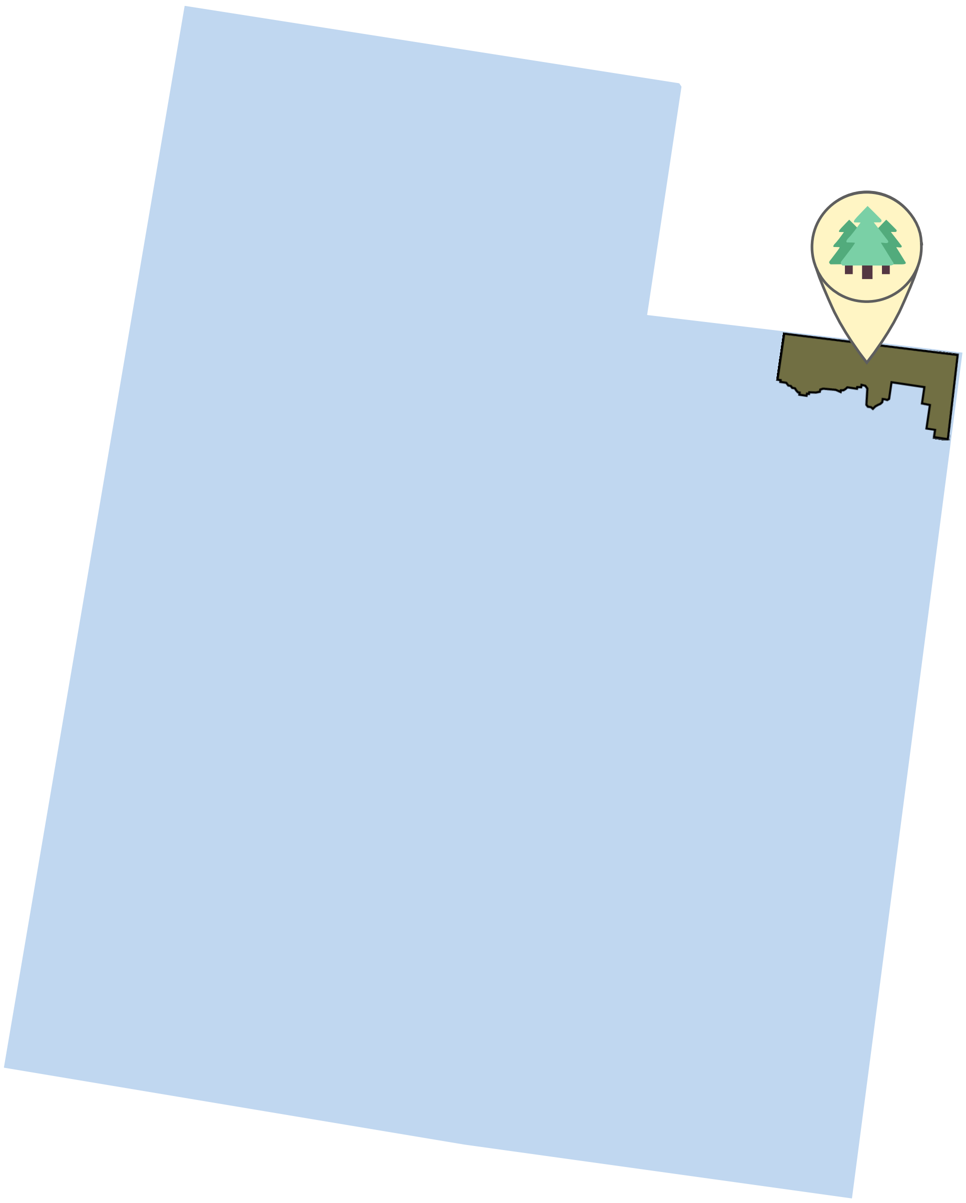
- Answer depends on…
  - What auxiliary data are available.
  - Appropriately modeling the relationship between the study variable and auxiliary data.
- Consider this question through the lens of a specific example:
  - Forest inventory

U.S. FOREST INVENTORY AND ANALYSIS (FIA)

- Tasked with monitoring status and trends in forested ecosystems across the U.S.
- It provides estimates of numerous forest attributes at a variety of subpopulations, such as county, state, and regional levels.
- Estimates are expected to be both unbiased and efficient, be computationally feasible for nationwide processing, and be easily explained to a broad user base.

DAGGETT COUNTY, UT

- County is the smallest estimation unit for FIA.
  - Many forest attributes are estimated.
    - EX: average trees per acre
- Discretize the region into equally sized units.



ESTIMATING FOREST ATTRIBUTES

Goal: Estimate the mean number of trees per acre.

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

$$\mu_y = N^{-1} t_y = \frac{1}{N} \sum_{i \in U} y_i$$

ESTIMATING FOREST ATTRIBUTES

FIA's sampling design: systematic sample

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

$$\pi_i = \pi$$

ESTIMATING FOREST ATTRIBUTES

The sample is drawn over a 10 year period.

x_1	x_2, y_2	x_3	x_4	x_5	x_6	x_7, y_7	x_8	x_9	x_{10}	x_{11}	x_{12}, y_{12}
x_{13}	x_{14}	x_{15}	x_{16}	x_{17}, y_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}, y_{22}	x_{23}	x_{24}
x_{25}	x_{26}	x_{27}, y_{27}	x_{28}	x_{29}	x_{30}	x_{31}	x_{32}, y_{32}	x_{33}	x_{34}	x_{35}	x_{36}
x_{37}, y_{37}	x_{38}	x_{39}	x_{40}	x_{41}	x_{42}, y_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}, y_{47}	x_{48}
x_{49}	x_{50}	x_{51}	x_{52}, y_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}, y_{57}	x_{58}	x_{59}	x_{60}

• • •

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

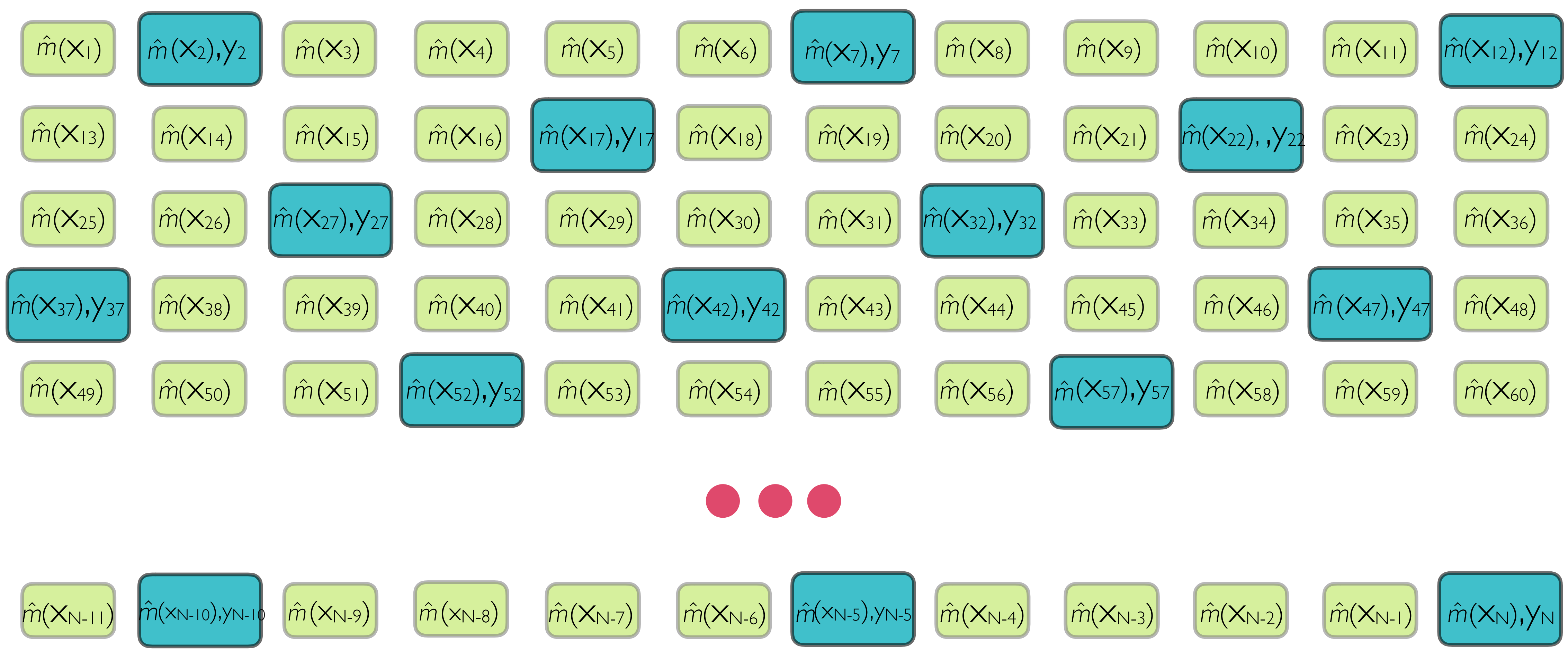
Number of field plots (n) = 80

Number of pixels (N) = 2,073,897

ESTIMATING FOREST ATTRIBUTES

Need to determine a good assisting model to construct estimator.

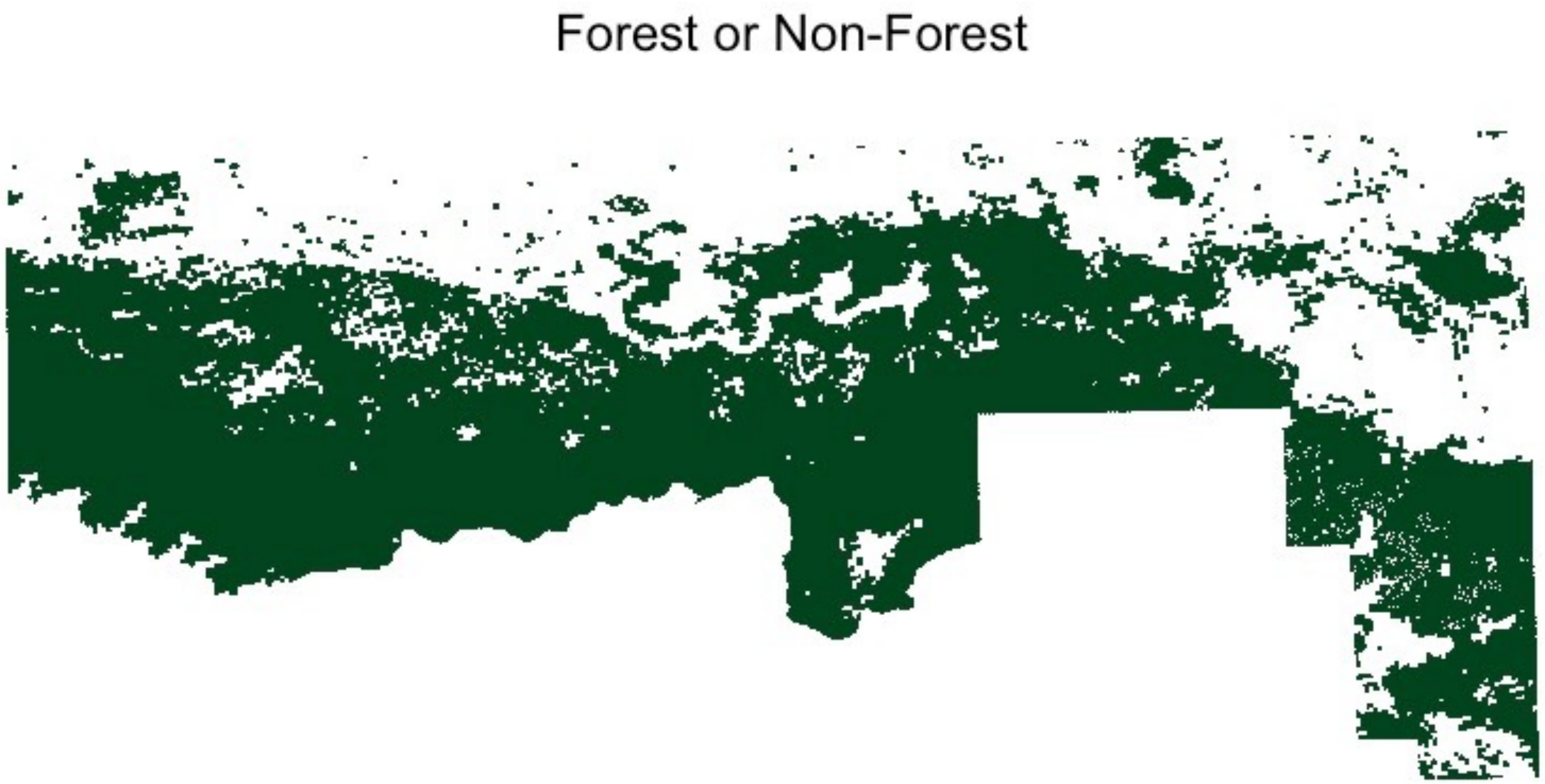
What auxiliary data are available?



$$N^{-1}\hat{t}_y = \frac{1}{N}\left(\sum_{i\in U}\hat{m}(x_i) + \sum_{i\in s}\frac{y_i - \hat{m}(x_i)}{\pi_i}\right)$$

ESTIMATING FOREST ATTRIBUTES

FIA currently uses only one auxiliary variable.

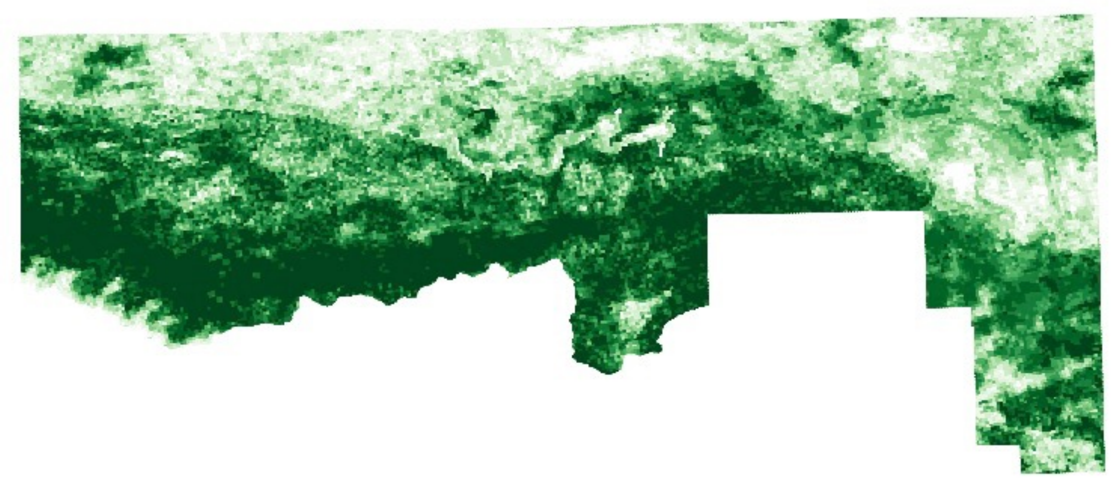


ESTIMATING FOREST ATTRIBUTES

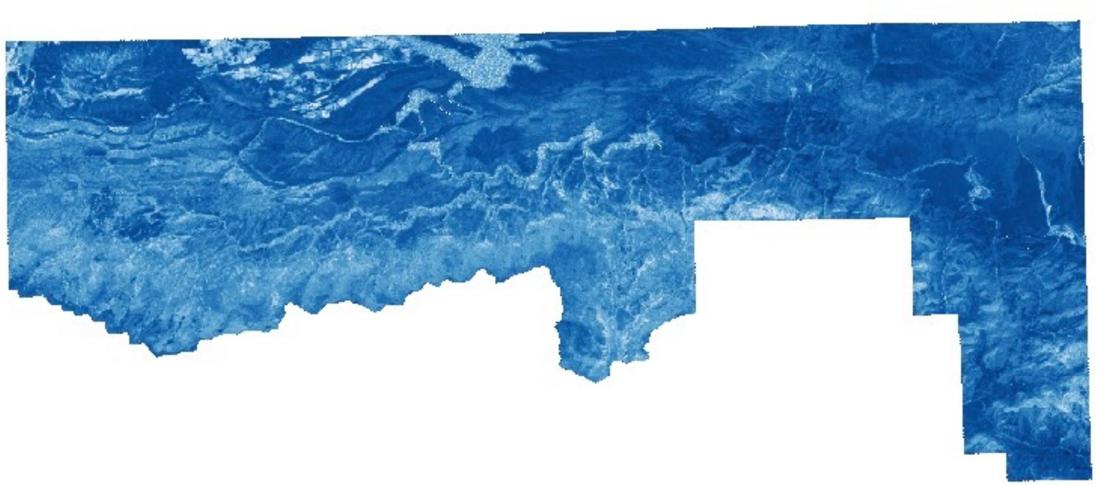
But, FIA has access to many auxiliary variables.

Forest or Non-Forest

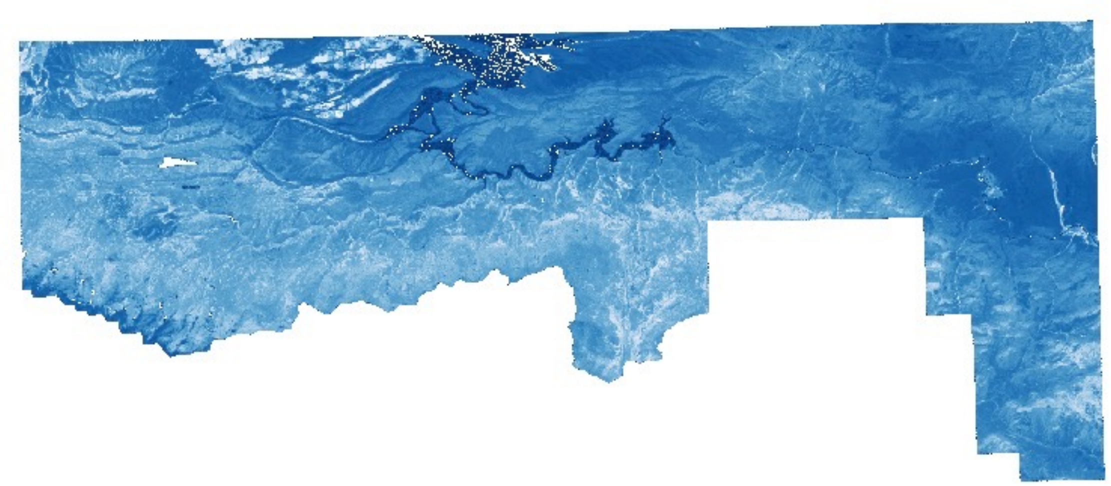
Probabilty of Forest



Normalized Burn Ratio



Normalized Difference Vegetation Index



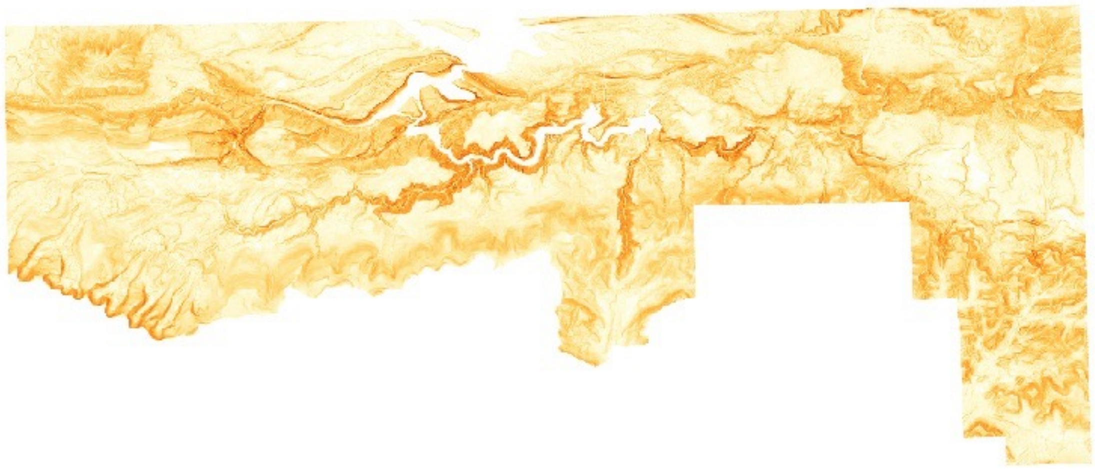
ESTIMATING FOREST ATTRIBUTES

But, FIA has access to many auxiliary variables.

Elevation

Eastness

Slope



ESTIMATING FOREST ATTRIBUTES

- FIA has access to many auxiliary 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!**

ESTIMATING FOREST ATTRIBUTES VIA THE ELASTIC NET

- Model:

$$y_i = m(\mathbf{x}_i) + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \epsilon_i$$

- Estimation criterion:

$$\hat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i \in s} \pi_i^{-1} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right] \right\}$$

  - Introduced in a non-survey context by Zou and Hastie (2005).
  - McConville et. al. (2017) extended lasso to survey case.

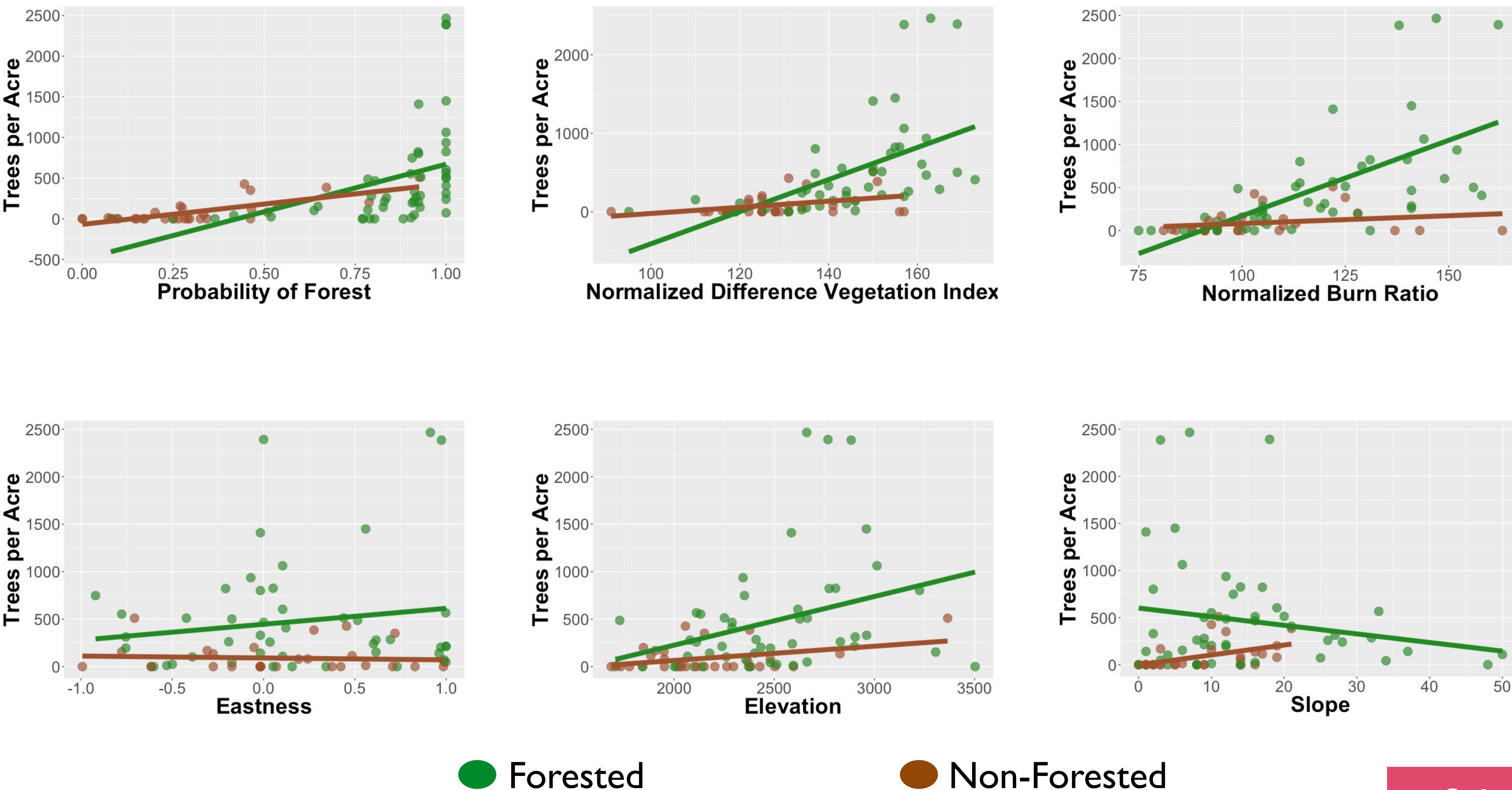
ESTIMATING FOREST ATTRIBUTES VIA THE ELASTIC NET

- Estimation criterion:

$$\hat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i \in s} \pi_i^{-1} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right] \right\}$$

  - Mixing parameter: α
  - Penalty parameter: λ
    - Non-negative value
    - As penalty parameter increases, estimates shrink toward zero.
    - Selected through cross validation.

ESTIMATING FOREST ATTRIBUTES



ESTIMATING FOREST ATTRIBUTES VIA THE ELASTIC NET

	Canopy Cover		Basal Area		Trees per Acre	
	Estimator	SE	Estimator	SE	Estimator	SE
HT	22.43	2.38	63.78	7.66	327.76	58.44
PS	22.89	2.01	64.58	7.19	336.35	55.82
REG	23.46	1.80	64.77	9.75	310.99	43.99
LASSO	23.42	1.62	64.54	8.20	316.55	43.90
ENET	23.63	1.62	66.62	7.81	339.17	45.73
RIDGE	23.51	1.58	65.30	8.18	318.96	42.93

MODEL-ASSISTED SURVEY ESTIMATION

- Allows us to combine multiple sources of information to estimation population quantities.
  - Utilizing a good model for the relationship between the study variable and auxiliary data can decrease the variance of the estimator.
  - Machine learning models allow for a flexible fit!

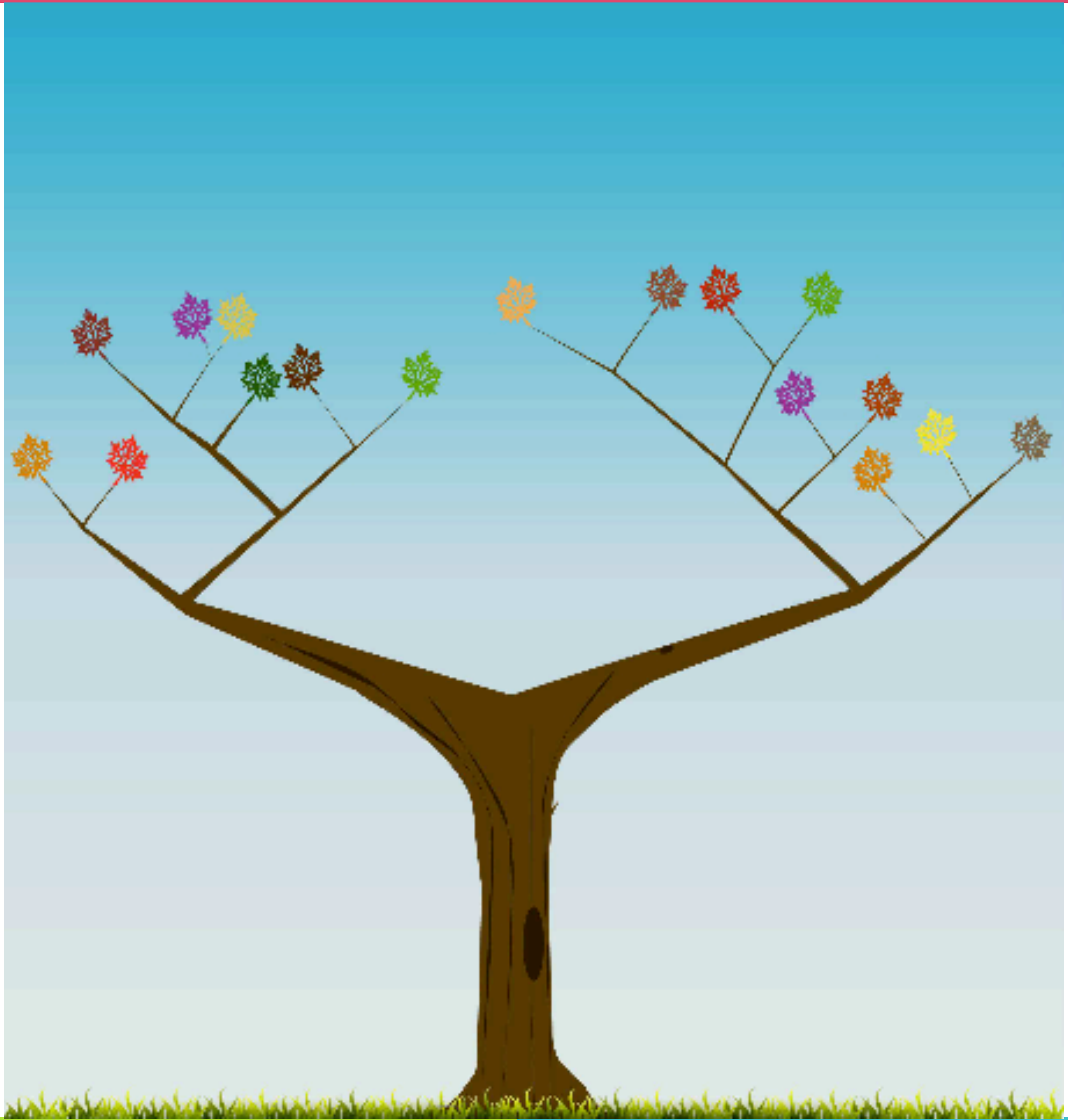
CURRENT RELATED WORK

- Estimating annual land use and land cover **change**.
    - Using photo interpreted data, ground data, and remote sensing data.
    - Trying a variety of model-assisted estimators.

REFERENCES

- 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., & Thompson, D. J. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association, 47(260), 663-685, 1952.
- 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.
- 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.
- Sarndal, C. E., B. Swensson, and J. Wretman. Model-Assisted Survey Sampling. Springer-Verlag, New York, 1992.
- Zou and Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

QUESTIONS?



QUESTIONS?