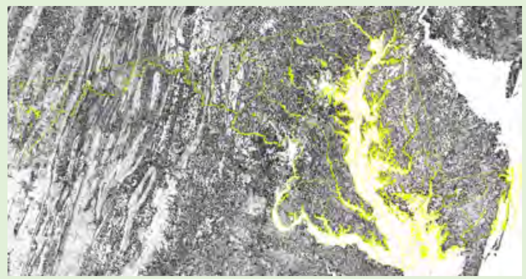
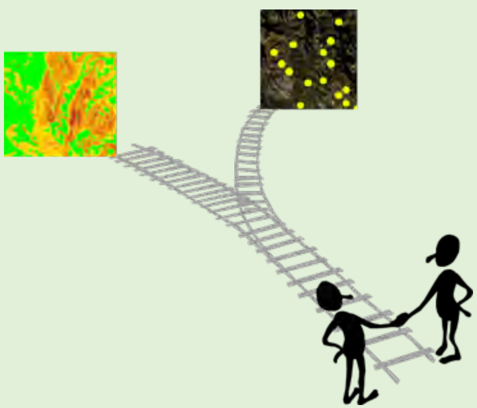
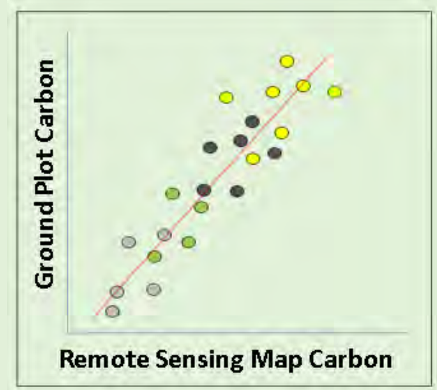
Small area estimation of forest carbon using carbon maps and model-assisted regression in Maryland

Forest Inventory and Analysis
We are the Nation's Forest Census
Welcome!







Andrew Lister (USDA Forest Service, Forest Inventory and Analysis (FIA)
George Hurtt and Lei Ma (University of Maryland)
Ty Wilson (USDA Forest Service, Forest Inventory and Analysis)

Objectives

- Investigate benefits of using model-assisted regression (MAR) over traditional post-stratification (PS) in FIA
- Assess MAR “hyperparameters” that affect operationalizing MAR in FIA

Outline

- Background on FIA
- Background on PS and MAR
- Results of study comparing PS and MAR
- Final thoughts, caveats, **questions for YOU**!

What does FIA do?

Land variables
- Forest type
- Site class
- Stand size
- Physiographic class
- Land use
- Land cover
- Owner type/class

…

Tree variables
- Species
- Diameter
- Height
- Merchantability
- Damage type
- Crown ratio
- Growth

…

Other
- Invasive plant cover
- Down woody debris
- Soils
- Regeneration

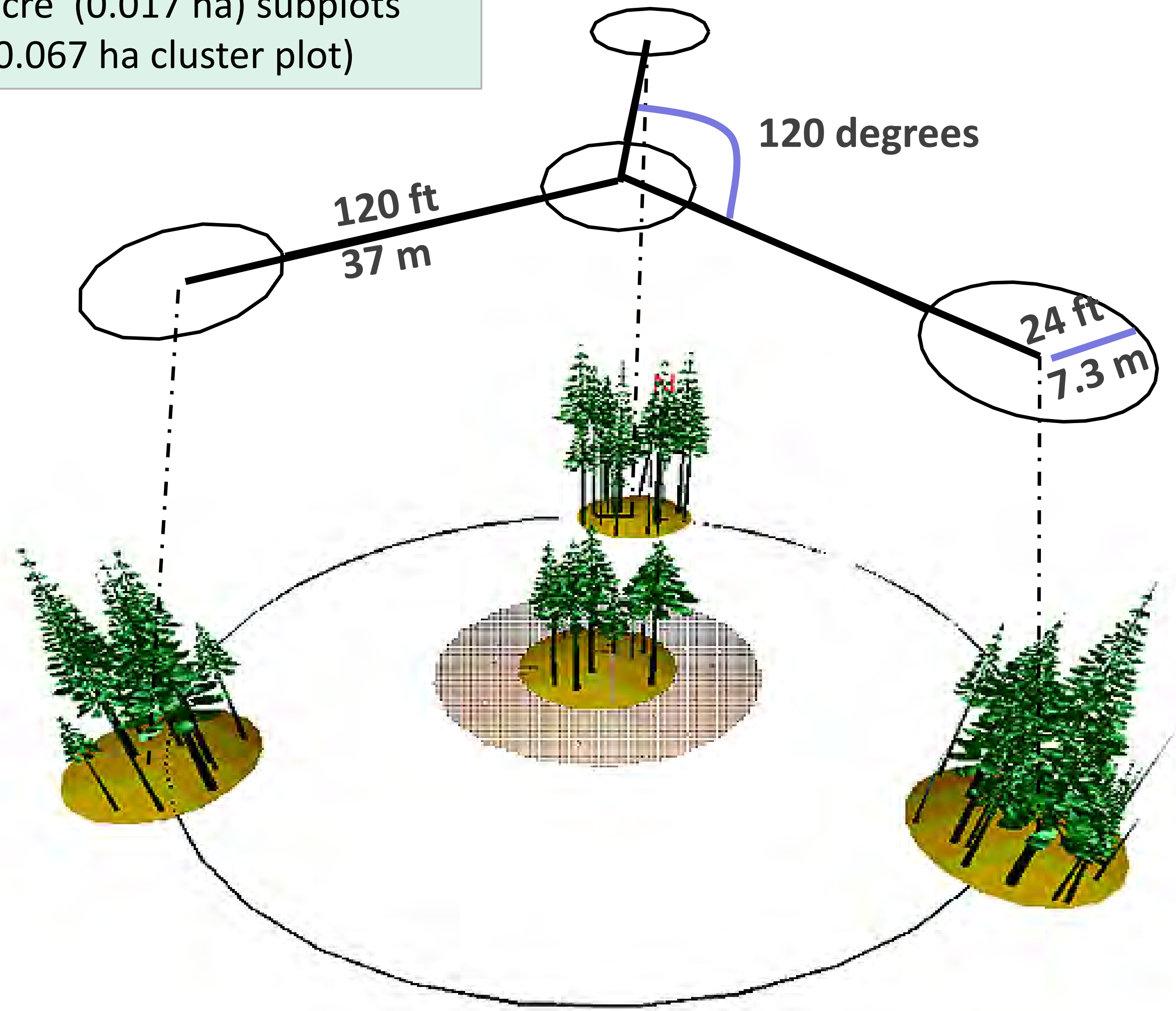
….

Data Processing
- Estimation
- Modelling
- GIS + Remote Sensing
- Reporting

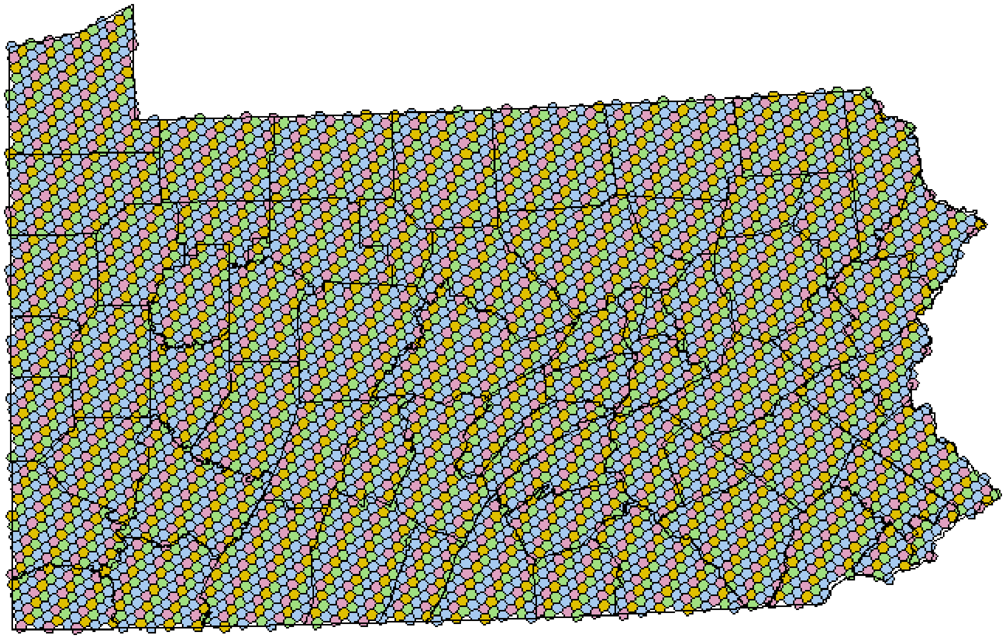
Reporting on the status of and trends in the Nation's forest resources; technology transfer to NFS and partners; basic and applied research; information for policy….

FIA Plot Design

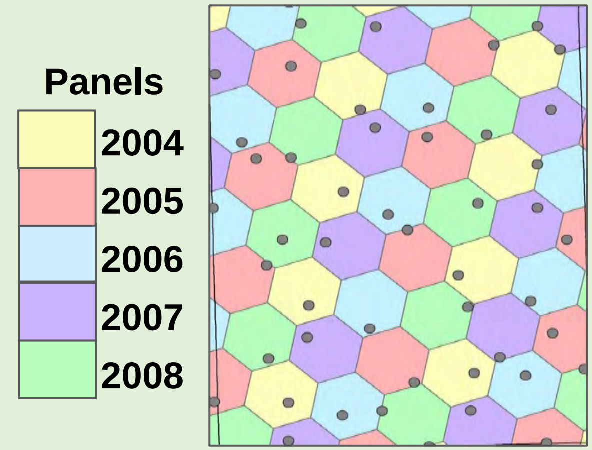
4 X 1/24 acre (0.017 ha) subplots
1/6 acre (0.067 ha cluster plot)



FIA Sample Design



Annual Inventory:



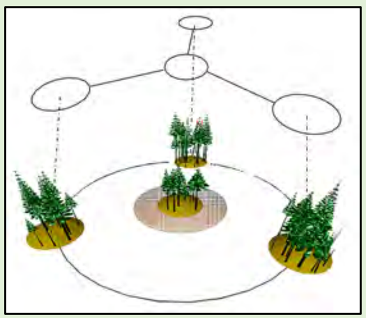
Any panel or cycle represents a full measurement of state.

Sample Intensity = 1 sample location per hexagon (5,933 ac, 2401 ha)

Inventory Cycle Length = Between 1/5 and 1/10 of the plots will be measured each year

>300,000 plots at full implementation!!

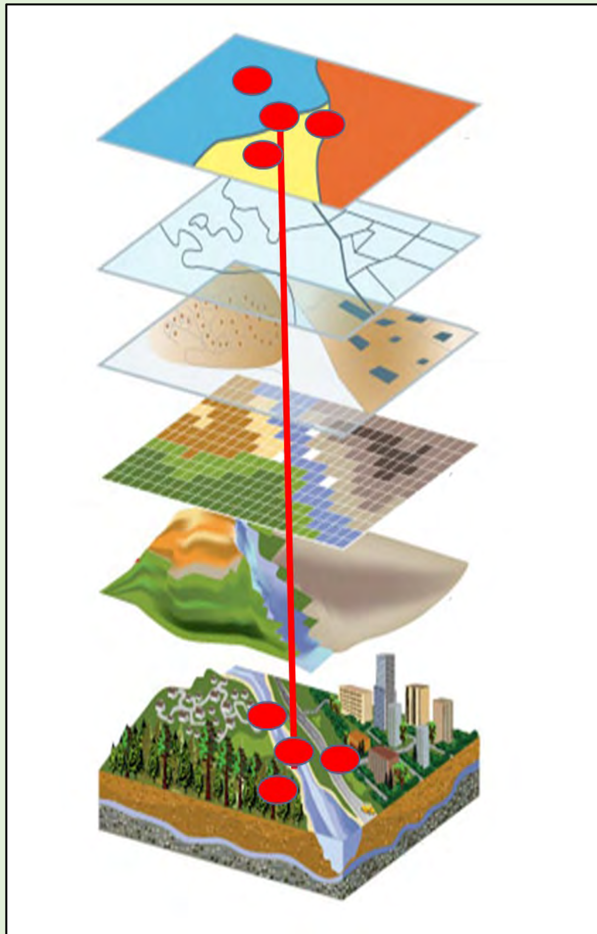
How does FIA use remote sensing and GIS?

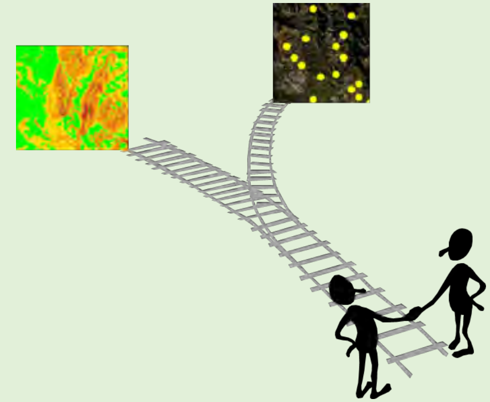


GPS

+

Plots labeled with **GIS and Remote Sensing data**





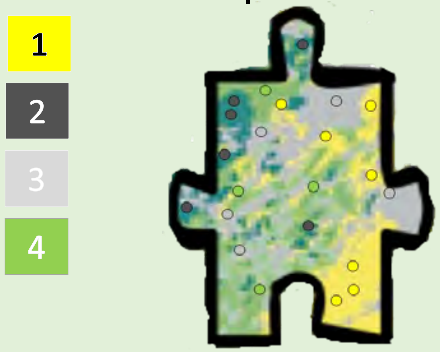
Improved estimation through stratification and other model-assisted estimation methods

Maps for small area estimation, additional context for tabular estimates

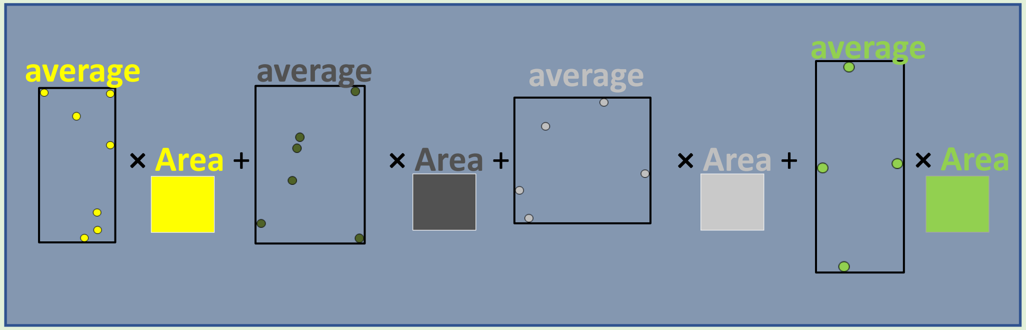
Traditional technique in FIA: Post-Stratification (PS)

Categorical

Stratum map

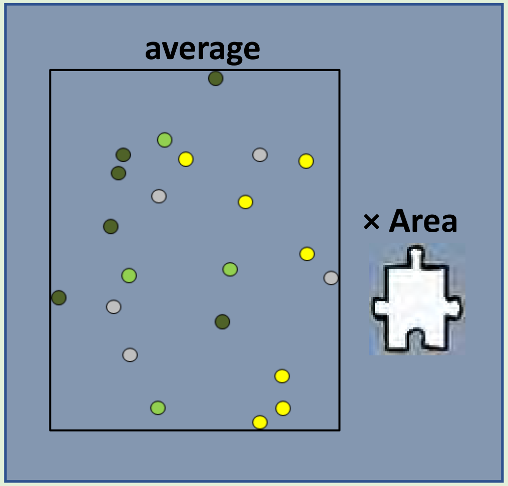


Weighted averaging procedure:



Vs.

Simple Random Sampling – No remote sensing:



Post-stratification gives smaller confidence intervals than Simple Random Sampling.

Smaller confidence interval = better, less expensive estimates in smaller areas.

PS estimators of mean and total

$$\bar{Y}_{hd} = \frac{\sum_{i=1}^{n_h} y_{hid}}{n_h}$$

The mean of the attribute (y) by stratum (h) for domain d

Number of plots in stratum *h*

$$\bar{Y}_d = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_{hd}$$

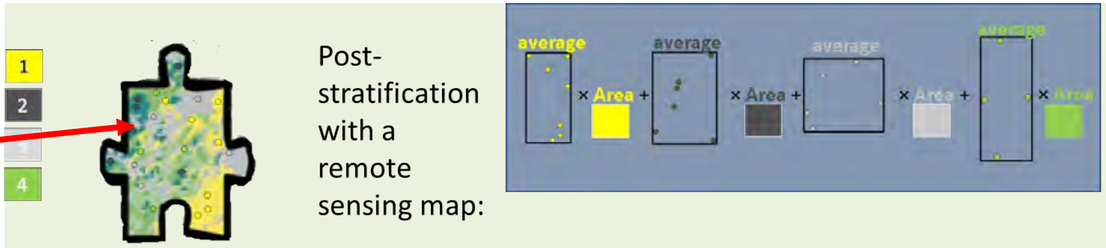
The weighted mean of the attribute in the domain

Stratum Weight (W_h)

$$\hat{Y}_d = A_T \bar{Y}_d$$

Estimate of total: weighted average X the Total Area (A_T)

N, N_h, A_T *and* y_{hi}



PS estimator of variance of total

$$v\left(\overline{Y}_{hd}\right)=\frac{\sum_{i}^{n_h} y_{hid}^{2}-n_h\overline{Y}_{hd}^{\;2}}{n_h\left(n_h-1\right)}$$

Variance of the mean of the attribute y in stratum h in domain d

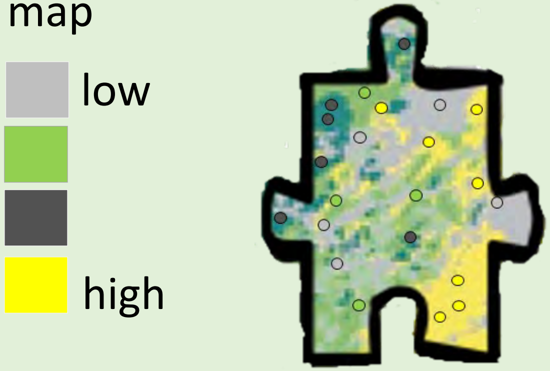
$$v\left(\hat{Y}_d\right)=\frac{A_T^2}{n}\left[\sum_{h}^{H} W_h n_h v\left(\overline{Y}_{hd}\right)+\sum_{h}^{H}\left(1-W_h\right)\frac{n_h}{n} v\left(\overline{Y}_{hd}\right)\right]$$

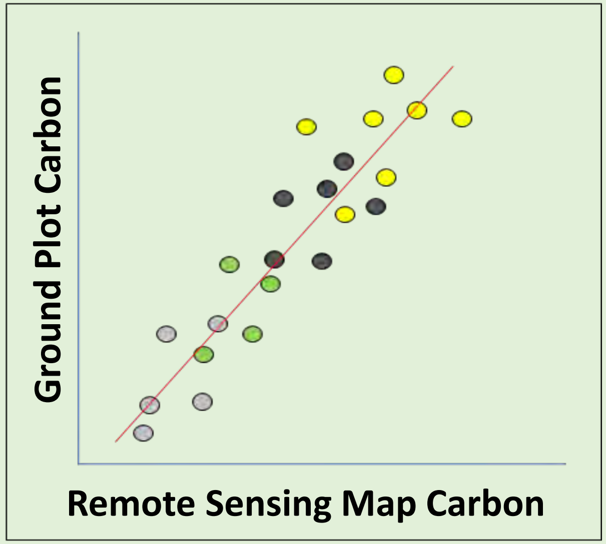
The variance of the mean, weighted with the stratum weight ($W_h = N_h/N$)

A penalty to address the fact that the number of plots in a stratum is not pre-determined, i.e., is a RV

Newer technique in FIA: Model Assisted Regression (MAR)

Continuous volume map





To calculate estimates and confidence intervals, need:

1. Linear* relationship between map and ground data
2. map value for each plot
3. sum of all pixel values in the map

*Can be used with other model types if df can be estimated.

Goal: a better estimate (smaller confidence interval) than PS or SRS

MAR estimator of mean

$$\bar{Y}_{MAR} = \bar{y} + b(\bar{X} - \bar{x})$$

$\bar{y} = the\ mean\ value\ of\ the\ attribute\ y$

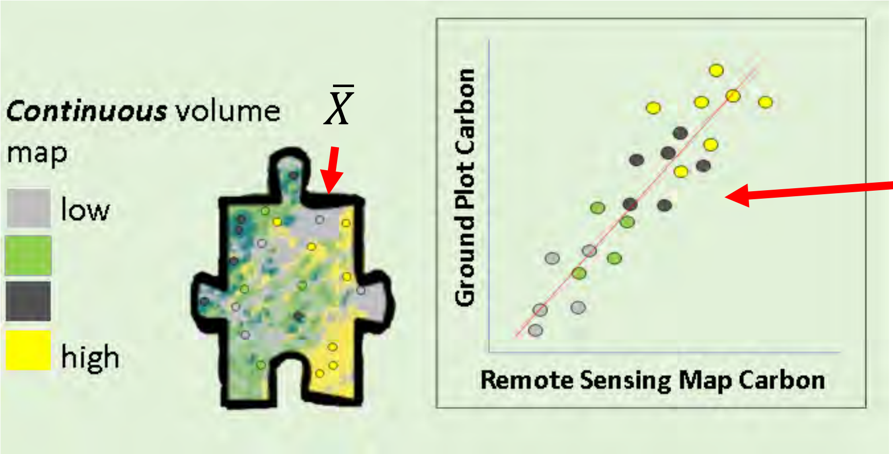
$b = the\ slope\ of\ the\ linear\ regression\ of\ y\ on\ x$

$\bar{X} = the\ mean\ value\ of\ the\ map\ value\ x\ in\ the\ (sub)population$

$\bar{x} = the\ mean\ value\ of\ the\ map\ value\ x\ at\ the\ locations\ of\ the\ y\ values$

$$\bar{Y}_{MAR} = a + b(\bar{X})$$

$a = the\ y\ intercept\ of\ the\ linear\ regression\ of\ y\ on\ x$



$\bar{y}, \bar{x}, a, b, and\ a + b(\bar{X})$

All you're doing is adjusting the mean of the pixels using the ground plots! (or you're adjusting the mean of the ground plots using the mean of the pixels….)

MAR estimator of variance of mean

$$v(\bar{Y}_{MAR}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-2)}$$

Variance of the mean of the attribute y from MAR

$$v(\bar{Y}_{MAR}) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n(n-2)}$$

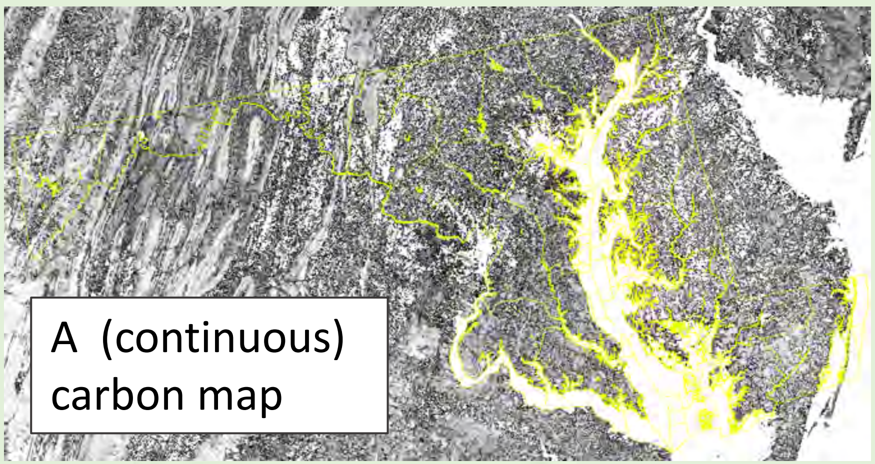
Equivalently, the variance of the residuals of the simple linear regression ÷ n (i.e., MSE/n)

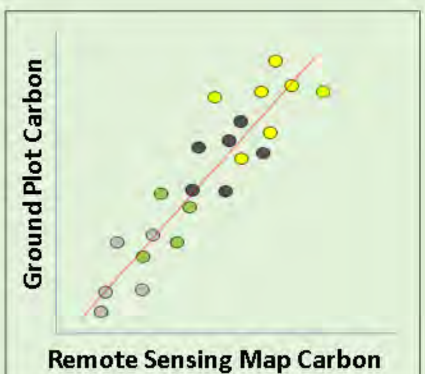
$$v(\bar{Y}_{MAR}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 * (n-1)}{n(n-2)}(1 - R^2)$$

$$v(\bar{Y}_{MAR}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cancel{*(n-1)}}{n\cancel{(n-2)}}(1 - R^2)$$

If we know the simple random sample variance and the R^2 of the regression of y on x, for large n, we know about how much using MAR will reduce the variance!

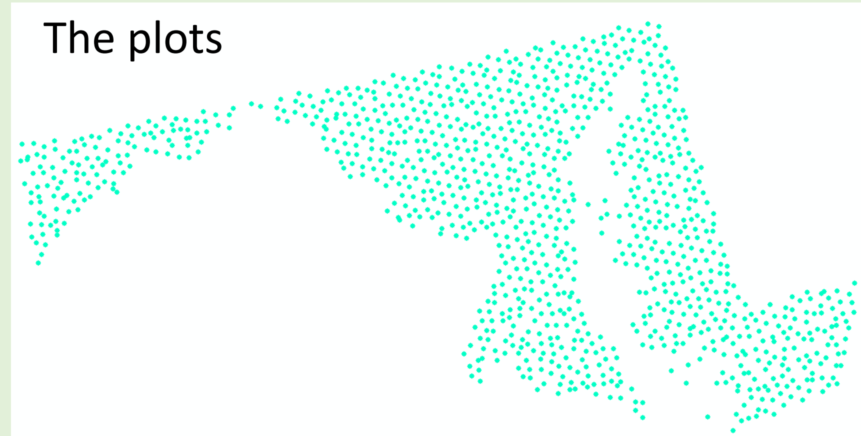
So what did *I* do? *I compared* $v(\bar{Y})$ from MAR with that from PS

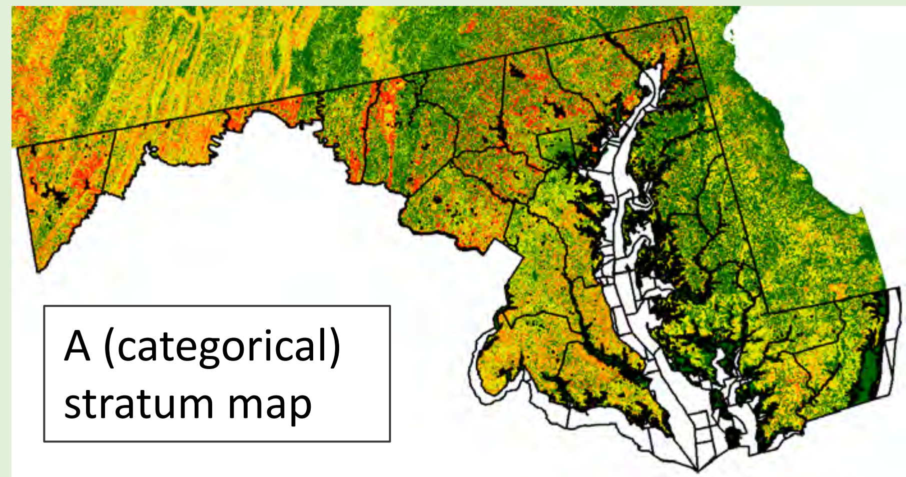




$$\bar{Y}_{MAR} = \bar{y} + b(\bar{X} - \bar{x})$$

$$v(\bar{Y}_{MAR}) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n(n-2)}$$





$$\bar{Y}_d = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_{hd}$$

$$v(\hat{Y}_d) = \frac{A_T^2}{n}\left[\sum_h^H W_h n_h v(\bar{Y}_{hd}) + \sum_h^H (1 - W_h)\frac{n_h}{n} v(\bar{Y}_{hd})\right]$$

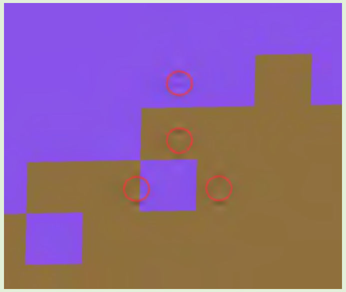
Some details:

The issue:

FIA uses subpopulations formed by combinations of county group, US Census Bureau inland water/land, and Public/Private ownership layer.

There are MANY forest carbon maps available. Some had predictions of C in nonforest areas.

There is a spatial mismatch between an FIA plot (covers 4 pixels) and a single pixel (the x_i in the equations).

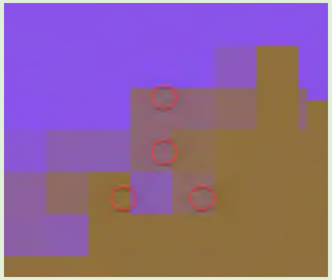


So I:

Did standard FIA subpopulation-based estimation, and did the MAR estimates both with and without subpopulations

Used 2 maps: a **University of Maryland** carbon map from a NASA CMS Carbon mapping project (Hurtt et al.) and the carbon map from **FIA's BigMap** project (Wilson et al.). **Applied a forest/nonforest mask** to the UMD maps.

Ran the MAR in 2 ways: 1. center subplot vs. original map, and 2. center subplot vs. the original map with a **3x3 pixel low-pass filter** (a 3x3pixel mean kernel) was applied.



RESULTS:

Lower is Better!

Map	Type	slope	y int	R square	estimate of total carbon (short tons)	sampling error	relative efficiency	# additional plots needed by PS to achieve the MAR sampling error
PS	subPop	-	-	-	101928140	3.38%	1.000	0
UMD	**subPop**	**0.21**	**-1.63**	**0.46**	**100710240**	**3.24%**	**1.046**	**92**
UMD	**state**	**0.22**	**-1.59**	**0.51**	**103356418**	**3.09%**	**1.095**	**194**
UMD	subPop - FNF mask	0.21	3.22	0.48	99862034	3.23%	1.047	95
UMD	state - FNF mask	0.22	2.68	0.55	100392470	3.06%	1.107	221
UMD	**subPop - 3x3 filter**	**0.25**	**-4.26**	**0.51**	[illegible]	[illegible]	[illegible]	**194**
UMD	**state - 3x3 filter**	**0.25**	**-3.80**	**0.56**	[illegible]	[illegible]	[illegible]	**318**
Big Map	subPop	9.00	1.63	0.56	[illegible]	[illegible]	[illegible]	341
Big Map	state	8.61	1.45	0.59	[illegible]	[illegible]	[illegible]	395
Big Map	**subPop - 3x3 filter**	**9.20**	**0.75**	**0.56**	[illegible]	[illegible]	[illegible]	**423**
Big Map	**state - 3x3 filter**	**9.17**	**0.74**	**0.61**	**103741091**	**2.77%**	**1.222**	**483**
					estimate of total carbon (short tons)	**sampling error**	**relative efficiency (SRS vs PS)**	**# additional plots needed by SRS to achieve the PS sampling error**
SRS - no map	state	-	-	-	95026485	4.82%	0.702	1010

$$Sampling\ Error = \frac{\sqrt{\frac{v(\bar{\bar{Y}})}{n}}}{\bar{\bar{Y}}} * 100$$

RESULTS:

Higher is Better!

Map	Type	slope	y int	R square	estimate of total carbon (short tons)	sampling error	relative efficiency	# additional plots needed by PS to achieve the MAR sampling error
PS	subPop	-	-	-	101928140	3.38%	1.000	0
UMD	**subPop**	**0.21**	**-1.63**	**0.46**	**100710240**	**3.24%**	**1.046**	**92**
UMD	**state**	**0.22**	**-1.59**	**0.51**	**103356418**	**3.09%**	**1.095**	**194**
UMD	subPop - FNF mask	0.21	3.22	0.48	99862034	3.23%	1.047	95
UMD	state - FNF mask	0.22	2.68	0.55	100392470	3.06%	1.107	221
UMD	**subPop - 3x3 filter**	**0.25**	**-4.26**	**0.51**	**100073120**	**3.09%**	[illegible]	**194**
UMD	**state - 3x3 filter**	**0.25**	**-3.80**	**0.56**	**103255099**	**2.94%**	[illegible]	**318**
Big Map	subPop	9.00	1.63	0.56	103591553	2.91%	[illegible]	341
Big Map	state	8.61	1.45	0.59	102449083	2.86%	1.184	395
Big Map	**subPop - 3x3 filter**	**9.20**	**0.75**	**0.56**	**104769239**	**2.83%**	**1.196**	**423**
Big Map	**state - 3x3 filter**	**9.17**	**0.74**	**0.61**	**103741091**	**2.77%**	**1.222**	**483**
					estimate of total carbon (short tons)	**sampling error**	**relative efficiency (SRS vs PS)**	**# additional plots needed by SRS to achieve the PS sampling error**
SRS - no map	state	-	-	-	95026485	4.82%	0.702	1010

$SampErr_{PS}$ / $SampErr_{MAR}$

RESULTS:

Map	Type	slope	y int	R square	estimate of total carbon (short tons)	sampling error	relative efficiency	# additional plots needed by PS to achieve the MAR sampling error
PS	subPop	-	-	-	101928140	3.38%	1.000	0
UMD	**subPop**	**0.21**	**-1.63**	**0.46**	**100710240**	**3.24%**	**1.046**	**92**
UMD	**state**	**0.22**	**-1.59**	**0.51**	**103356418**	**3.09%**	**1.095**	**194**
UMD	subPop - FNF mask	0.21	3.22	0.48	99862034	3.23%	1.047	95
UMD	state - FNF mask	0.22	2.68	0.55	100392470	3.06%	1.107	221
UMD	**subPop - 3x3 filter**	**0.25**	**-4.26**	**0.51**	**100073120**	**3.09%**	**1.094**	**194**
UMD	**state - 3x3 filter**	**0.25**	**-3.80**	**0.56**	**103255099**	**2.94%**	**1.151**	**318**
Big Map	subPop	9.00	1.63	0.56	103591553	2.91%	1.161	341
Big Map	state	8.61	1.45	0.59	102449083	2.86%	1.184	395
Big Map	**subPop - 3x3 filter**	**9.20**	**0.75**	**0.56**	**104769239**	**2.83%**	**1.196**	**423**
Big Map	**state - 3x3 filter**	**9.17**	**0.74**	**0.61**	**103741091**	**2.77%**	**1.222**	**483**

					estimate of total carbon (short tons)	sampling error	relative efficiency (SRS vs PS)	# additional plots needed by SRS to achieve the PS sampling error
SRS - no map	state	-	-	-	95026485	4.82%	0.702	1010

RESULTS:

Map	Type	slope	y int	R square	estimate of total carbon (short tons)	sampling error	relative efficiency	# additional plots needed by PS to achieve the MAR sampling error
PS	subPop	-	-	-	101928140	3.38%	1.000	0
UMD	**subPop**	**0.21**	**-1.63**	**0.46**	**100710240**	**3.24%**	**1.046**	**92**
UMD	**state**	**0.22**	**-1.59**	**0.51**	**103356418**	**3.09%**	**1.095**	**194**
UMD	subPop - FNF mask	0.21	3.22	0.48	99862034	3.23%	1.047	95
UMD	state - FNF mask	0.22	2.68	0.55	100392470	3.06%	1.107	221
UMD	**subPop - 3x3 filter**	**0.25**	**-4.26**	**0.51**	**100073120**	**3.09%**	**1.094**	**194**
UMD	**state - 3x3 filter**	**0.25**	**-3.80**	**0.56**	**103255099**	**2.94%**	**1.151**	**318**
Big Map	subPop	9.00	1.63	0.56	103591553	2.91%	1.161	341
Big Map	state	8.61	1.45	0.59	102449083	2.86%	1.184	395
Big Map	**subPop - 3x3 filter**	**9.20**	**0.75**	**0.56**	**104769239**	**2.83%**	**1.196**	**423**
Big Map	**state - 3x3 filter**	**9.17**	**0.74**	**0.61**	**103741091**	**2.77%**	**1.222**	**483**
					estimate of total carbon (short tons)	sampling error	relative efficiency (SRS vs PS)	# additional plots needed by SRS to achieve the PS sampling error
SRS - no map	state	-	-	-	95026485	4.82%	0.702	1010

Take Home Points

MAR, compared head-to-head with standard FIA PS, leads to better precision.

There are many dials to turn when doing this, including pre-processing the default carbon maps (e.g., f/nf mask, 3x3 filter)

We have been talking about operationalizing this for years now.. How?

This will help entities (like those interested in carbon accounting) get better information for smaller areas for less money.

Definitional consistency between the attribute (carbon) and the map are important to ensure.

As we can see, the slope, intercept, and predictions are calculated with simple algebra. VERY compatible with our SQL-based estimation system!

```
=SUM(IF($K$2:$K$983=O2,($L$2:$L$983-P2)^2))
```

P	Q	R	S	T	U	V	W	X
xbar	ybar	n	sum(xi-xbar)^2	sum(yi-ybar)^2	sum((xi-xbar)*(yi-ybar))	slopeGlobal	interceptGlobal	sum(ycap-ybar)^2
73.44093686	14.75247746	982	5048482.074	486987.7994	1123756.693	0.2191	-1.5041	242378.3747

There's a few big buts related to domain estimation and nonresponse…

Questions? Answers to my questions? Come by the USDA Forest Service Forest Inventory and Analysis Table and we can discuss!

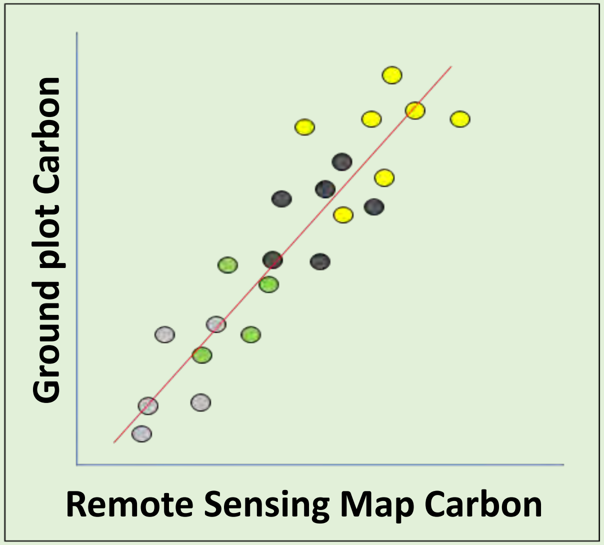
Forest Inventory and Analysis
We are the Nation's Forest Census
Welcome!

https://www.fia.fs.usda.gov/

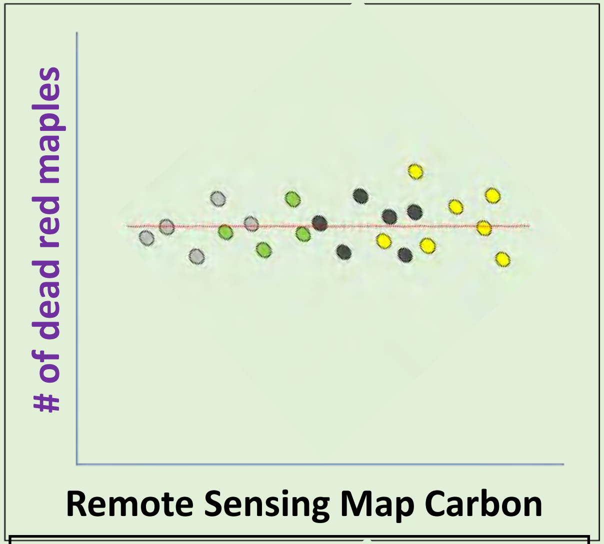
Andy Lister
andrew.lister@usda.gov

Big but 1:

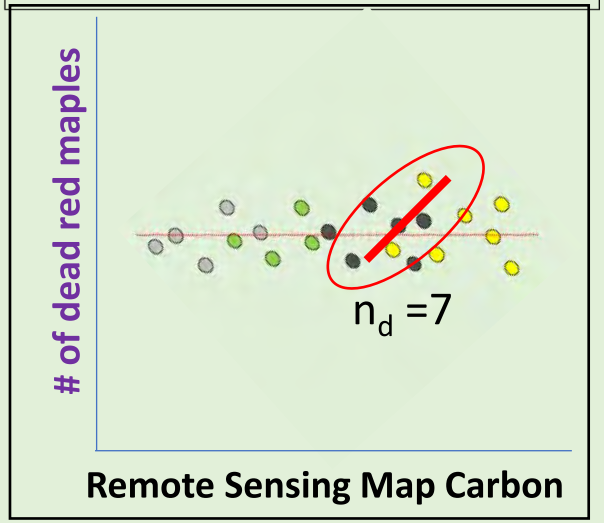
The MAR estimate is computed for each cell of a table.. What about analysis domain estimation?



We might need a new map for every cell in the table, and maps will likely be quite bad for strange attributes.



Uh oh…

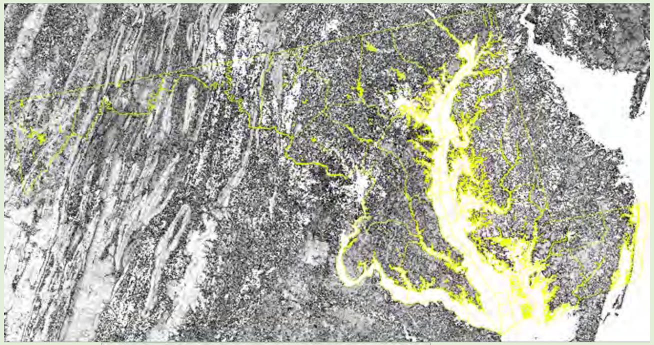


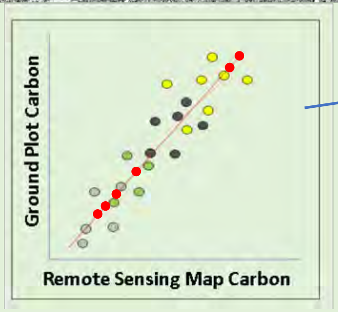
Question for YOU: maybe we make our MAR with just the n_d plots to calculate a mean?

Big but 2:

There is LOTS of nonresponse in our survey.

Question for YOU: Can we use MAR to improve upon current method (stratum mean applied to each missing value for the estimate, $n_{response}$ for the variance)?





$\hat{y}$ replaces y_{nr}

Maybe we differentially weight the imputed $\hat{y}$ values, or we use them to calculate the estimate but we use $n_{response}$ to calculate variance?