Applications of Small Area Estimation Methods in Forest Inventories and Modeling
Center for Intensive Planted-forest Silviculture

Bryce Frank, Temesgen Hailemariam & Francisco Mauro

February 2025

Outline

Part 1 (Introduction)

- Regression estimator
- Modelling at the unit level

Part 2 (Applications for Unit-level Models in Forestry)

- Small Area Estimation R packages and Data
- SAE Direct Estimation
- Class Exercise and Running Codes I

Part 3 (Applications for Area-level Models in Forestry)

- Small Area Estimation R packages and Data
- SAE Direct Estimation
- Class Exercise and Running Codes II

Part 4 (Variable Selection for SAE)

- Variable Selection for SAE
- Closing Thoughts and Future Direction

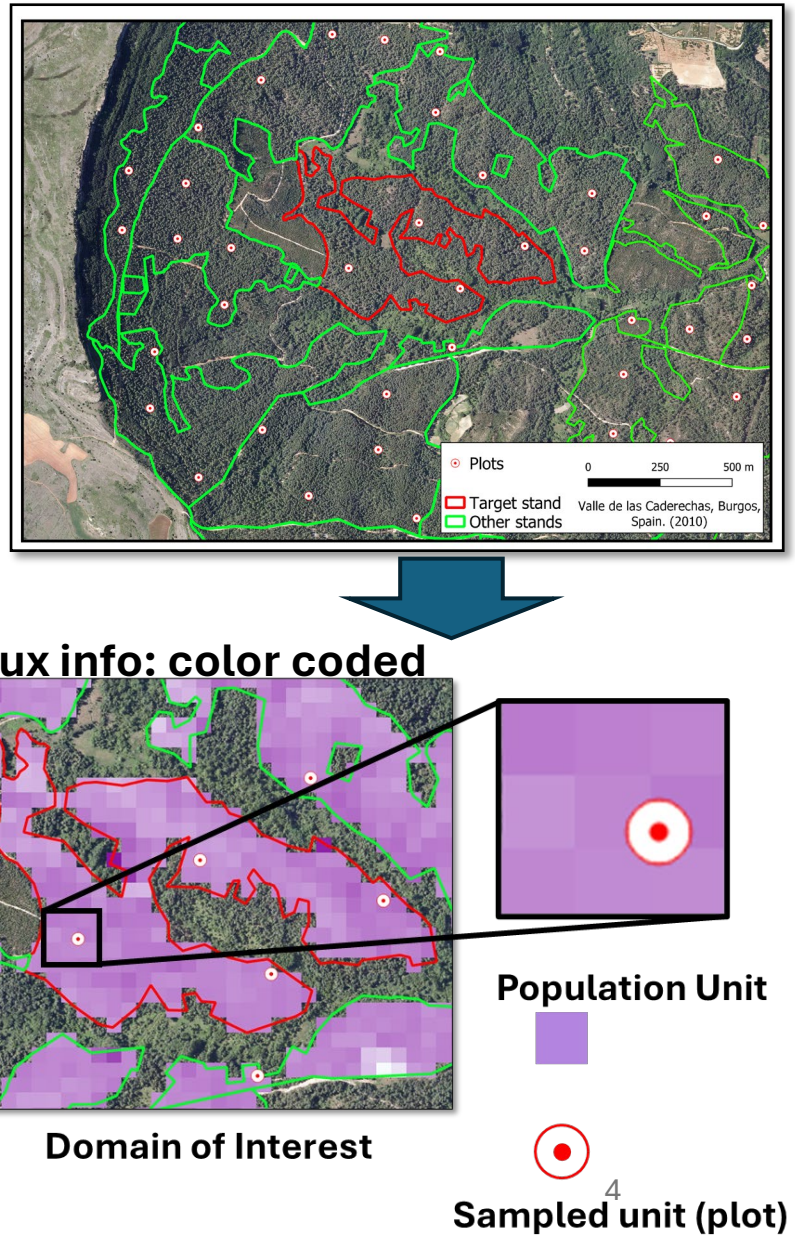
Part 1. Introduction

- The objective of this section is to **introduce the small area estimation problem**.
- To that end, we will **revisit classical sampling techniques (i.e., design-based)**.
- Then we will move forward to **understand how model-based estimation works**.
- The introduction concludes presenting the small area estimation problem in the **context of model-based estimation**.

Part 1. Introduction

To focus on the main ideas of the course we are going use some simplifications and common notation.

- **Study area**: A forest → Population (set of population units)→Enclosed in green polygons
- **Domains of interest**: Stands → Subpopulation (Subsets of units of the population, typically based on a delineation). →E.g., red polygon.
- **Population unit**: Pixels → grid cells with available auxiliary information (e.g. from lidar) with sizes ~ 1/10th acre
- **Sampled units**: Field plots → assumed equivalent to population units ~1/10th acre



Part 1. Estimators based on sampling design

When working with classical sampling methods (i.e., design based):

- **Each population unit has an UNKNOWN BUT FIXED value for the variable of interest. This implies that target parameters such as means or totals are UNKNOWN BUT FIXED quantities.**
- **Each population unit may have KNOWN BUT FIXED values for one or more auxiliary variables.**

This starting point is very reasonable and **do not imply any strong assumption about the population**.

We cannot measure all population units so we will estimate parameters of interest about our population or about subpopulations using sampling.

Part 1. Estimators based on sampling design

When working with classical sampling methods (i.e., design based). We have to do five steps.

1. **Define our target population parameter:** e.g., Mean volume in the forest, total volume in the forest, mean volume in one stand… **We cannot know it. It is a population parameter and we cannot measure the entire population We have to sample**.
2. **Establish randomized sampling design:** e.g. simple random sampling, stratified…
3. **Select estimator**: Sample mean, stratified mean…
4. **Collect the sample**: We go to the field and measure units according to our design
5. **Calculate our estimates and uncertainties:** Using the collected data, we calculate our estimators and uncertainty for the chosen sampling design.

We decide what the sampling design is and what estimator we are going to use!!!

Part 1. Estimators based on sampling design

The two most important ideas so far are:

We do not make any strong assumption about the population.

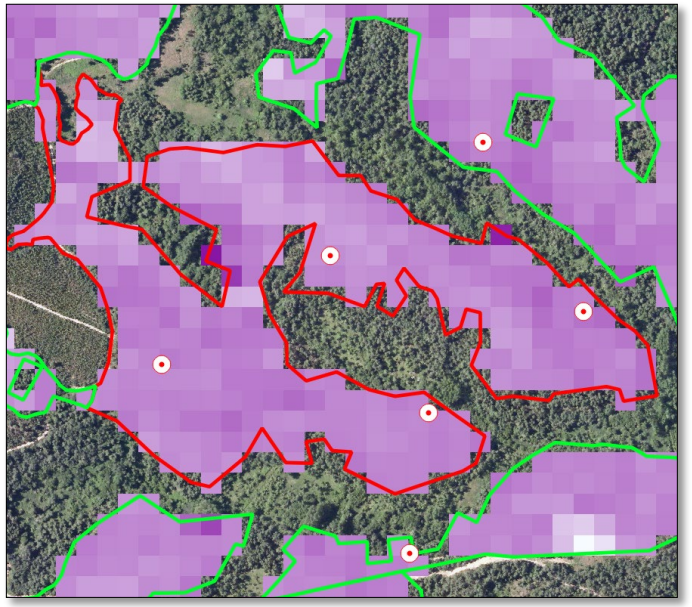
- **Each population unit has an UNKNOWN BUT FIXED value for the variable of interest. This implies that target parameters such as means or totals are UNKNOWN BUT FIXED quantities.**
- **Each population unit may have KNOWN BUT FIXED values for one or more auxiliary variable.**

<u>We decide what the sampling design is and what estimator we are going to use!!!</u>

Part 1. Estimators based on sampling design

Example:

1. **Target population parameter:** We want to estimate the mean volume ($\bar{Y}$) in the red polygon (mean of the volumes of all pixels)
2. **Establish sampling design:** We decide to do simple random sampling.
3. **Select estimator**: Our estimator will be the sample mean.
4. **Collect the sample**.
5. **Calculate our estimates and uncertainties:**



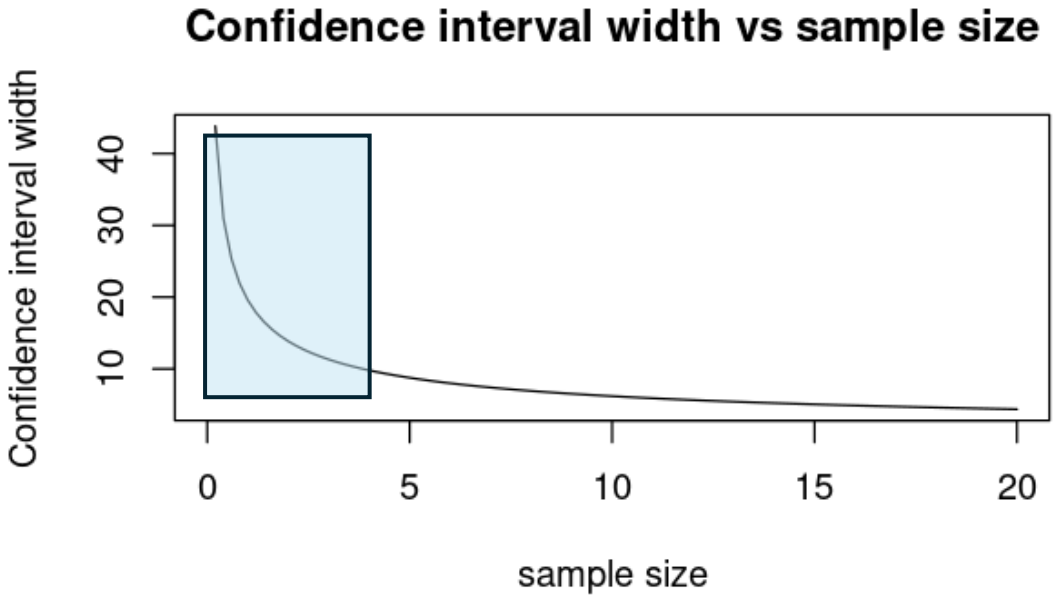
$$\hat{\bar{Y}} = \frac{1}{n} \sum_{i\ in\ sample} y_i \qquad \hat{V}(\hat{\bar{Y}}) = \frac{1}{n} \frac{\sum_{i\ in\ sample} (y_i - \hat{\bar{Y}})^2}{n-1} \qquad CI(\hat{Y}) = \hat{\bar{Y}} \pm 1.96 \sqrt{\hat{V}(\hat{Y})}$$

Part 1. Estimators based on sampling design

Uncertainty:

$$CI(\hat{Y}) = \hat{\bar{Y}} \pm 1.96\sqrt{\hat{V}(\hat{Y})} = \hat{\bar{Y}} \pm 1.96\sqrt{\frac{1}{n}\frac{\sum_{i\ in\ sample}(y_i - \hat{\bar{Y}})^2}{n-1}} =$$

$$\hat{\bar{Y}} \pm 1.96\ sd(sample)\sqrt{\frac{1}{n}}$$



In design-based methods, the uncertainty (confidence interval width) tends to be **proportional to the inverse of the square root of the sample size**.

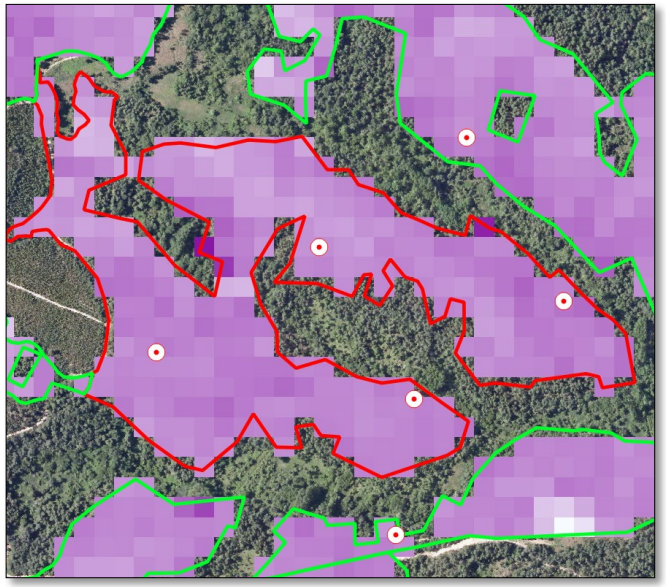
In many applications our sample sizes are in the blue area (large uncertainty)

Part 1. Estimators based on sampling design

What can we do if the error is large?

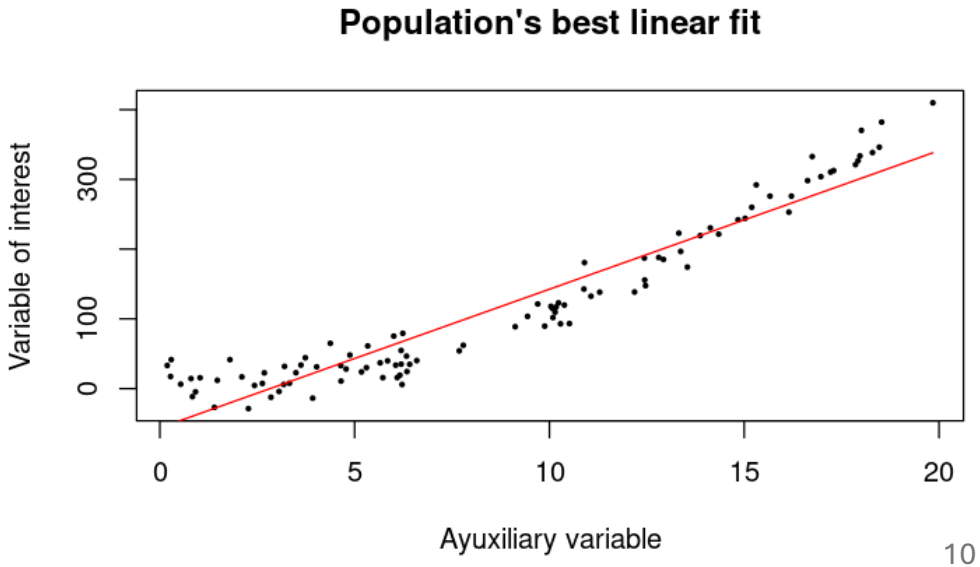
- **If there is auxiliary information available we can use regression estimators**.

All population units have a fixed and unknown value for the variable of interest (e.g., Volume) and values for some auxiliary variables (e.g., Lidar P95 purple)



For that population there is line of best fit.

- **We do not assume that the data is generated by that linear relationship, in this case the data seems to generate from something like a parabola.**
- **The line is just the line of best fit for that population , but in this case gets reasonably close.**

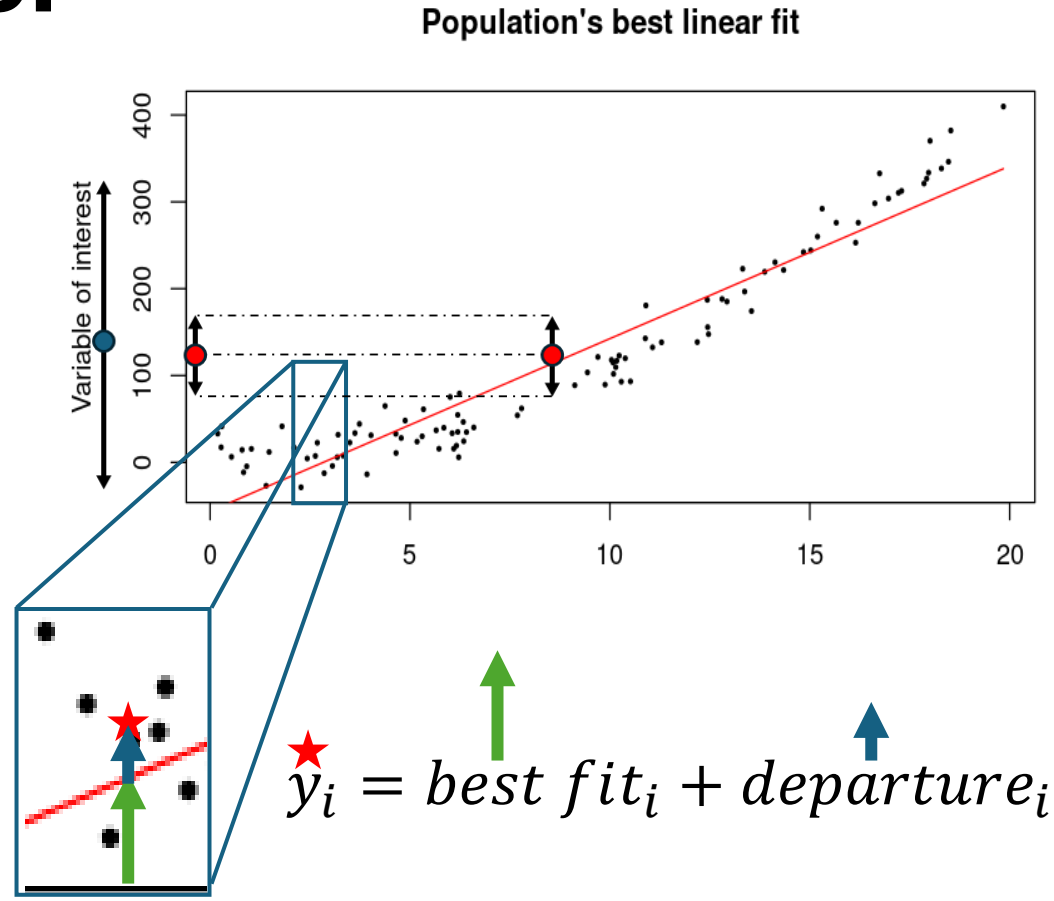


Part 1. Estimators based on sampling design

If we knew the line of best fit (**fixed**) we could calculate the adjusted value for any population unit and **only consider the departures** from that line (**also fixed but unknown**)

To estimate the population mean, **the only thing we would need to estimate is the mean of the departures**.

If the line of best fit represents the data well, **then the variability that is left (departures) would be smaller than the variability of the variable of interest**.



$$y_i = best\ fit_i + departure_i$$

$$\bar{Y} = \frac{1}{N}\sum_{all\ units} best\ fit_i + \frac{1}{N}\sum_{all\ units} departure_i$$

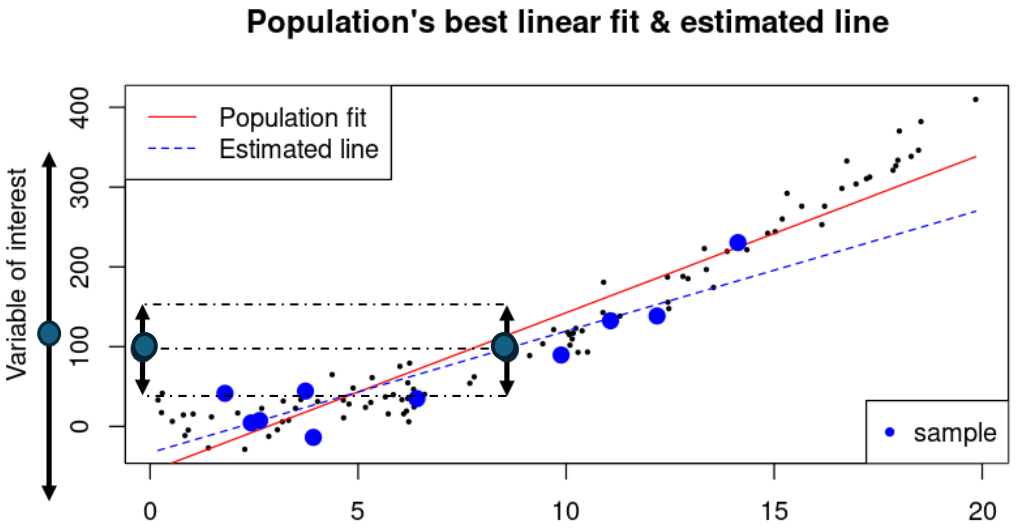
known

estimated

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{all\ unit} best\ fit_i + \frac{1}{n}\sum_{i\ in\ sample} departure_i$$

Part 1. Estimators based on sampling design

In practice, we do not know the line of best fit (red). What **we** do is**, estimate that line (blue) with the sample (for SRS and systematic use OLS)**.



We **estimate the population mean as the mean of fitted values for the entire population** (we have the auxiliary information for all units) **plus the mean of the departures with respect to the sample fit for the sample**.

$$y_i = sample\ fit_i + departure_i$$

Sample fit and the mean of the departures for the sample are estimates with uncertainty. If the blue line approximates well the data we increase precision.

$$\hat{\bar{Y}}_{regression} = \frac{1}{N} \sum_{all\ unit} sample\ fit_i + \frac{1}{n} \sum_{i\ in\ sample} departure_i$$

Part 1. Estimators based on sampling design

For large sample sizes regression estimators are unbiased

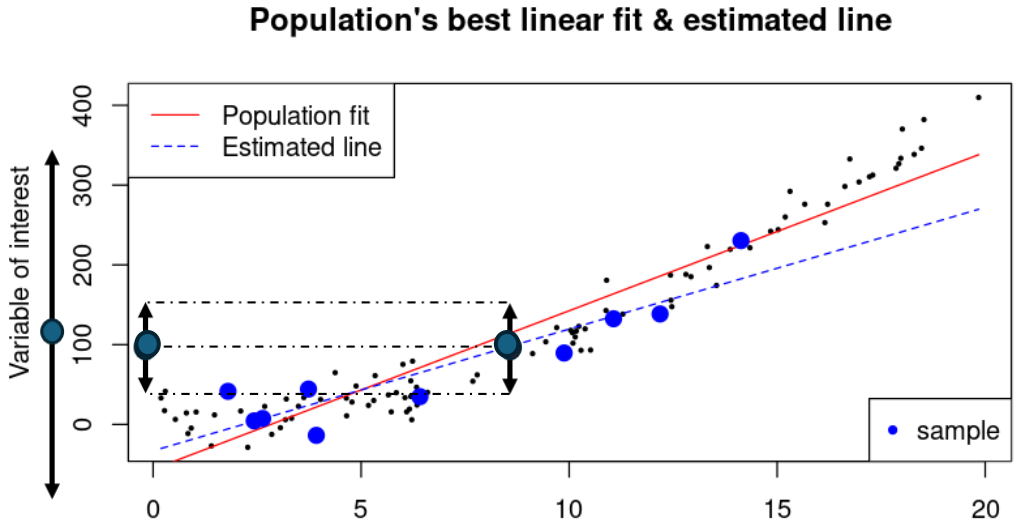
They are NOT based on the assumption of an existing linear relationship. In fact, the relation in the figures is a parabola.

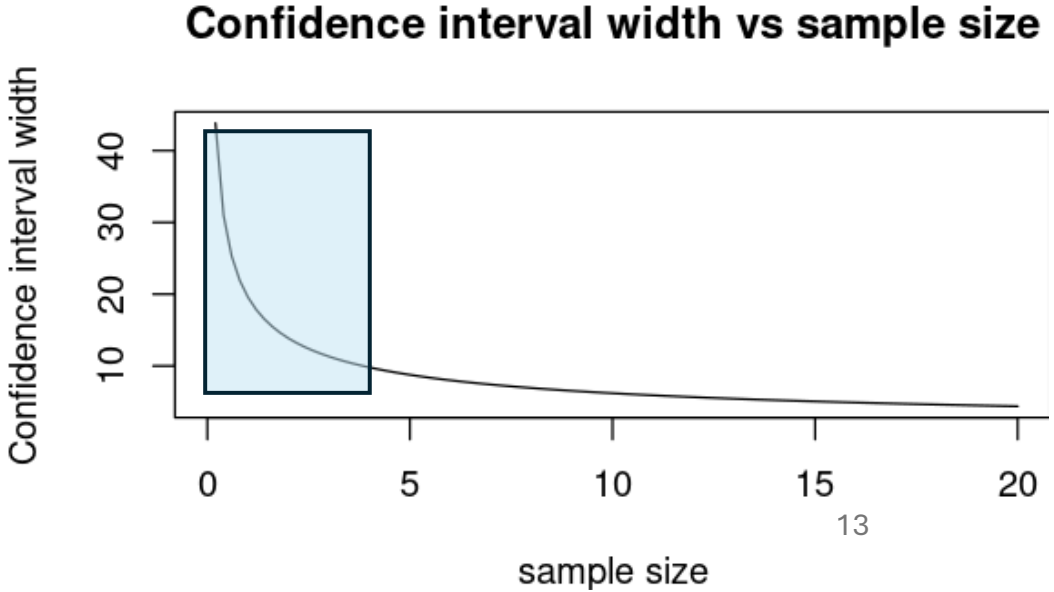
They just take as starting point the fit between auxiliary information and variable of interest for the population.

That fit is a fixed characteristic of the population.

$$y_i = sample\ fit_i + departure_i$$

$$\hat{\bar{Y}}_{regression} = \frac{1}{N} \sum_{all\ unit} sample\ fit_i + \frac{1}{n} \sum_{i\ in\ sample} departure_i$$

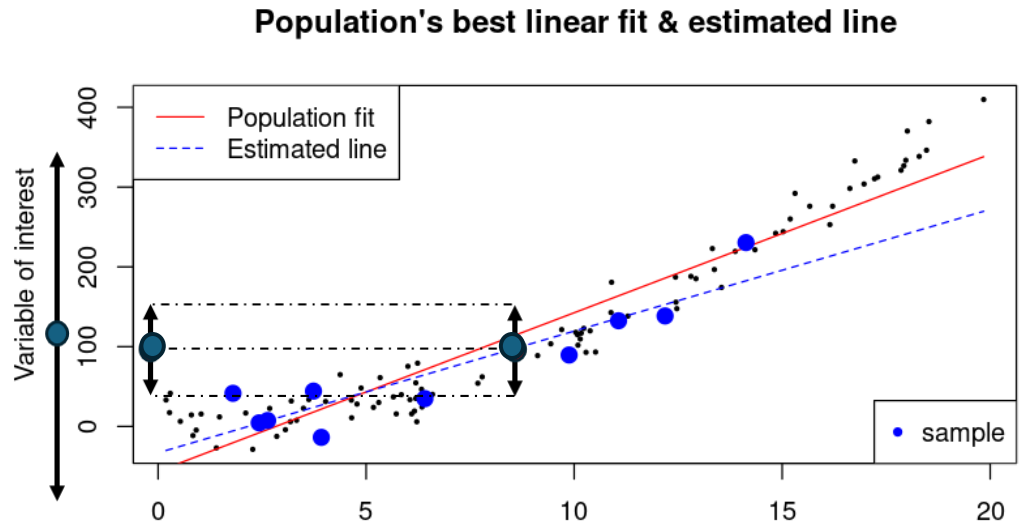




Part 1. Estimators based on sampling design

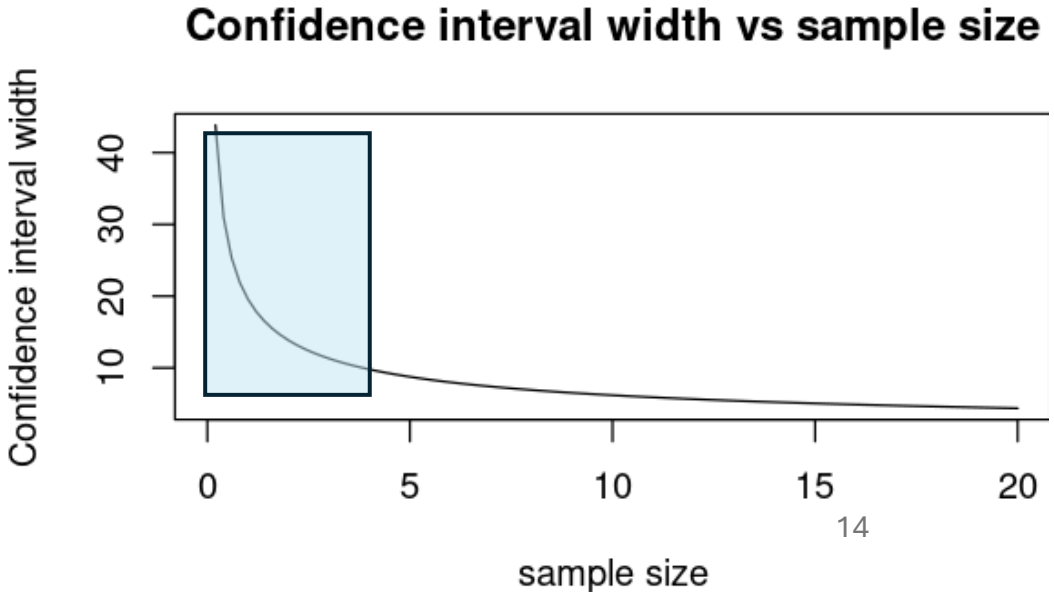
If the fit is strong the regression estimator works well (low variance)

If the fit is not strong the regression estimator has large variance but remains unbiased for large samples



$$y_i = sample\ fit_i + departure_i$$

$$\hat{\bar{Y}}_{regression} = \frac{1}{N}\sum_{all\ unit} sample\ fit_i + \frac{1}{n}\sum_{i\ in\ sample} departure_i$$



Part 1. Estimators based on sampling design

Code session 1A

The main conclusion after the code session 1 is that the regression estimator allows using auxiliary information making the estimates more precise.

However, there is limit. If the sample size becomes too small precision degrades and the variance gets large.

The estimator is unbiased even if the data does not follow a linear relationship. In that case, the estimator will do worse but it will remain unbiased. If the relationship can be approximated well by a linear function, it will do well.

Part 1. Estimators based on a model

There is an alternative way of estimating quantities for a population. This is what it is called model-based estimation.

In model-based estimation we assume that the population has some structure (it follows some model) and all our inferences are based on the assumption that such model holds. This is very different from the regression estimator for which we did not assume any relationship.

The fact that we will assume that the population follows some model will change completely the way we work.

Part 1. Estimation based on a model

In model-based estimation we assume that **every unit has an associated random variable with a distribution dictated by the model**. (Randomness comes from the model)

Unlike with techniques based on the sampling design, will not look at how our estimates changes if we take different samples. We will analyze the uncertainty that arise from the random components of the model.

All our inferences will be conditioned to:

1. **The sample that we observe**
2. **The model that we postulate for the population.**

Part 1. Estimation based on a model

In model-based estimation we conceptualize our population in a different way:

1. **For all population units, the variable of interest is a random variable following a specific model.**
2. **Target parameters such as means or totals will be sums or means of random variables→ Target parameters will also be random quantities with an inherent uncertainty.**

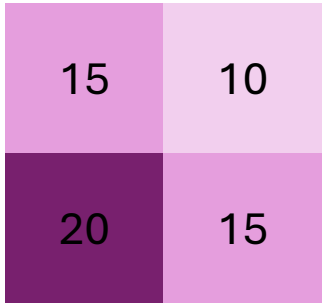
This starting point entails a very important and potentially risky assumption, **the model holds→To use these techniques, we will need a step to validate the model that we postulate for our population.**

Part 1. Estimation based on a model

In model-based estimation we conceptualize our population in a different way:

1. **For all population units, the variable of interest is a random variable following a specific model.**
2. **Target parameters such as means or totals will be sums or means of random variables→ Target parameters will also be random quantities with an inherent uncertainty.**

Example:



P95

Model:

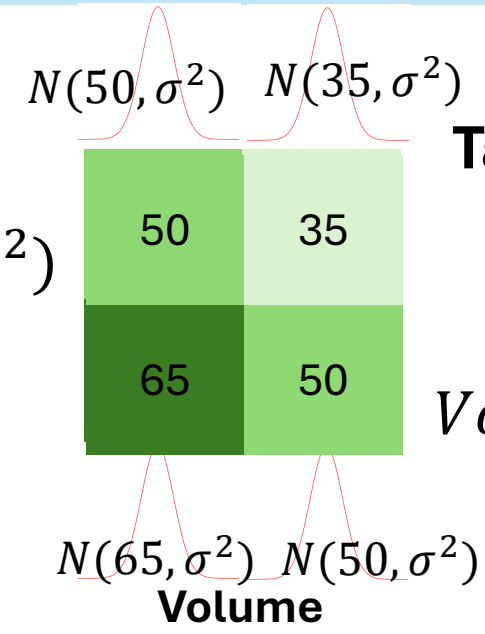
$$Vol_i = 3P95_i + 5 + \varepsilon_i \; ; \; \varepsilon_i \sim N(0, \sigma^2)$$

$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

$$Or$$

$$Vol_i \sim N(3P95_i + 5, \sigma^2)$$

$$Cov(Vol_i, Vol_j) = 0$$



Volume

Target parameter (total):

$$Vol = \sum Vol_i$$

$$Vol \sim N(\sum 3P95_i + 5 \, , 4\sigma^2)$$

Part 1. Estimation based on a model

When working with model-based methods we will have a similar workflow but with some differences.

1. **Define our target population parameter:** e.g., Mean volume in the forest, mean volume in one stand… **We cannot know it and we will need a sample. This parameter will be a random variable**
2. **Collect a sample covering the maximum variability:** The sample should not be collected purposively, i.e., targeting the variable of interest.
3. **Postulate a model for the population using the sample data.**
4. **Validate the model**: Check that model assumptions hold.
5. **Calculate our estimates and uncertainties assuming the model holds:** Using the collected data, calculate estimators and uncertainty based on the model.

In this case, we do not have control on the model. We can postulate and test different models but they might or might not hold. We don't control them!!

Part 1. Estimation based on a model

Code session 1B

In model-based estimation we consider the randomness that arises from a model that we assume that holds for the population.

We need to test the model.

A model that does not account for between stand variations can result in biased estimates

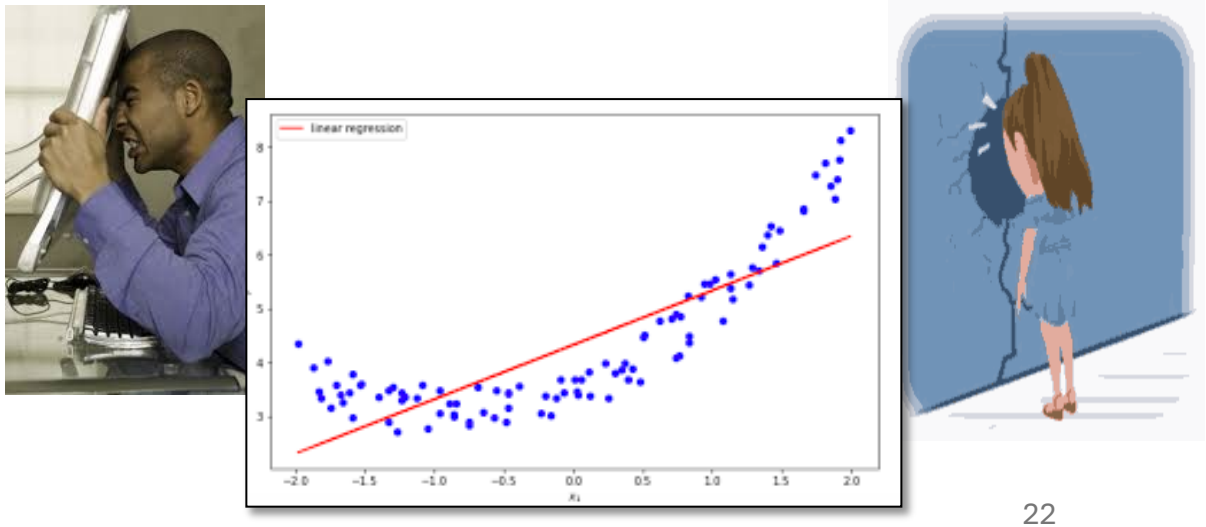
Part 1. Estimation based on a model

Model based estimators **are based on the assumption that the model holds** (importance of validation).

Those are my principles, and if you don't like them..., well, I have others.

Groucho Marx

Model assumptions are my principles, and **if my data does not like them ..., well, I need to find other model.**



Part 2. SAE, unit-level model

?

Part 2. SAE, unit-level model

In this section we will see the basic unit-level model for small area estimation

We will introduce the unit-level model and compare it with a linear model with fixed effects for stands

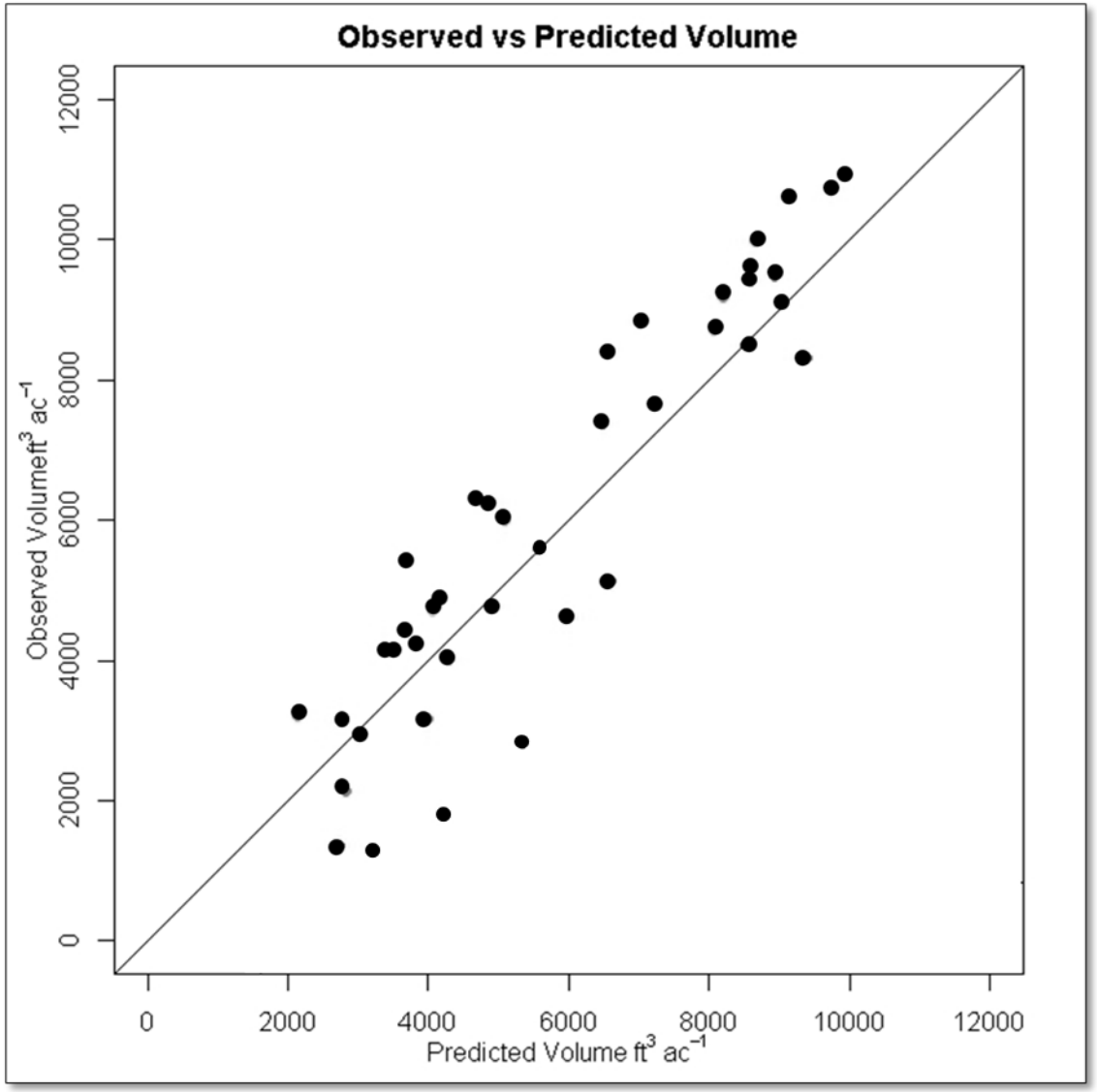
We will see code examples to obtain stand-level estimates using the unit level model using the R sae package.

We will see examples of how to fit the unit-level model using general R packages for mixed-effects models.

Part 2. SAE, unit-level model

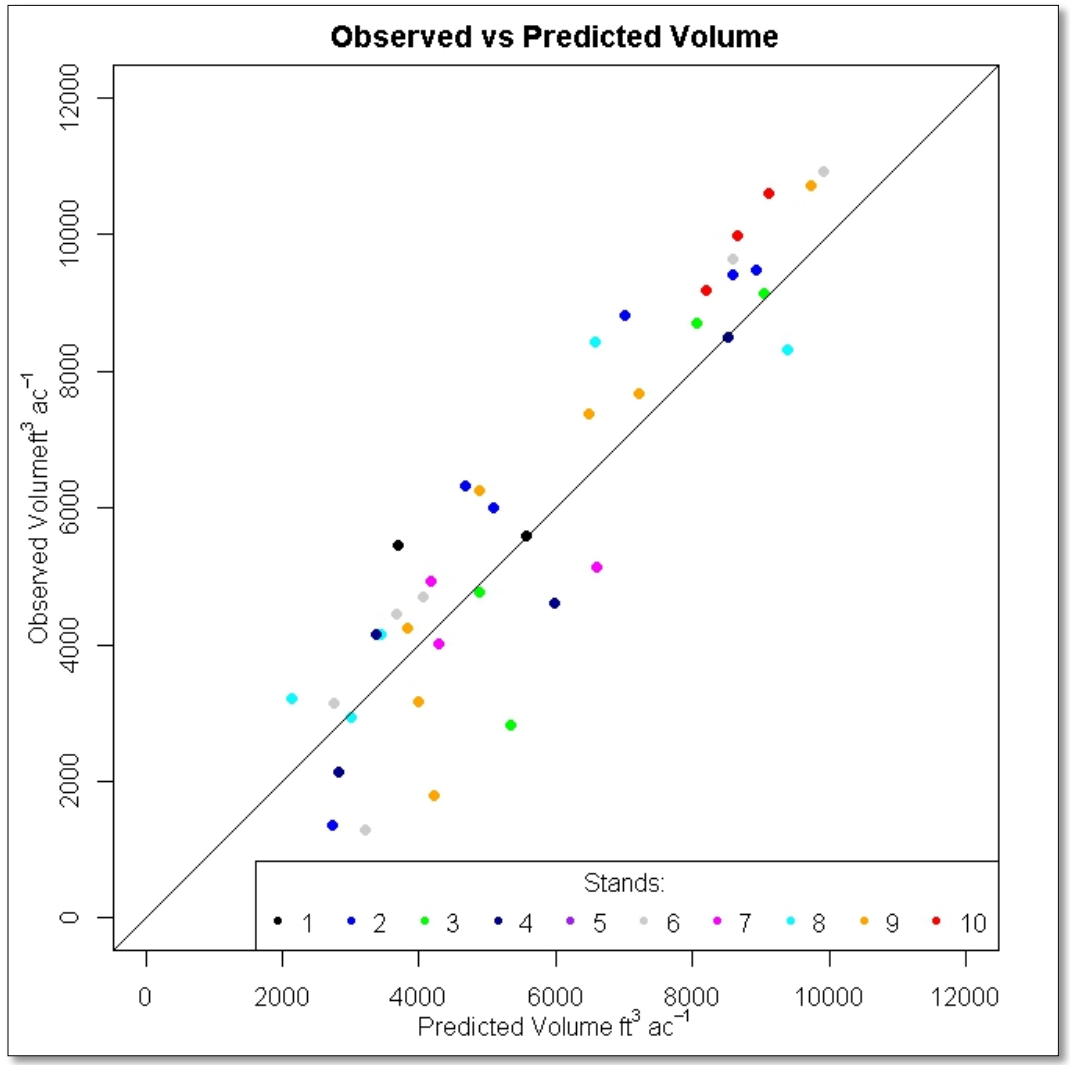
?

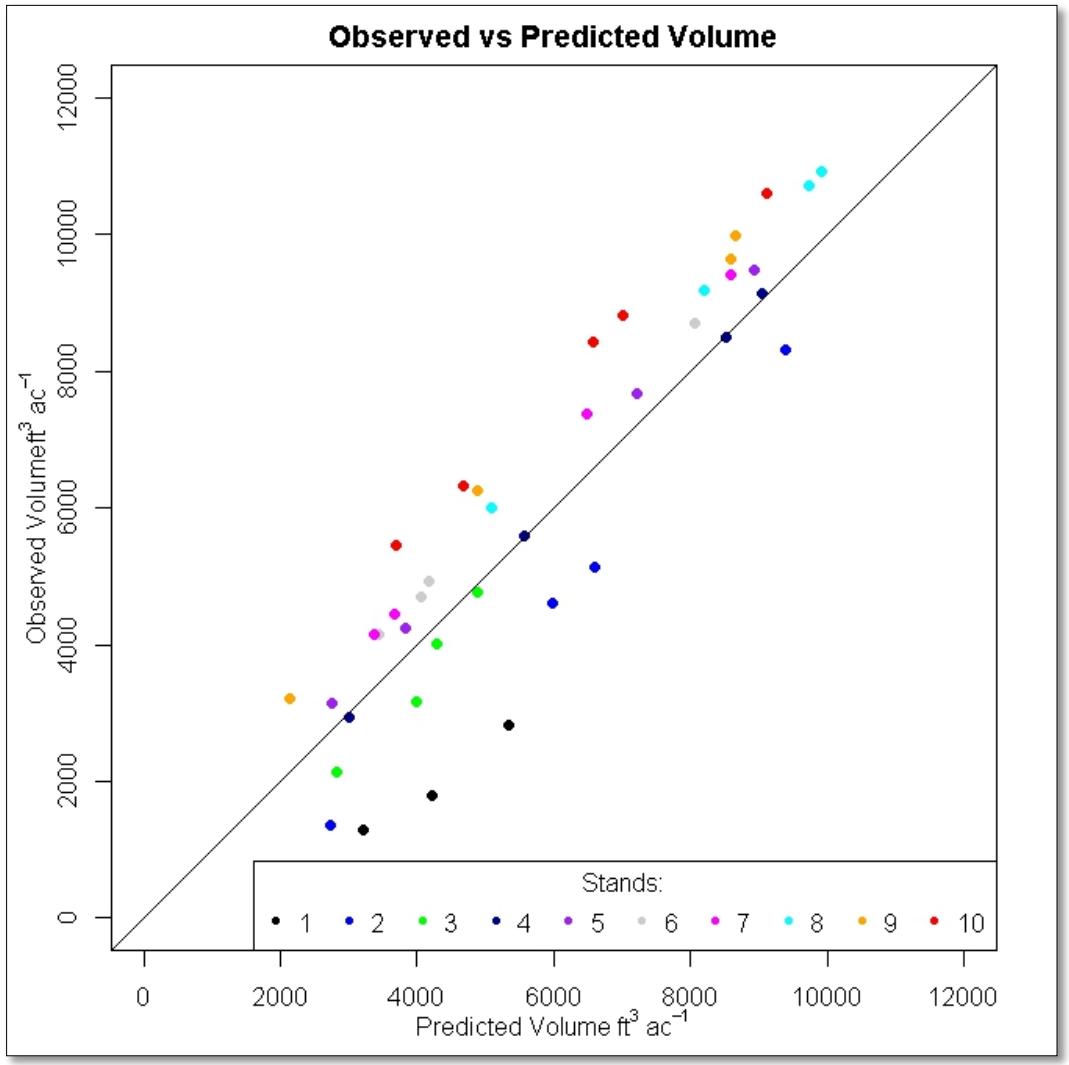
Part 2. SAE, unit-level model



Part 2. SAE, unit-level model

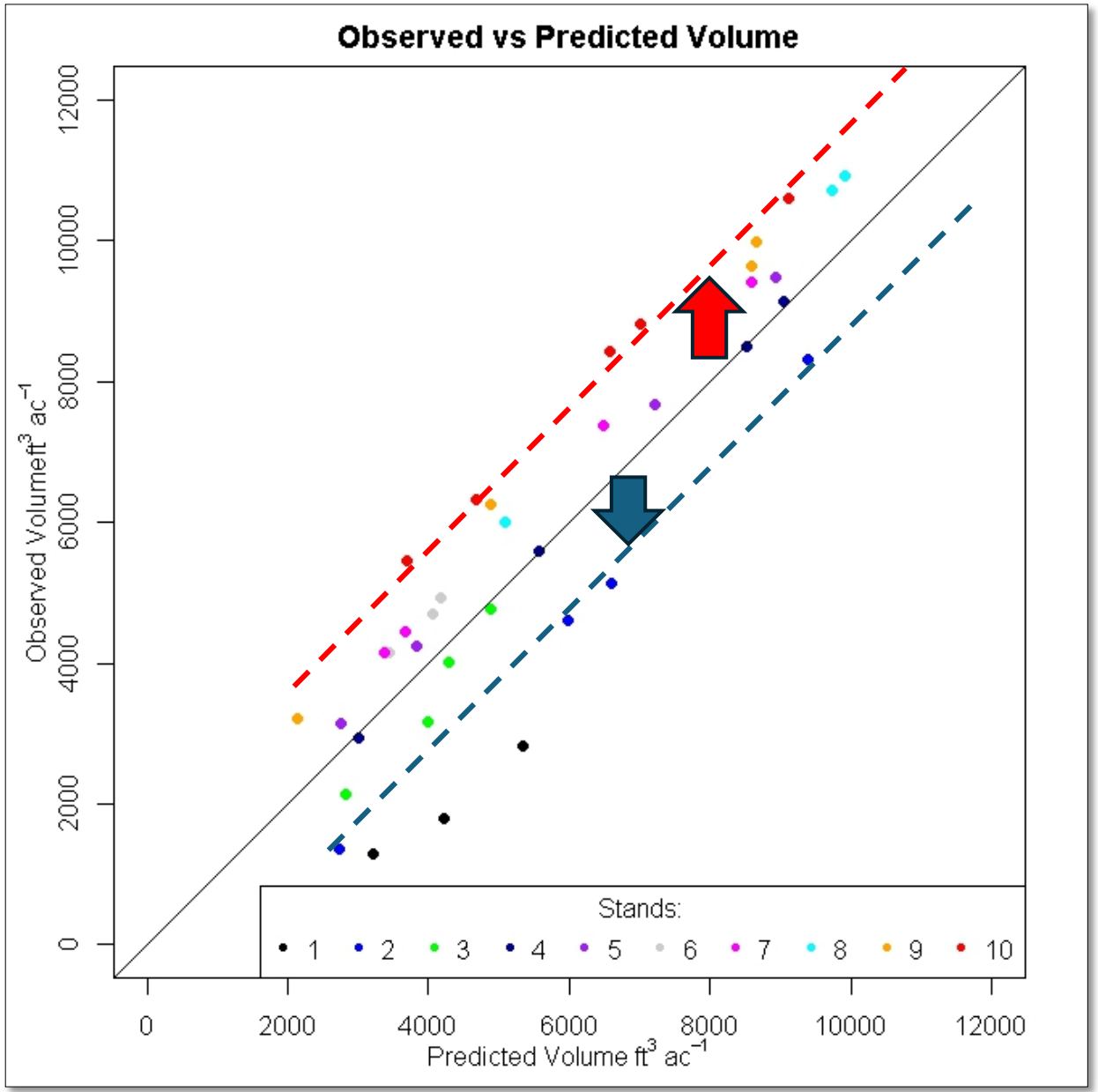
2 extreme cases for our model





Part 2. SAE, unit-level model

Solution: We need a model that “lifts” or “pushes down” the regression line depending on the stand



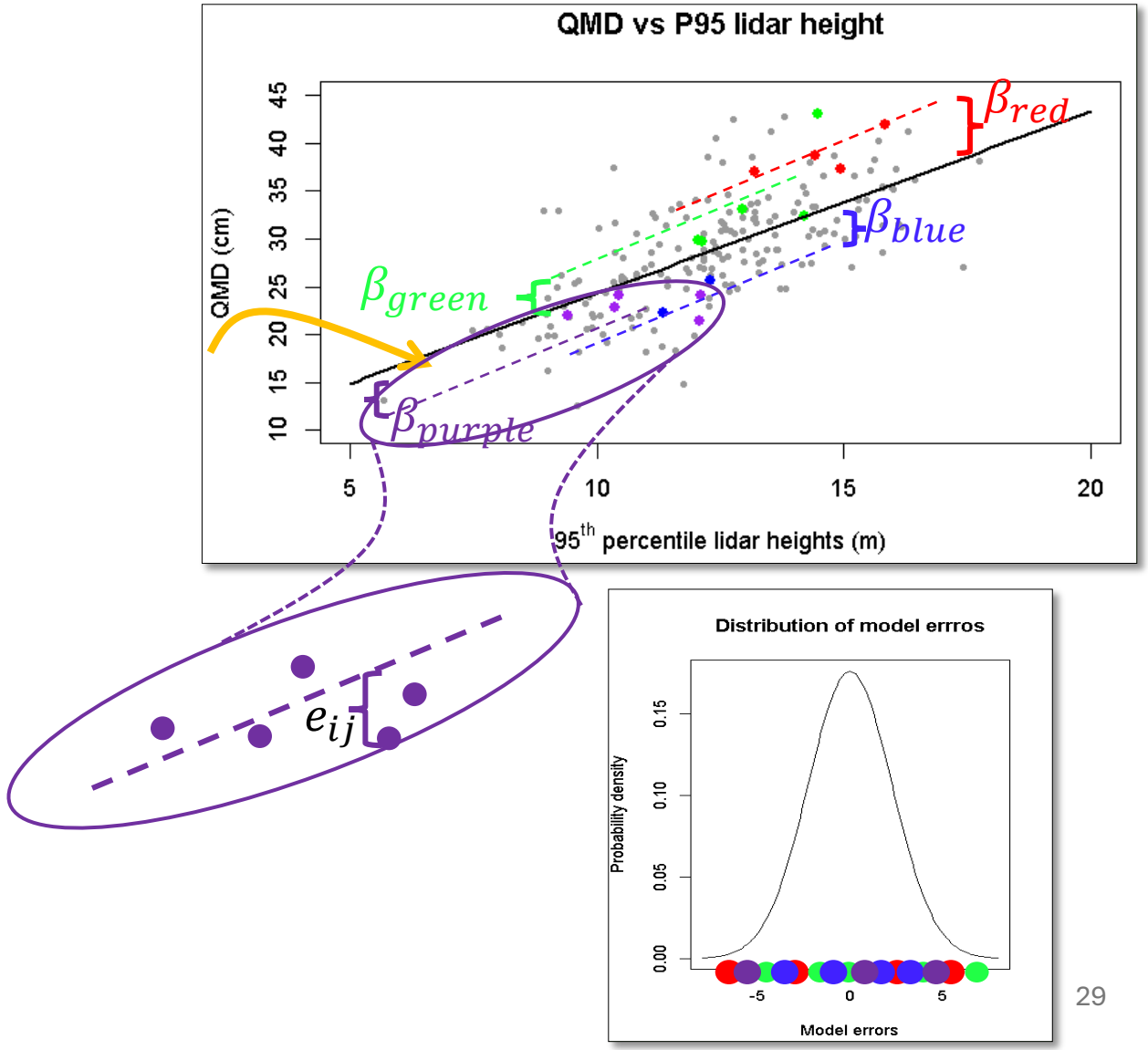
Part 2. SAE, unit-level model

- $y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \beta_i + e_{ij}$ Unexplained by aux info
- y_{ij} value of variable of interest for unit j in stand i, β_i stand shift

- $e_{ij}\ iid\ N(0, \sigma_e^2)$ Noise

- **The model implies an assumption about e_{ij}, but β_i can take any value, no assumption about its value.**

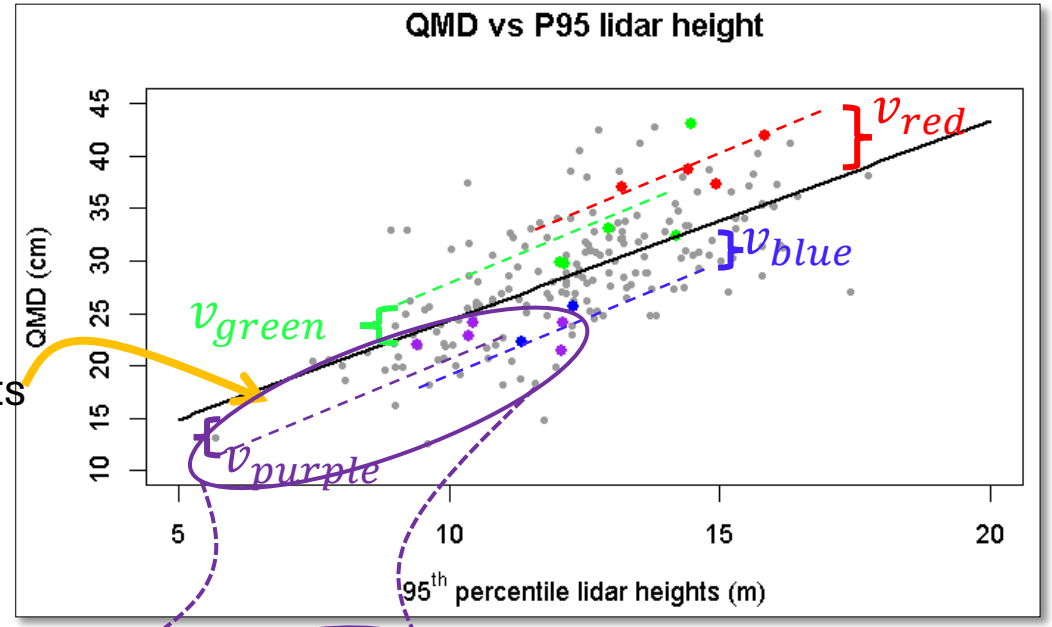
- **The coefficients β_i are similar to stand means→ Large variance**



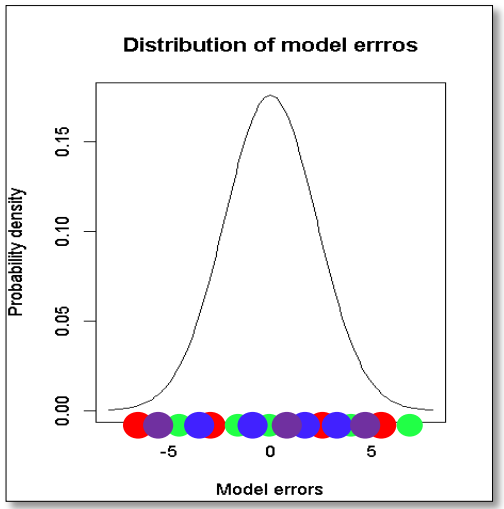
Part 2. SAE, unit-level model

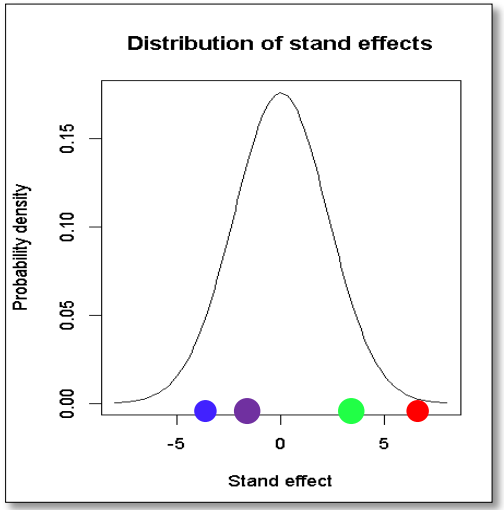
- $y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}$ + Unexplained by aux info $v_i + e_{ij}$

Fixed effects (indirect)



- y_{ij} value of variable of interest for unit j in stand i
- Stand effects $v_i \; iid \; N(0, \sigma_v^2)$ $\perp$ Noise $e_{ij} \; iid \; N(0, \sigma_e^2)$
- **<u>The model implies an assumption about v_i and e_{ij}</u>**

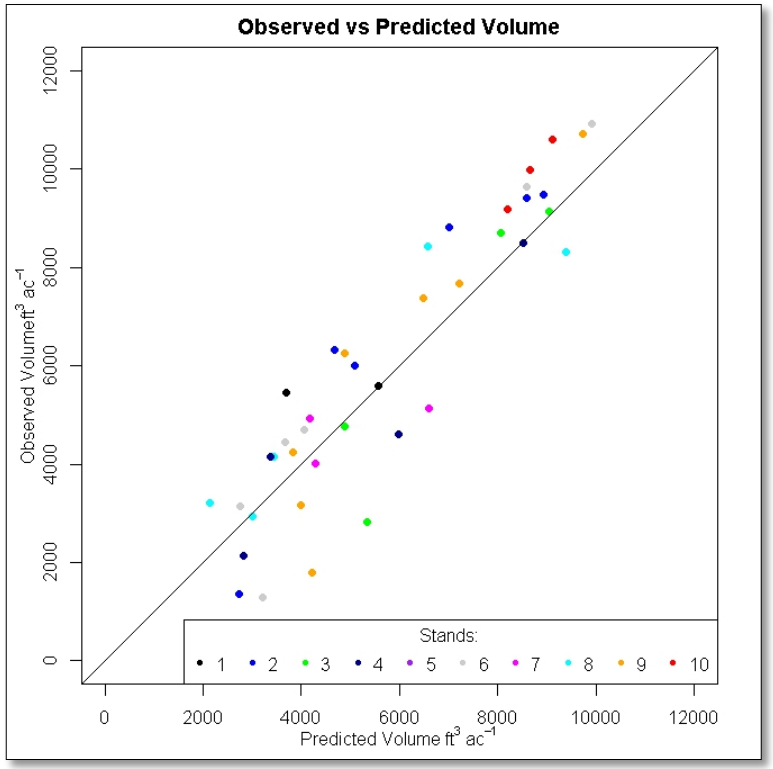


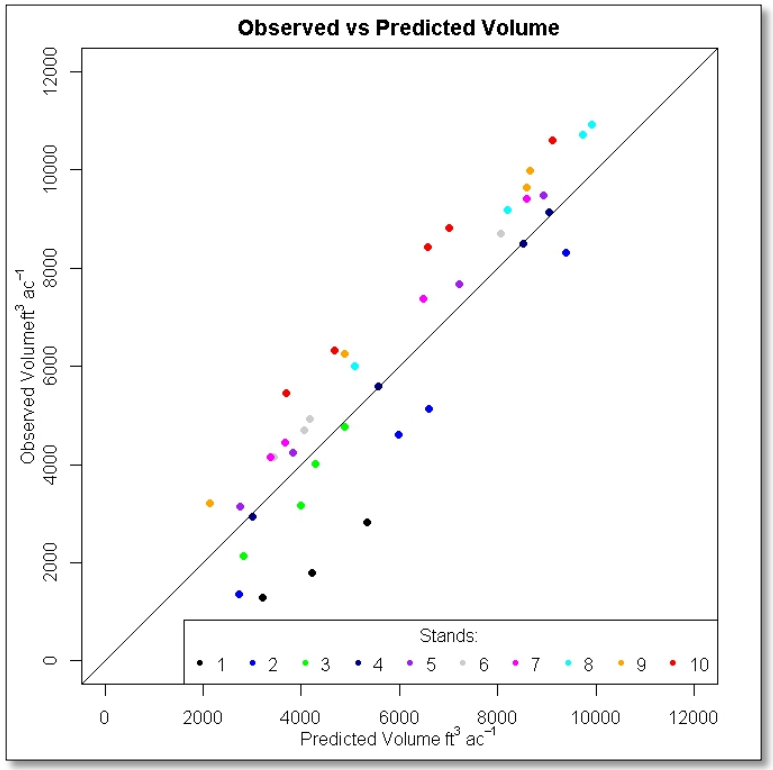


Part 2. SAE, unit-level model

Small when the between-area variability is small (we favor *global* information, indirect component)

Large when the between-area variability is large (we favor *local* information)



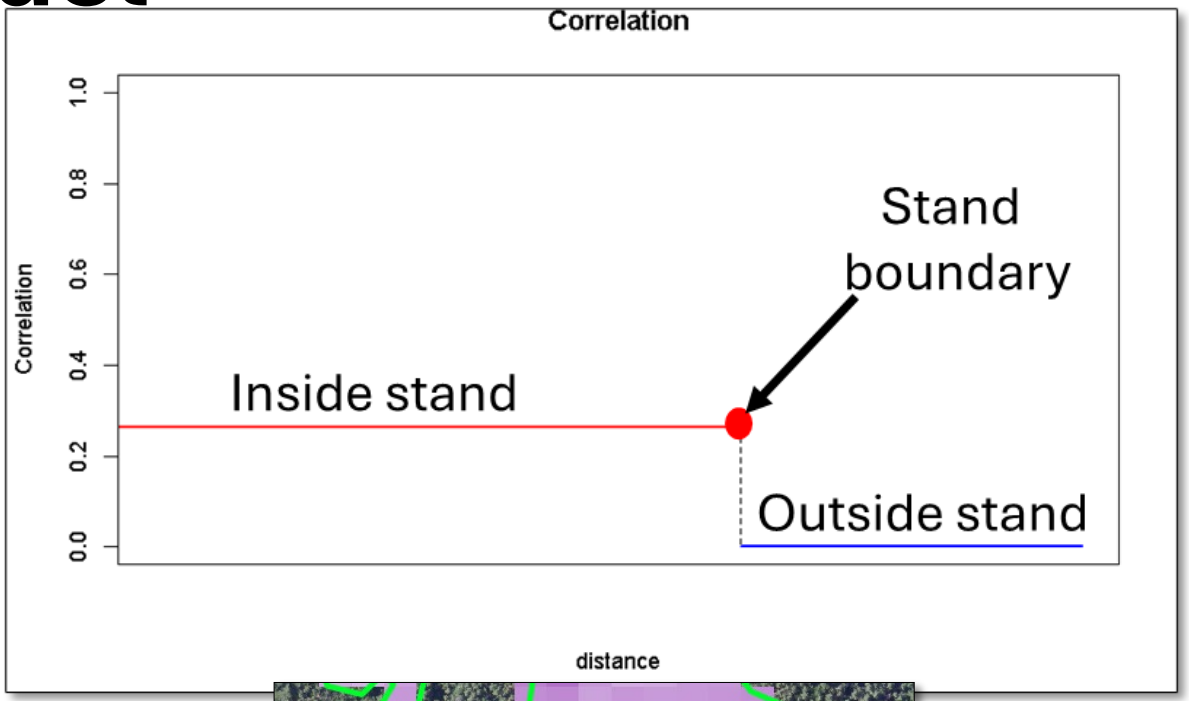


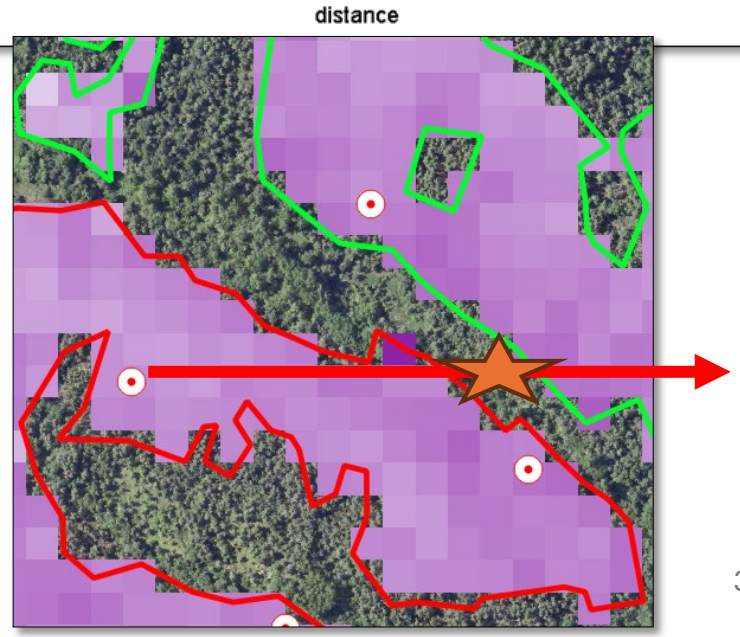
$$\frac{\hat{\sigma}_v^2}{\hat{\sigma}_e^2} = 0 \longrightarrow \frac{\hat{\sigma}_v^2}{\hat{\sigma}_e^2} = 1$$

Part 2. SAE, unit-level model

- For two units, the parts not explained by the model are correlated if both units belong to the same stand.
- The correlation is stronger when stand effects v_i have a large variance σ_v^2 (compared to σ_e^2)

$$Cor(v_i + e_{ij}, v_k + e_{km}) = \begin{cases} \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2}, if\ i = k \\ 0 \quad if\ i \neq k \end{cases}$$





Part 2. SAE, unit-level model

- If our model holds, the mean (μ_i) value for stand i, is:
  - $\bar{Y}_i = \frac{1}{N_i}\left[\left(\sum_{j \in stand\ i} \boldsymbol{x}_{ij}^T\right)\boldsymbol{\beta} + v_i + \sum_j e_{ij}\right]$

    Fixed component component

    Random
- $\hat{\bar{Y}}_i$ (EBLUP) is obtained minimizing the mean square error ($MSE = E\left[\bar{Y}_i - \hat{\bar{Y}}_i)^{\,2}\right]$), requires estimating the model parameters $\boldsymbol{\beta}$,σ_v^2 and σ_e^2.

- $\hat{\bar{Y}}_i = \frac{1}{N_i}\sum_j \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$, with $\hat{\beta}$, ${\hat{\sigma}_v}^2$ and ${\hat{\sigma}_e}^2$ obtained from, e.g., REML, and $\hat{v}_i$ is obtained from $\hat{\beta}$, ${\hat{\sigma}_v}^2$ and ${\hat{\sigma}_e}^2$

- v_i follow a distribution → **Structure that helps reducing the number of parameters**

Part 2. SAE, unit-level model

Unit-level Model

- **Model**
  - $y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij}$
- **Model parameters**
  - $\boldsymbol{\beta}$ *One per auxiliary variable plus intercept*
  - σ_v^2 *Variance of stand effects*
  - σ_e^2 *Variance of model errors*
- **Estimate of stand mean (total)**
  - $\hat{\bar{Y}}_i = \bar{\boldsymbol{x}}_{i,S}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$

Regression with dummy vars

- **Model**
  - $y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \beta_i + e_{ij}$
- **Model parameters**
  - $\boldsymbol{\beta}$ *One per auxiliary variable plus intercept*
  - β_i *Stand effects*
  - σ_e^2 *Variance of model errors*
- **Estimate of stand mean (total)**
  - $\hat{\bar{Y}}_i = \bar{\boldsymbol{x}}_{i,S}^T\hat{\boldsymbol{\beta}} + \hat{\beta}_i$

Example with 16 stands & 1 auxiliary variables

If we have 1 auxiliary variable for the unit level model we have to estimate 4 parameters (β_0, β_x, σ_v^2 and σ_e^2)

For the same problem the regression with dummy variables implies estimating 19 parameters (β_x, σ_e^2**, plus 16** β_i)

Part 2. SAE, unit-level model

Can we estimate the uncertainty of $\hat{\bar{Y}}_i$?

- The answer is yes.
- Approximately unbiased estimators of the **MSE** of $\hat{\mu}_i$ involve parametric bootstrapping or analytical derivations based on three quantities:
  - **First term**, obtained assuming that β and the variance parameters (σ_v^2, and σ_e^2 for unit level models, σ_v^2 and σ_v^2 and ψ_i for area level models) are known.
  - **Second term** because we need to estimate β (assuming variance parameters are known).
  - **Third term** because we need to estimate the variance parameters. Becomes small when the total number of small areas is large $o(m^{-1})$.
- Confidence intervals approximated as $\hat{\bar{Y}}_i \pm 2 * \sqrt{MSE}$

Part 2. SAE, unit-level model

Code session 3

If we want a fixed effect model with coefficients for every stand we need large sample sizes. If sample sizes are small estimates will have large variance because we do not assume any structure for the stand effects

The unit-level model reduces the number of parameters that we have to estimate while introducing stand effects.

The reduction in the number of parameters implies assuming some structure for the random effects and that is what helps in the estimation process.

Part 3. SAE, area-level model

In this section we will see the basic area-level model for small area estimation

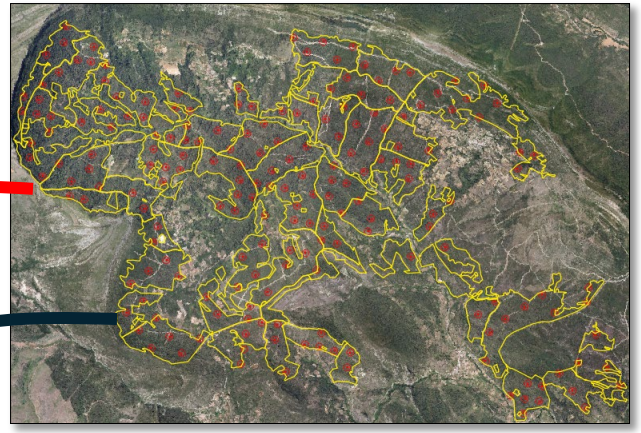
We will introduce the area-level model and compare it to the unit-level model.

We will see code examples to obtain stand-level estimates and uncertainty metrics using the are-level model using the R sae package.

We will discuss applications of the area level model

Part 3. SAE, area-level model

Field plots in stands, but bad GPS



STAND_ID	PLOT_ID	VOL	P95
1	1_1	127.8	20.2
1	1_2	135.9	23.1
1	1_3	147.3	24.8
2	2_1	110.1	18.9
2	2_2	108.5	18.7

Direct estimates (stand means) & their variance

STAND_ID	MEAN_VOL	VAR_VOL	MEAN_P95
1	137	31.99	22.7
2	104.3	33.64	18.8

Part 3. SAE, area-level model

- "Area-level" combine two components

1. Model True mean ~ Auxiliary information

$\bar{Y}_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + v_i$ *(we cannot fit this model, we don't know* $\bar{Y}_i$*)*

$v_i \sim N(0, \sigma_v^2)$

2. Assumptions about the direct estimator $\hat{\bar{Y}}_{i,D}$ (e.g. stand exams)

$\bar{Y}_i = \hat{\bar{Y}}_{i,D} + e_i$ *(e.g.* μ_i *this holds for large samples in SRS)*

$e_i \sim N(0, \psi_i^2)$ *(We estimate* ψ_i^2 *from sampling theory)*

3. Combine 1 and 2

$\hat{\bar{Y}}_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + v_i + e_i$ *(Note: only one index for both,* v_i *and* e_i*, but we can estimate* ψ_i^2 *and separate it from* σ_v^2*)*

Part 3. SAE, area-level model

- We obtain $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}_v^2$ using, for example, REML
- One SAE method for this purpose looks like

$$\hat{\mu}_i = \phi_i \underbrace{\left[\hat{\bar{Y}}_{i,D}\right]}_{(1)} + (1-\phi_i)\underbrace{\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}}_{(2)}$$

1. A direct estimator with large variance and small bias
   - The sample mean
2. An indirect estimator with small variance and (likely) large bias
   - The predicted mean using only fixed-effects

Part 3. SAE, area-level model

- What is ϕ_i?

$$\phi_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\psi}_i^2}$$

- Must be between 0 and 1
  - Large when the between-area variability is large (we favor *local* information)
  - Small when the between-area variability is small (we favor *global* information)
- This solution also minimizes the *MSE* for our area-level prediction

Part 3. SAE, area-level model

Code session 4

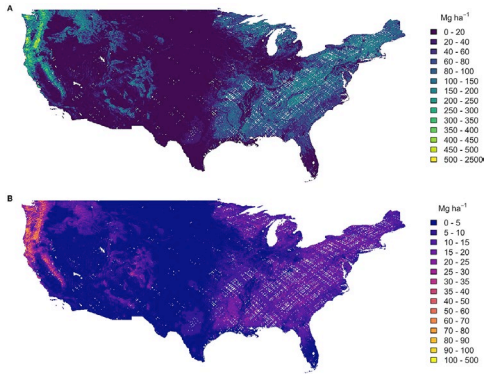
The area-level model relates means or totals for the stand and auxiliary information aggregated at the stand level (zonal stats for the stand)

With the area-level model our stand estimates are weighted averages of a synthetic (i.e., global) component that we assume is the same for the stand and a local component (sample means for the stand)

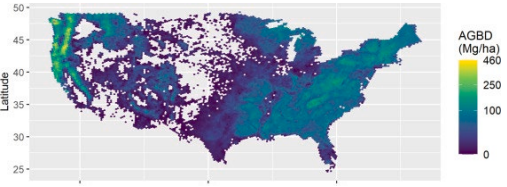
We reduce significantly the amount of information we have to use, but at the expense of loosing spatial detail.

Part 4. Future directions

National forest inventories



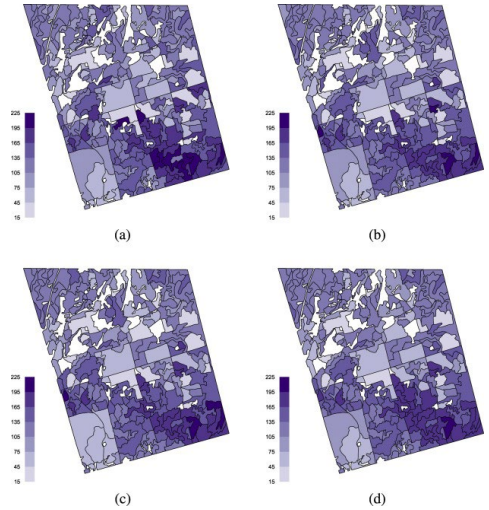
[May et al. 2023]



[Bruening et al. 2023]

- GEDI auxiliary information
- Multi-sensor auxiliary information
- Area level models
- AGB & Carbon

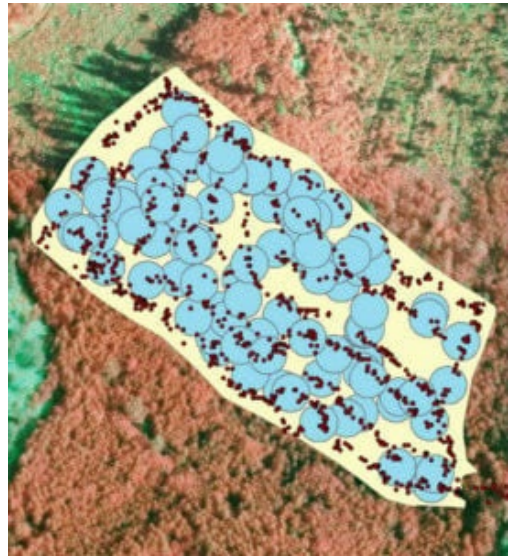
Stand level estimation



[Ver Planck et al. 2019]

- Multivariate models
- Spatial models
- Bayesian models
- Disaggregation of responses (e.g. dbh distributions)

Introducing harvester data



[Saukkola et al. 2019]

- Validation & development of lidar models
- Allometries

Frontiers Special Issue [Small Area Estimation in Forest Inventories: New Needs, Methods, and Tools | Frontiers Research Topic]

Part 4. Variable selection

Code session 5

There are many alternative methods for variable selection, in the code we present one that allows reducing the problem to a small list of candidate models that we evaluate sequentially.

As a personal recommendation, fully automatic methods imply some danger, always check the models in the candidate list using plots, tests, etc...

Thanks for your attention!!

francisco.mauro@uva.es