



Food and Agriculture
Organization of the
United Nations

SUSTAINABLE
DEVELOPMENT
GOALS

Small Area Estimation with Unit Level Models

Day 3 (Session 2)



**2022 Virtual Training on Data Disaggregation and Small Area Estimation for the SDGs –
20-23 September 2022, BPS and BAPPENAS Indonesia, CBS Nepal**

Unit-level SAE Models: Introduction

Small-Area estimation with unit-level approaches:

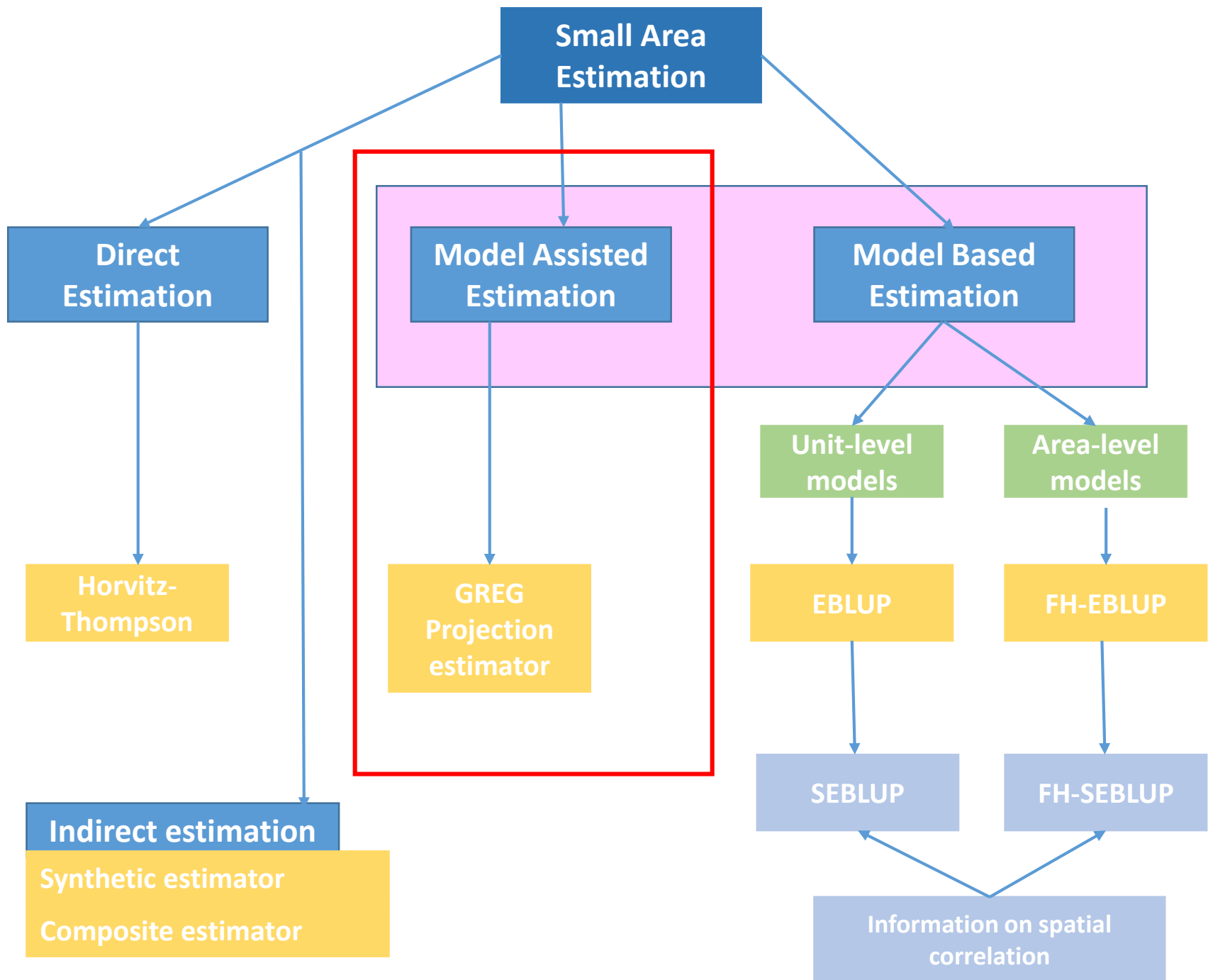
- Assume the availability of unit-specific auxiliary data $x_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ for each population element j and in each small area i .

Main characteristics:

- Unit-level approaches **allow producing estimates with a smaller MSE** compared to those obtained with direct estimators, but also with area-level approaches (assuming that the model is correct).
- Similarly to area-level approaches, **unit-level models consider the unexplained heterogeneity among the domains**, and allows producing estimates also for out-of-sample domains
- The estimator can be seen as a weighted average between the survey regression estimator and a regression-synthetic component. The **weight** assigned to the survey regression component **increases with increasing sample size**
- Contrarily to area-level approaches, sampling weights are not taken into account. In addition, unit-level approaches have stricter data requirements: the additional information needs to be available for all units in the population, and the auxiliary variables need to share the same definition.



A RECAP ON THE GENERALIZED REGRESSION ESTIMATOR



The GREG Estimator for domain estimation

Notation:

- Data: $\{y_{ij}, x_{ij}\}$, with j being the j – th unit in sample s , and i being the i – th small area
- $X_i = (X_{i1}, \dots, X_{ip})^T$ (known) population totals for auxiliary variables in the small area i
- Generalized REGression Estimator:

$$\hat{Y}_{i,GR} = \hat{Y}_i + (\bar{X}_i - \hat{X}_i)^T \hat{\beta}_i$$

- $\hat{\beta}_i = (\sum_{j \in s_i} w_{ij} x_{ij} x_{ij}^T / c_{ij})^{-1} (\sum_{j \in s_i} w_{ij} x_{ij} y_{ij} / c_{ij})$ with c_{ij} being a specified positive constant.

The GREG Estimator for domain estimation

Main characteristics:

- **GREG estimators** attempts to improve the precision of the traditional HT estimator by borrowing strength from relevant covariates through an adjustment of the initial sampling weights.
- This estimator is still approximatively design-unbiased, and should allow decreasing the design-variance and, thereby, the overall MSE.

Quiz time!



Question 1 – MULTI-SELECT

The GREG estimator is:

- A. The Generalized Regressive Estimator
- B. Approximately design-unbiased
- C. The Generalized Regression Estimator
- D. A model based estimator

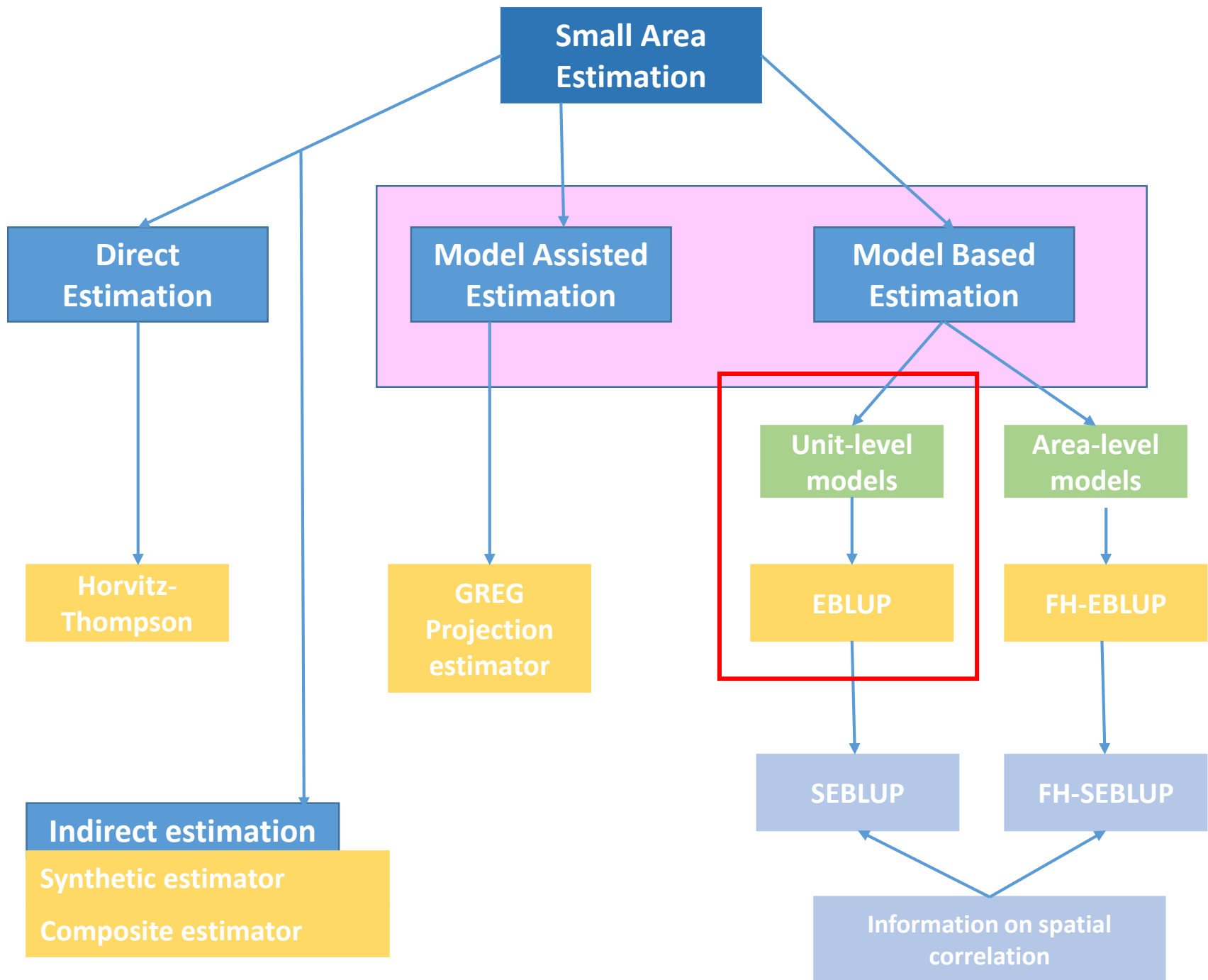
Question 1 – MULTI-SELECT

The GREG estimator is:

- A. The Generalized Regressive Estimator
- B. Approximately design-unbiased**
- C. The Generalized Regression Estimator**
- D. A model based estimator

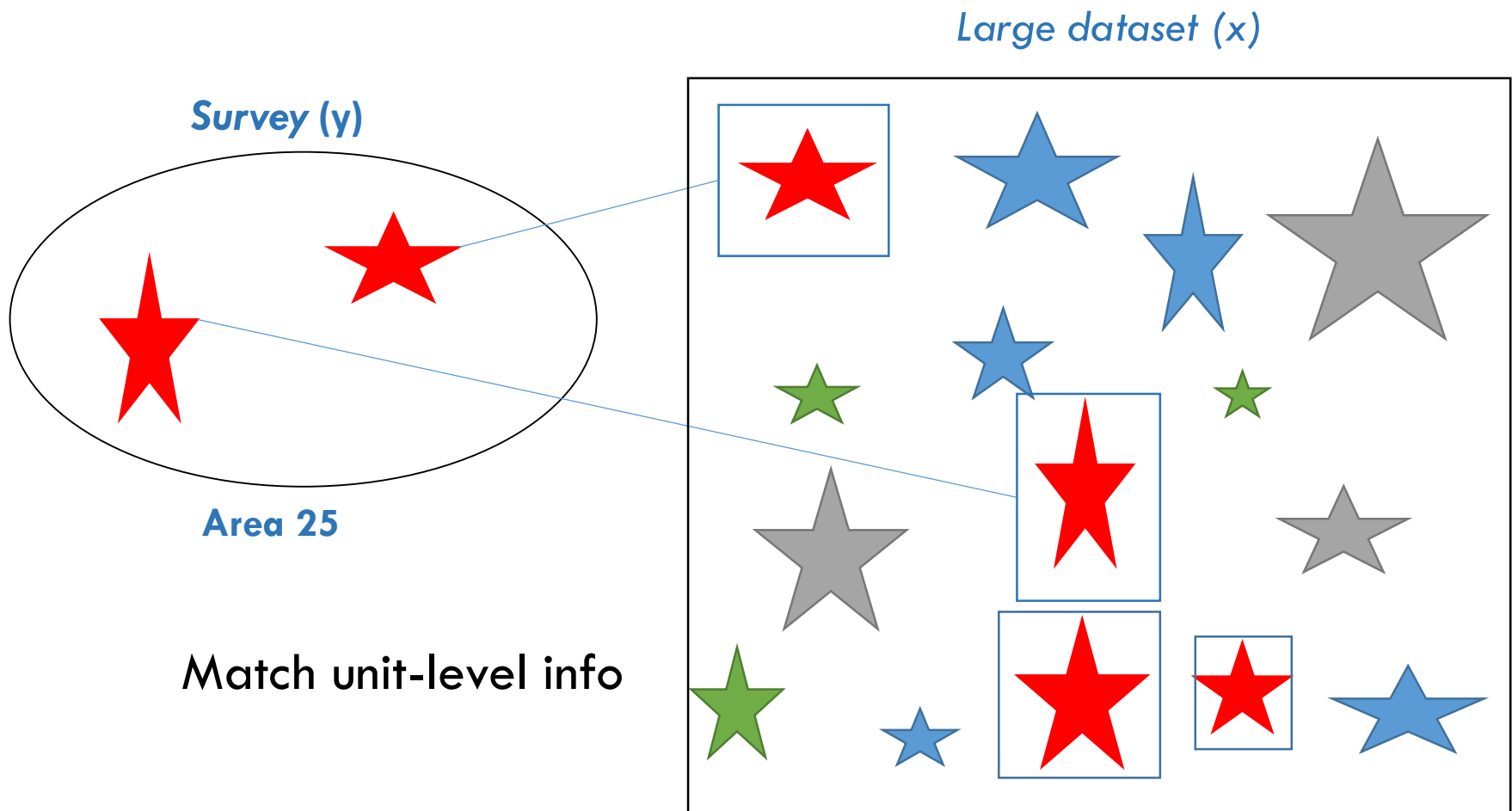


BASIC UNIT-LEVEL SAE MODEL



Model Based Approach – Unit Level

Unit-level approaches: these approaches are usually to be preferred when unit-level data is available. Contrarily to area-level approaches, this type of models do not consider the sampling design.



EBLUP with Unit-level approach

The underlying framework is as usual:

- The target population is partitioned into D domains: $\Omega = \bigcup_{i=1}^D \Omega_i$
- Ω_i ($i = 1, \dots, D$) is the i -th domain with population size N_i
- Sample data available on the target variable y such that $\mathbf{y} = [\mathbf{y}'_s, \mathbf{y}'_o]$.
- $\mathbf{y}_s$ is the vector of the n observed units, while $\mathbf{y}_o$ is the vector of $N - n$ out-of-sample units, where N is the overall size of the target population.
- D parameters of interest to be estimated θ_i ($i = 1, \dots, D$)
- X is the $p \times N$ matrix of auxiliary variables that are assumed to be known for all units in the population

EBLUP with Unit-level approach (2)

Superpopulation model for y :

$$y = X\beta + Zu + e$$

Alternatively, the same model can be expressed as:

$$y_{ij} = x_{ij}\beta + u_i + e_{ij}$$

Additional assumption: the model should hold both for the population and the sample

- $X\beta$ is the baseline component of the model
- Zu formalizes the difference between small areas

EBLUP with Unit-level approach (3)

Starting from the linear regression model for grouped individuals:

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + e_{ij}$$

- HP1: $\beta_{0i} = \beta_0 + u_{0d}$ and $\beta_{1i} = \beta_1 + u_{1d}$
- HP2: $E(u_{0d}) = E(u_{1d}) = 0$
- HP3: $V(u_{0d}) = \sigma_{u_0}^2$; $V(u_{1d}) = \sigma_{u_1}^2$; $Cov(u_{0d}, u_{1d}) = \sigma_{u_{01}}$

This leads to:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 * x_{ij}}_{X\beta} + \underbrace{u_{0d} + u_{1d} * x_{ij}}_{Zu} + e_{0ij}$$

$V(e_{0ij}) = \sigma_{e0}^2$

Additional HP: independence between the error term and the random effects

EBLUP with Unit-level approach (4)

Random intercept model:

- Assume that only the intercept is a random component on the second level.
- This represent the most common choice among small area estimation approaches
- It implies $u_{1d} = 0$ and hence

$$y_{ij} = x_{ij}^T \beta + u_i + e_{ij}$$

Under these assumptions, the covariance of two units j and j' is:

$$\text{Cov}(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_{u_0}^2 + \sigma_{e_0}^2 & \text{if } i = i' \text{ and } j = j' \\ \sigma_{u_0}^2 & \text{if } i = i' \text{ and } j \neq j' \\ 0 & \text{if } i \neq i' \text{ and } j \neq j' \end{cases}$$

Hence, units belonging to the same small area are assumed to be correlated, whereas units from different small areas are supposed to be independent

The Battese Harter and Fuller Model

The [BHF model](#) is the most common unit-level SAE model – also known as [nested error linear regression model](#)

The assumption on the covariance structure seen in the previous slide leads:

- $y_{ij} = x_{ij}^T \beta + u_i + e_{ij}$ with $i = 1, \dots, D$ and $j = 1, \dots, N_i$ (1)
- $u_i \stackrel{\text{iid}}{\sim} N(0, \sigma_u^2)$
- $e_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma_e^2)$
- $x_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ is a $p \times 1$ column vector of the covariates for the j – th unit within the i – th area
- β can be estimated with usual GLS, i.e.

$$\hat{\beta} = \left[\sum_{i=1}^D \sum_{j=1}^{n_d} x_{dj} (x_{dj} - \gamma_d \bar{x}_d)^T \right]^{-1} \left[\sum_{d=1}^D \sum_{j=1}^{n_d} (x_{dj} - \gamma_d \bar{x}_d)^T y_{dj} \right]$$

The Battese Harter and Fuller Model (2)

Under the BHF model, the small area means are

$$\hat{\mu}_i = \gamma_r' y_s + \gamma_r' [X_r \hat{\beta} + V_{rs} V_{ss}^{-1} (y_s - X_s \hat{\beta})]$$

For small sampling fractions is possible to particularize the BLUP formula for the target parameter

$$\hat{\mu}_i^{blup} = \begin{cases} x_{ij} \hat{\beta} + \gamma_i^w (\hat{\bar{Y}}_i^{dir} - \hat{\bar{X}}_i^{dir} \hat{\beta}) & \text{if } Y_{dj} \in r, \\ Y_{ij} & \text{if } Y_{dj} \in s. \end{cases} \quad (2)$$

Where

$$\gamma_d^w = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{w_d}}$$

Upper-case notation refers to population values while lower-case notation to sample values; a hat indicates that the variable is estimated while the bar denotes the mean

The Battese Harter and Fuller Model (3)

$$\hat{\mu}_i^{blup} = (1 - f_i) \left[\bar{X}_i \hat{\beta} + \gamma_i^w \left(\hat{Y}_i^{dir} - \hat{X}_i^{dir} \hat{b} \right) \right] + f_i [\hat{Y}_i + (\bar{X}_i - \hat{X}_i) \hat{\beta}] \quad (3)$$

Expressions (2) and (3) are equivalent

As (3) indicates, the **BHF estimator may be seen as a composite estimator** of

- The survey regression estimator
- The EBLUP of the regression synthetic component

Note that σ_u^2 and σ_e^2 are unknown and therefore $\hat{\mu}_i^{blup}$ is not calculable.

If we substitute σ_u^2 and σ_e^2 by convenient estimators, we have the Empirical Best Linear Unbiased Predictor (EBLUP) denoted by $\hat{\mu}_i^{eblup}$.

$$\hat{\mu}_i^{eblup} = (1 - f_i) \left[\bar{X}_i \hat{\beta} + \hat{\gamma}_i^w \left(\hat{Y}_i^{dir} - \hat{X}_i^{dir} \hat{b} \right) \right] + f_i [\hat{Y}_i + (\bar{X}_i - \hat{X}_i) \hat{\beta}]$$

$$\hat{\mu}_i^{eblup} = \frac{1}{N_i} \sum_{j \in s \cap n} Y_{ij} + \frac{1}{N_i} \sum_{j \in r \cap n} \{x_{ij} \hat{\beta} + \hat{\gamma}_i^w (\hat{Y}_i^{dir} - \hat{X}_i^{dir} \hat{\beta})\}$$

The Battese Harter and Fuller Model (4)

For finite populations with non-negligible sampling fractions, expressions (2) and (3) should be replaced with:

$$\hat{\mu}_i^{blup} = \frac{1}{N_i} [\sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} \hat{y}_{ij}] \quad (4) \text{ with}$$

$$\hat{y}_{ij} = x_{ij}^T \hat{\beta} + \hat{u}_i \quad (5)$$

Equation (4) is a general representation of an empirical best predictor (EBP) as well. The EBP generally comprises two parts

1. The sum of the observations for the sampled units
2. The sum of the predictions for the non-sampled units

After producing SAE estimates, the following step consist in assessing their MSE, that is evaluated with standard software.

Recap on BHF model and EBLUP estimators

- The unit-level EBLUP is a **model-based estimator** that can be used also when only the **average population values** of auxiliary variables are known.
- In many applications, the EBLUP perform better than usual design based estimators in terms of **coefficient of variation** (estimates have smaller confidence intervals).
- The **shrinkage factor** $\hat{\gamma}_i$ plays an important role in balancing the **survey regression component** with the **regression synthetic component**. When $\hat{\sigma}_u^2$ is small, also $\hat{\gamma}_i$ is small and, as a consequence, a higher weight is assigned to the regression synthetic part of the estimator. The same applies to small values of the area specific sample size n_d .

Recap on BHF model and EBLUP estimators (2)

- The estimator resulting from the **BHF model** is **not design-consistent for general survey designs**, as the survey regression estimator does not take the sampling weights into account.
- In addition, the **EBLUP** is **not model unbiased when conditioning on u_d** , as this would imply assuming fixed intercepts in the different areas.

Recap on BHF model and EBLUP estimators (3)

Main limitations

- The **assumption of normality is needed** both for area and individual effects. However, sensitivity analysis can show that the model is robust against non-normality when the symmetry of these distribution holds.
- The **EBLUP** is not design-unbiased, meaning that under complex survey design the estimates could suffer from bias.
- The parameters of interest in out of sample areas cannot be estimated. The EBLUP requires at least two observations per area.
- Extension of this model are not easy to implement, due to the complex derivation of the MSE.

Some extensions of the basic unit-level model

Some extensions/improvement of the **basic EBLUP**

- Spatial processes (CAR and SAR models)
- Time processes
- Spatio-temporal processes
- Binary and count models
- Multiple random-effect (e.g. R package mind)

Quiz time!



Question 1 – MULTI-SELECT

The Battese Harter and Fuller unit level model

- A. Assumes normality only for the error terms
- B. Is the most common SAE unit-level model
- C. Assumes normality of both the error terms and the random effects
- D. Is called nested error linear regression model

Question 1 – MULTI-SELECT

The Battese Harter and Fuller unit level model

- A. Assumes normality only for the error terms
- B. Is the most common SAE unit-level model**
- C. Assumes normality of both the error terms and the random effects**
- D. Is called nested error linear regression model**

Question 2 – MULTI-SELECT

The EBLUP resulting from a unit-level SAE model is:

- A. A weighted average of the survey regression estimator and the regression-synthetic part
- B. Can be seen as a model-based composite estimator
- C. Is design-unbiased
- D. Is a model assisted estimator

Question 2 – MULTI-SELECT

The EBLUP resulting from a unit-level SAE model is:

- A. A weighted average of the survey regression estimator and the regression-synthetic part**
- B. Can be seen as a model-based composite estimator**
- C. Is design-unbiased
- D. Is a model assisted estimator

Question 3 – SINGLE-SELECT

Is it possible to predict the parameter of interest in out-of-sample domains using unit-level models?

- A. Yes
- B. No

Question 3 – SINGLE-SELECT

Is it possible to predict the parameter of interest in out-of-sample domains using unit-level models?

- A. Yes
- B. No

Question 4 – SINGLE-SELECT

How many auxiliary variables can be included in unit-level models?

- A. 1
- B. 10
- C. Indefinite number
- D. Maximum $n-1$, where n is the number of units in the sample

Question 4 – SINGLE-SELECT

How many auxiliary variables can be included in unit-level models?

- A. 1
- B. 10
- C. Indefinite number
- D. Maximum $n-1$, where n is the number of units in the sample**

Question 5 – SINGLE-SELECT

Can area-level auxiliary information be included within unit-level models?

- A. Yes
- B. No

Question 5 – SINGLE-SELECT

Can area-level auxiliary information be included within unit-level models?

A. Yes

B. No

Unit-level models: extensions to non-linear indicators

The BHF model only supports the estimation of means and totals

Suitable extensions, based on the nested error linear regression model allow the estimation of **non-linear indicators such as proportions**

Two are the most popular approaches:

- The World Bank or **ELL method**
- The **Empirical best predictor**

Both approaches are commonly used for **poverty mapping**

Unit-level models: extensions to non-linear indicators

General idea of these extensions:

- The survey data is used to fit a model relating the variable of interest and the auxiliary information at the unit level which results into estimates of $\hat{\beta}$, $\hat{\sigma}_u^2$, $\hat{\sigma}_e^2$ of the model parameters.
- The model relation, the estimated parameters and the auxiliary information from the additional data source at unit level are used to produce predictions of the variable of interest for every unit in every estimation domain.
- The predicted values are then used to estimate the parameter of interest at the required domain level.

Unit-level models: extensions to non-linear indicators (2)

While the basic **BHF unit-level model** only requires means of the auxiliary information, the ELL and the EBP require auxiliary information for all units in all domains. Consequently, the higher flexibility in possible indicators comes along with stronger data requirements

Method	PROS	CONS
ELL	▪ Allows estimating any indicator expressed as a function of the variable of interest ▪ More accurate when the number of domains is large and there are non-sampled domains	▪ Assumes homogeneity in the small domains ▪ Results can be affected by individual outliers
EBP	▪ Allows estimating any indicator expressed as a function of the variable of interest ▪ Better performance compared to ELL regarding the MSE, in case of significant heterogeneity between domains	▪ The sampling design is not considered ▪ Results can be affected by individual outliers or lack of normality ▪ It assumes homogeneity in clusters

Thank you!

