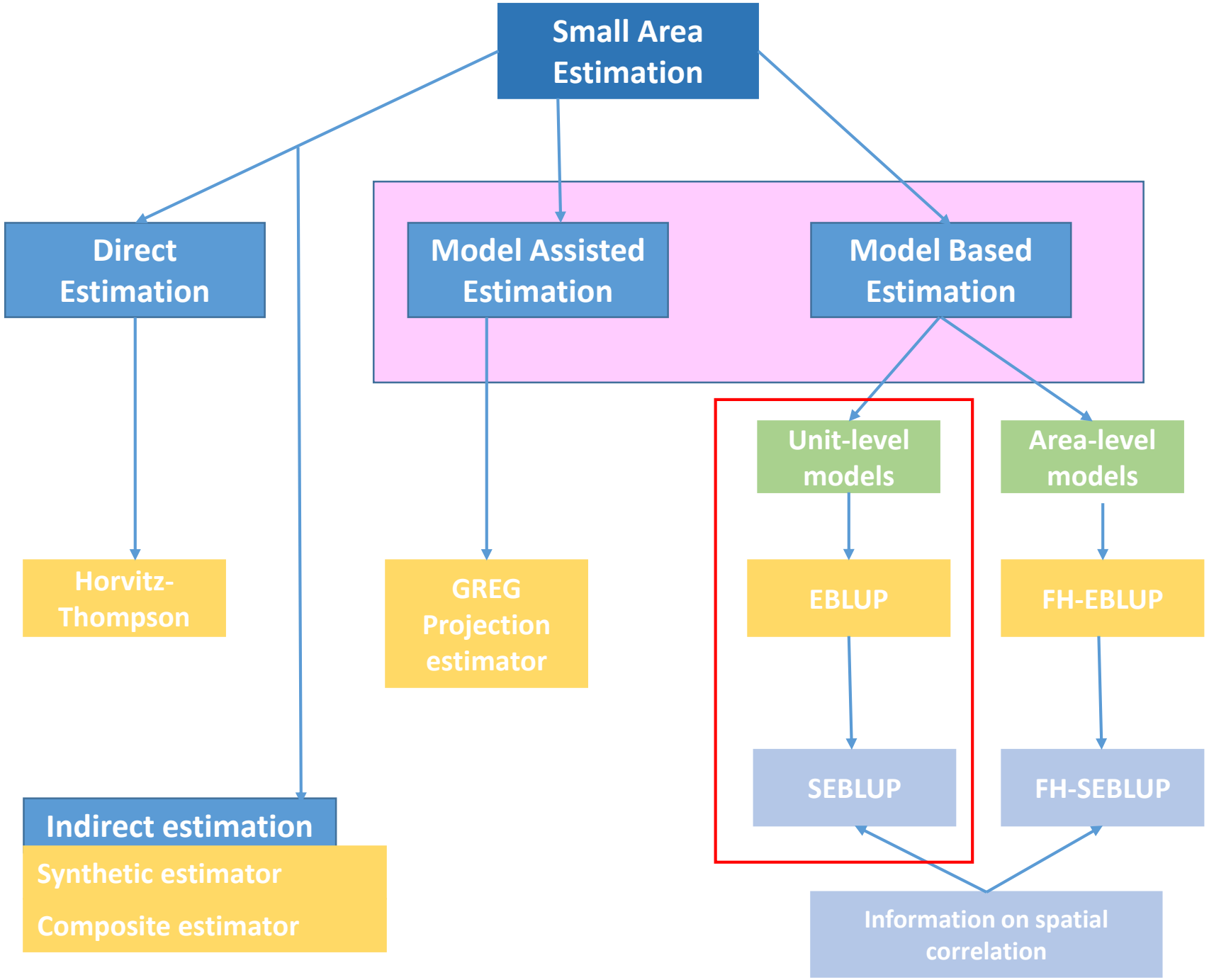
Food and Agriculture Organization of the United Nations

SUSTAINABLE DEVELOPMENT GOALS

Small Area Estimation with Unit Level Models

2022 Virtual Training on Data Disaggregation and Small Area Estimation for the SDGs –
22-25 November 2022, *INStaD Benin, INSTAT Mali, Stats South Africa, Statistics Botswana*


Small Area Estimation
Direct Estimation
Model Assisted Estimation
Model Based Estimation
Unit-level models
Area-level models
Horvitz-Thompson
GREG Projection estimator
EBLUP
FH-EBLUP
SEBLUP
FH-SEBLUP
Indirect estimation
Synthetic estimator
Composite estimator
Information on spatial correlation

Unit-level SAE Models: Introduction

Small-Area estimation with unit-level approaches:

- Assume the availability of unit-specific auxiliary data $x_{ij} = \left(x_{ij1}, \ldots., x_{ijp}\right)^T$ 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 of the mean 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
- $X_i = (X_{i1}, \ldots, X_{ip})^T$ (known) population totals for auxiliary variables in the small area i
- Generalized REGression Estimator:

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

- $\hat{\beta}_i = \left(\sum_{j \in s_i} w_{ij} x_{ij} x_{ij}^T / c_{ij}\right)^{-1} \left(\sum_{j \in} w_{ij} x_{ij} y_{ij} / c_{ij}\right)$ 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.

BASIC UNIT-LEVEL SAE MODEL

Model Based Approach – Unit Level

Unit-level approaches for means and totals:

- Very popular models in **poverty mapping** which is one of the most common applications of SAE

- **Unit-level data is assumed:** the variable of interest (e.g. the household income, the total quantity produced, etc.) is contained in the survey data, and auxliary variables with predictive power are available in the unit-level survey data. The same auxiliary variables need also to be available as domain means from an additional data source that is not affected by sampling error.

- **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**: $\boldsymbol{\Omega = \cup_{i=1}^{D} \Omega_i}$
- $\boldsymbol{\Omega_i}$ $(i = 1, \dots, D)$ is the **i-th domain** with population size $\boldsymbol{N_i}$
- Sample data available on the target variable y such that $\boldsymbol{y = [y_s', y_o']}$.
- y_s is the vector of the n observed units, while y_o is the vector of $N - n$ out-of-sample units, where N is the overal size of the target population.
- D parameters of interest to be estimated θ_i $(i = 1, \dots, D)$
- X is the pxN matrix of auxiliary variables that are assumed to be known for all units in the population

EBLUP with Unit-level approach (2)

Basic unit-level model – known as nested error linear regression model or BHF model:

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

Where the covariance of two units j and j' is:

$$Cov(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_{u_0}^2 + \sigma_{e0}^2 \; if \; i = i' \; and \; j = j' \\ \sigma_{u_0}^2 \; if \; i = i' \; and \; j \neq j' \\ 0 \; if \; i \neq i' \; and \; j \neq j' \end{cases}$$

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

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

ASSUMPTIONS:

- $u_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$
- $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$
- $x_{ij} = \left(x_{ij1,\ldots,} x_{ijp}\right)^T$ is a $px1$ column vector of the covariates for the $j-th$ unit within the $i-th$ area
- β can be estimated with usual GLS, i.e.

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

The Battese Harter and Fuller Model

The **Empirical Best Unbiased Predictor (EBLUP)** can be expressed as a weighted average (or linear combination) of the survey regression estimator (the GREG) and a regression-synthetic part:

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i\left[\bar{y}_i + (\bar{X}_i^T\hat{\beta} - \bar{x}_i^T\hat{\beta})\right] + (1-\hat{\gamma}_i)\bar{X}_i^T\hat{\beta}$$

- $\bar{y}_i$ is the sample mean of the variable of interest
- $\bar{X}_i^T$ and $\bar{x}_i^T$ are the means of the auxiliary information from the survey and the additional data source
- $\hat{\beta}$ is the vector of regression parameters
- $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$ is a weight measuring the amount of unexplained between-area variability to the total variability.

NB: when the sample size increases, the weight of the survey regression estimator increases

The Battese Harter and Fuller Model (4)

For finite populations with non-negligible sampling fractions, $f_i = \frac{n_i}{N_i}$, the EBLUP can be expressed as:

$$\hat{\theta}_i^{EBLUP} = f_i \bar{y}_i + (\bar{X}_i - f_i \bar{x}_i)^T \hat{\beta} + (1 - f_i)\left[\hat{\gamma}_i (\bar{y}_i - \bar{x}_i^T \hat{\beta})\right]$$

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 performs 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.

- 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 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 4 – 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 $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{u}}^{\boldsymbol{2}}$, $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\boldsymbol{2}}$ 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!