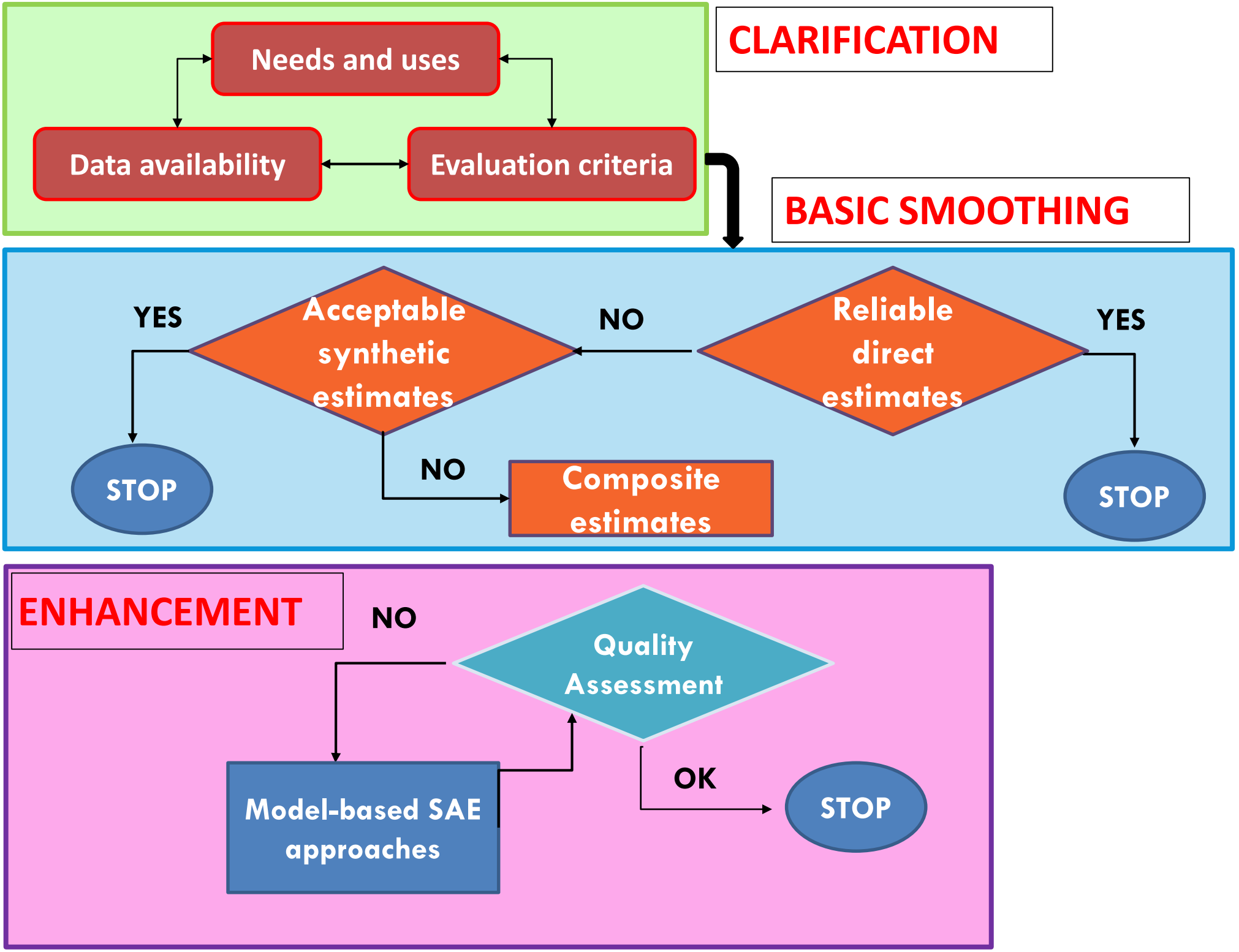
Food and Agriculture Organization of the United Nations

SUSTAINABLE DEVELOPMENT GOALS

Direct and indirect estimators for data disaggregation of SDG indicators

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*


CLARIFICATION
Needs and uses
Data availability
Evaluation criteria
BASIC SMOOTHING
YES
Acceptable synthetic estimates
NO
Reliable direct estimates
YES
STOP
NO
Composite estimates
STOP
ENHANCEMENT
NO
Quality Assessment
Model-based SAE approaches
OK
STOP

Data disaggregation in planned domains: direct estimators

- **Small Area Estimation** techniques often start from and/or include **direct estimators** → Good to provide a recap.

- **Horvitz-Thompson (HT) Estimator:** Most popular direct estimator to use if there is no auxiliary information available at the estimation stage.

Data disaggregation in planned domains: direct estimators (2)

Important Notation

Data: $\{y_j\}, j \in s$

- **HT estimator for the *Mean*:** $\hat{\bar{Y}} = \frac{\sum_{j \in s} w_j y_j}{\sum_{j \in s} w_j}$
- **HT estimator for the *Total*:** $\hat{Y} = \sum_{j \in s} w_j y_j$

Where:

- $w_j = \pi_j^{-1}$ is the basic **design weight**
- π_j is the **inclusion probability** of unit j, i.e. the probability of having the unit j included in the sample $s \in S$
- The weights w_j are **independent** from the y_j

Data disaggregation in planned domains: direct estimators (3)

- Let's **partition** the target population into D **domains**: $\boldsymbol{\Omega} = \bigcup_{i=1}^{D} \boldsymbol{\Omega_i}$

Where:

- $\boldsymbol{\Omega_i}$ $(i = 1, \dots, D)$ is the **i-th domain** with population size $\boldsymbol{N_i}$
- $\boldsymbol{s_i}$ is the **sample of units selected from the i-th domain** having size $\boldsymbol{n_i}$

The statistics of interest in the i-th domain can be formulated as:

- **Total:** $Y_i = \sum_{j=1}^{N_i} y_j = \sum_{\Omega_i} y_j$
- **Mean:** $\bar{Y}_i = Y_i / N_i$

- Data $\{y_{ij}\}; j \epsilon s_i; j = 1, \dots, n_i; i = 1, \dots, D$
- HT estimator of the mean in the i-th domain:

$$\hat{\bar{Y}} = \frac{\sum_{j \in s_i} w_{ij} y_{ij}}{\sum_{j \in s_i} w_{ij}}$$

Where $\boldsymbol{y_{ij}}$ are the **observed values** and $\boldsymbol{w_{ij}}$ indicates the **sampling weight** of unit j in domain (area) i.

Data disaggregation in planned domains: direct estimators (4)

- The estimator $\hat{\bar{Y}}_i$ is **design unbiased** → $\boldsymbol{B\left(\hat{\bar{Y}}_i\right) = E(\hat{\bar{Y}}_i) - \bar{Y}_i = 0}$

- $V\left(\hat{\bar{Y}}_i\right) = (1 - \frac{n_i}{N_i})\frac{S_i^2}{n_i}$, where $S_i^2 = \frac{\sum_{j \in s_i}(y_{ij} - \bar{Y}_i)}{n_i - 1}$ is the sample variance (needs to be estimated).

The magnitude of $V\left(\hat{\bar{Y}}_i\right)$ depends on three factors: $\frac{n_i}{N_i}$; S_i^2; and n_i

When the sample size is small (e.g. in unplanned domains), the design variance is likely to be large.

- **Possible solution:** use auxiliary data – whenever possible – to improve the reliability of estimates (decrease Variance).

- NB: There are many ways of using auxiliary information to improve estimates' precision (e.g. Ratio estimator, Regression estimator, Calibration estimator). In this course we only focus on the use of auxiliary information to produce **INDIRECT ESTIMATORS**

Quiz time!

Question 1 – MULTI-SELECT

The Horvitz-Thompson estimator for the estimation of totals and means:

A. Is the most popular direct estimator
B. Is the most popular indirect estimator
C. Can be used only when auxiliary variables are available
D. Is design-unbiased
E. Makes use of sampling weights

Data disaggregation in unplanned domains: indirect estimators

Let's consider the case in which direct estimates are not reliable:

- We can «reinforce» sample surveys: borrow strenght from related areas and/or related time periods.

- We need to enstablish a relationship between the target variable and the donors of strenght: **Implicit or explicit models**

Different types of indirect estimators:

- ***Domain Indirect:*** uses values from another domain but not from another time.
- ***Time Indirect:*** uses values from another time but not from another domain.
- ***Domain and Time Indirect:*** uses values both from another domain and time.

Data disaggregation in unplanned domains: indirect estimators

Indirect estimation approaches can be broadly classified into:

- **Model assisted Approaches:**
  - The main concern is unbiasedness. Estimators' properties are assessed with respect to the sampling design. This approach to indirect estimation is traditionally adopted because of its simplicity, applicability to general sampling designs, and potential of increasing estimates precision by borrowing information from similar small areas.

- **Model-based Approach:**
  - The finite population is treated as a random realization from a superpopulation and a suitable model for the variable of interest is proposed.

In this session, we focus on model-assisted approaches, while the reminder of the training will target model-based methods.

Data disaggregation in unplanned domains: indirect estimators

Under the model-assisted paradigm, there are two main categories of indirect predictors

- **Synthetic Estimators:**
  - A reliable direct estimator for a broad area, covering several small areas, is used to derive an indirect estimator for a small area.
  - *Produced under the assumption that the small areas have the same characteristics as the broad area.*

- **Composite Estimators:**
  - A linear combination between a direct estimator and a synthetic one *using a design-based approach or by assuming an explicit area or unit-level model.*
  - Represents a good compromise in terms of efficiency between the characteristics of the two components.

Indirect estimators: Synthetic

- **Simple assumption:** small and broad areas have the same characteristics
  - Example: the average crop yield is homogeneous in districts (small area) belonging to the same region (broad area)

- **Advantages of this approach:**
  - ✓ Simple and intuitive
  - ✓ Applicable to general sampling designs
  - ✓ Borrow strength from similar areas
  - ✓ Can be used to produce estimates for small areas with no sample observations

Indirect estimators: Synthetic (2)

Broad Area Ratio Estimator (BARE) of a Total

$$\widehat{Y}_{i,BA} = N_i * \frac{\sum_{j \in s} w_j y_j}{\sum_{j \in s} w_j} = N_i * \widehat{\bar{Y}}_{(BA)}$$

Where:

- N_i population size for i-th domain
- s sample
- w_j initial weight associated with j-th unit in the sample
- y_j value of the target value for j-th unit

The total value for the variable under study y for the large area is proportionally allocated in all small areas according to the population area sizes N_i.

Indirect estimators: Synthetic (2.1)

Example of BARE application

Let's consider the SDG Indicator 2.1.2 and suppose we want to estimate the total number of people in moderate or severe food insecurity at district level but the survey direct estimates have too poor precision.

When the only available additional information is the population sizes of the districts, then indirect estimates can be obtained by applying the broad area (regions) incidence of moderate or severe food insecurity to the district-level population.

Indirect estimators: Synthetic (3)

Linear Regression Estimator (LRE) of a Total

$$\widehat{Y}_{i,LRE} = N_i\left[\overline{y}_{s\,(BA)} + \left(\overline{X}_{i\,(BA)} - \overline{x}_{s\,(BA)}\right)\widehat{\beta}\right]$$

Where:

- $\overline{y}_s, \overline{x}_s$ sample mean of target variable and auxiliary variable
- $\overline{X}_i$ population mean of auxiliary variable
- $\widehat{\beta}$ linear regression parameters

- **ADVANTAGES:**
  - Easy computation: the only elements required are the local summary statistics of auxiliary variables and the parameters $\hat{\beta}$.
  - Can be used for estimation domains with no sample observations.
  - Unbiased, when the assumption that small areas have the same characteristics as the broad area is fulfilled.

- **NB:** Its performance depends on the relationship between the area-level parameter and the selected vector of auxiliary variables.

Indirect estimators: Synthetic (3.1)

Example of LRE application

Still considering the estimation of the total number of people in moderate or severe food insecurity at district level, suppose we have as auxiliary information the proportion of poor people both at district and broad area level (regions).

In this situation, the relationship between the variable of interest and the auxiliary variable can be exploited to obtain indirect estimates by simply:

- deriving the parameter $\hat{\beta}$ from any model used to formalize the relationship,
- then applying $\hat{\beta}$ to the elements of the formula in the previous slide.

Indirect estimators: Synthetic (4)

Ratio Synthetic Estimator (RSE)

$$\widehat{Y}_{i,RSE} = X_i * \frac{\sum_{j \in s} w_j y_j}{\sum_{j \in s} w_j x_j} = X_i \frac{\widehat{Y}_{(BA)}}{\widehat{X}_{(BA)}}$$

Where:

- X_i is the known total of the auxiliary variable for the small area i
- $\widehat{Y}_{(BA)}$ is the direct survey estimate of the total of the target variable at the broad area level.
- $\widehat{X}_{(BA)}$ is the direct survey estimate of the total of the auxiliary variable at the broad area level.

This synthetic estimator uses a broad area survey estimate $\widehat{Y}_{(BA)}$ and can be adopted when the value of a single auxiliary variable, X, is observed in the survey and its total is known from another source (e.g. a census, or a source that is not affected by sampling error) for each small area.

Indirect estimators: Synthetic (4.1)

Example of RSE application

Still considering SDG Indicator 2.1.2, suppose that from the survey we can estimate the total number of poor people in the broad area (regions) and that this total is known, from another source, for each district.

It is very similar to the previous example of synthetic estimator but now the ratio between the total number of people in moderate or severe food insecurity and the total number of poor people takes the place and the role of the parameter $\hat{\beta}$.

Indirect estimators: Synthetic (5)

Post-Stratified Synthetic Estimator (PSSE)

$$\widehat{Y}_{i,PSSE} = \sum_{g} N_{ig} * \frac{\widehat{Y}_g}{\widehat{N}_g}$$

Where:

- g cross-classifications or post-strata (e.g. $g = 1$ to 6 for three age groups by male and female)
- $\widehat{Y}_g$ direct survey national estimate (e.g. the Horvitz-Thomson estimate) of the target variable for cross-classification cell g;
- $\widehat{N}_g$ direct survey national estimate of the population size for cross classification cell g;
- N_{ig} known population size for cross classification cell g of the small area i.

Indirect estimators: Synthetic (5.1)

Example of PSSE application

Finally, let's suppose that a variable representing individuals' education level (with values from 0 to 3) is available in the survey, and that the population size at district-level for these four cross-classifications are known.

Then, indirect estimates can be obtained taking the cross-classifications into account.

Indirect estimators: Synthetic (6)

RECAP: Synthetic estimators are based on reliable estimates for a broad area including the small area of interest. They strongly rely on the assumption that small areas have similar characteristics of broad areas that include them.

MAIN ADVANTAGES:

- Can be applied when **sample data are not available** for the domain of interest – great advantage compared to **direct estimates**.
- Can be used even **when sampling was not involved** – broad area estimates should be based on a data source that is not affected by sampling error.
- The strenght of this approach is the **simplicity** – they are applicable to general sampling designs and allow borrowing strenght from areas with similar characteristics.

Indirect estimators: Synthetic (7)

LIMITATIONS

- Not appropriate when the **main assumption** (i.e., matching characteristics) is **not fullfilled**.
- **Over-shrinkage problem** - estimators generally display less between-area variation then they should.
- Very important that **good auxiliary information is available – synthetic estimators do not take into account possible misspecification of area level variables or correction in the variables**
- very important to take possible **selection effects** into consideration as far as possible, since they may cause systematic differences in the target variable between sample and population (*bias*). Adjustments to ensure coherence of estimates at different levels:

$$\hat{Y}_{i,adj} = \frac{\hat{Y}_i}{\sum_i \hat{Y}_i} \hat{Y}$$

A special case of synthetic estimator: the projection estimator

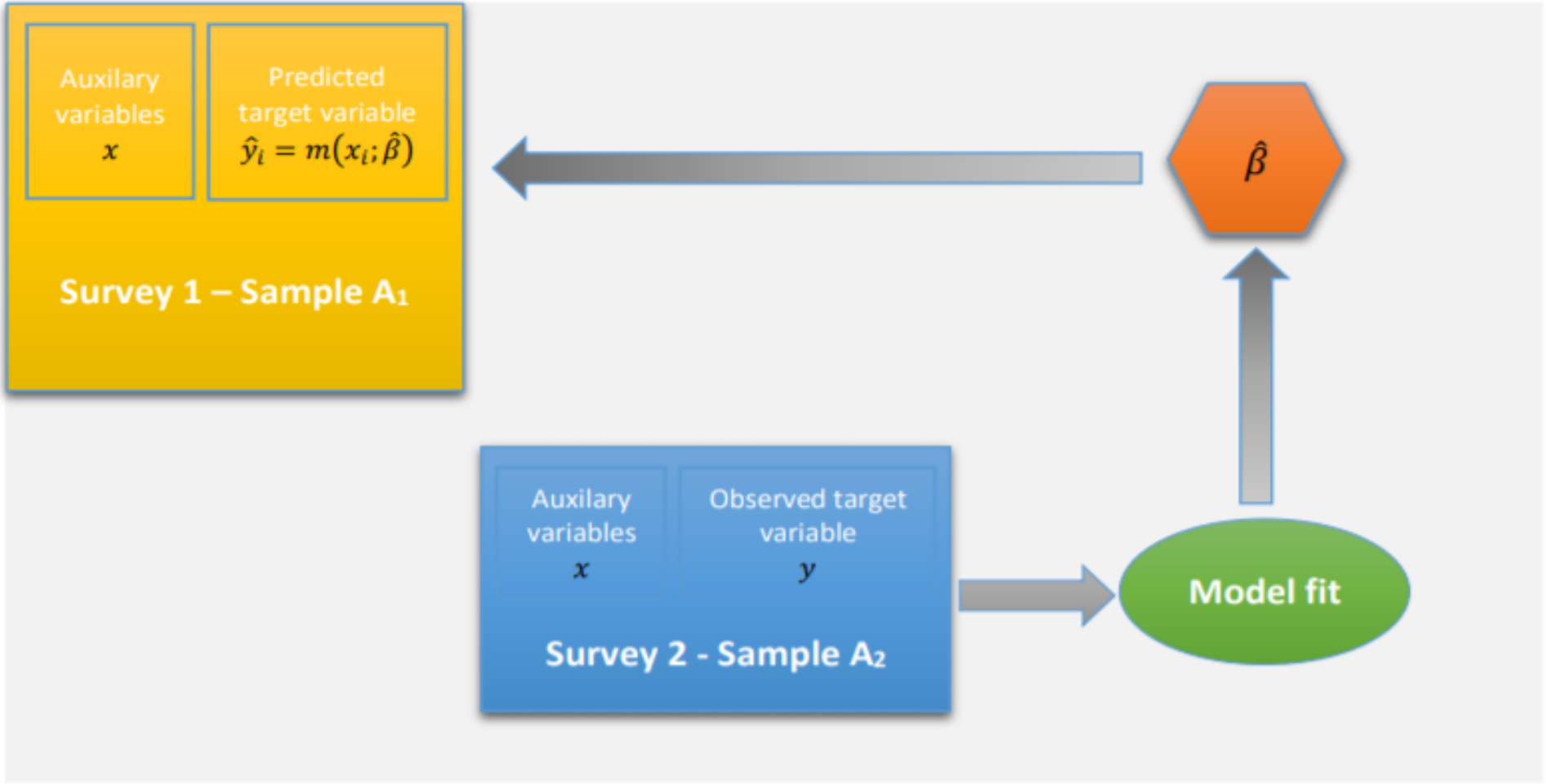
Projection Estimator: introduced by Kim and Rao as a model-assisted approach allowing to integrate data from two independent surveys

- **The first survey**, is characterized by a large sample A_1, but only collects auxiliary information or variables of general use (e.g. socio-economic variables);
- **The second survey** has a smaller sample A_2 but collects information on the target variable y, along with the same set of auxiliary variables available from A_1.

A special case of synthetic estimator: the projection estimator (2)

The total of variable y in the disaggregation domain i can be estimated as

$$\hat{Y}_{PR,i} = \sum_{j \in A_1} w_{j1} m(x_j;\, \hat{\beta}) \gamma_{ji}$$



Source: FAO, 2020.

Indirect estimators: Composite Estimator (CE)

Expressed as a **linear combination** of a direct estimator with a synthetic estimator:

$$\hat{Y}_{i,CE} = \phi_i \hat{Y}_{i,Dir} + (1 - \phi_i)\hat{Y}_{i,Sy}$$

Where:

- $\hat{Y}_{i,Dir}$ is the direct estimator for the $i-th$ small area
- $\hat{Y}_{i,Sy}$ is a synthetic estimator for the $i-th$ small area
- ϕ_i is a suitably chosen weight, with $0 \leq \phi_i \leq 1$

Objective: balancing the bias of the synthetic estimator with the instability (high variance) of the direct estimator. Possible examples are:

$\hat{Y}_{i,Dir}$: HT estimator
$\hat{Y}_{i,Sy}$: Linear regression estimator, other synthetic estimators

Indirect estimators: Composite (2)

Fundamental step for the implementation of composite estimators: **selection of ϕ_i** (also called shrinkage factor). Various approaches are available:

- **Approach 1:** the choice of ϕ_i is based on the minimization of $D^{-1}\sum_{i=1}^{D} MSE(\hat{Y}_{i,C})$. The optimal solution is given by

$$\phi^* = \frac{\sum_i MSE(\hat{Y}_{i,Sy})}{\sum_i (MSE(\hat{Y}_{i,Sy}) + V(\hat{Y}_{i,Dir}))}$$ which is estimated with $$\hat{\phi}^* = 1 - \frac{\sum_i \hat{V}(\hat{Y}_{i,Dir})}{\sum_i (\hat{Y}_{i,Sy} - \hat{Y}_{i,Dir})^2}$$

- **Approach 2:** the choice of ϕ_i depends on the domain sample size

$$\phi_i^* = \begin{cases} 1 & if\ \hat{N}_i \geq \delta N_i \\ \dfrac{\hat{N}_i}{(\delta N_i)} & otherwise \end{cases}$$

Where $\hat{N}_i$ is the sum of the design weights in the domain, N_i is the known population size in the domain, and $\delta$is subjectively chosen to control the contribution of the synthetic estimator.

Indirect estimators: Composite (5)

ADVANTAGES:

- Can be useful in domains in which a direct estimator has a **large variance**.
- Can be useful in surveys when analysed **domains vary** very much in terms of sample size.
- When the weights depend only on the sub-sample sizes, it is possible to estimates for **many** target variables at the same time.
- **Easy to implement** and not difficult to understand by the users.

LIMITATIONS:

- Problem of how to establish the **value of the weight**.
- Problem on how to provide **measures of error** for a given small area – for example, for bias.
- **Over-shrinkage problem** - estimators generally display less between-area variation then they should.
- Do not take into consideration **errors in auxiliary variables**. Adjustments to ensure coherence of estimates at different levels: $\hat{Y}_{d,adj} = \frac{\hat{Y}_d}{\sum_d \hat{Y}_d} \hat{Y}$

Quiz time!

Question 1 – SINGLE-SELECT

Synthetic estimators:

A. Allow assigning a broad area estimate to all the small areas included in the considered broad area
B. Are usually more precise than direct small area estimates
C. Can be affected by strong bias
D. Are based on the assumption that small areas within a given broad area are homogeneous
E. All of the above

Question 2 – SINGLE-SELECT

Composite estimators:

A. Allow balancing the bias of direct estimators and the instability of synthetic estimators
B. Are, for example, BARE estimators
C. Are linear combinations of direct and synthetic estimators
D. All of the above

Thank you!