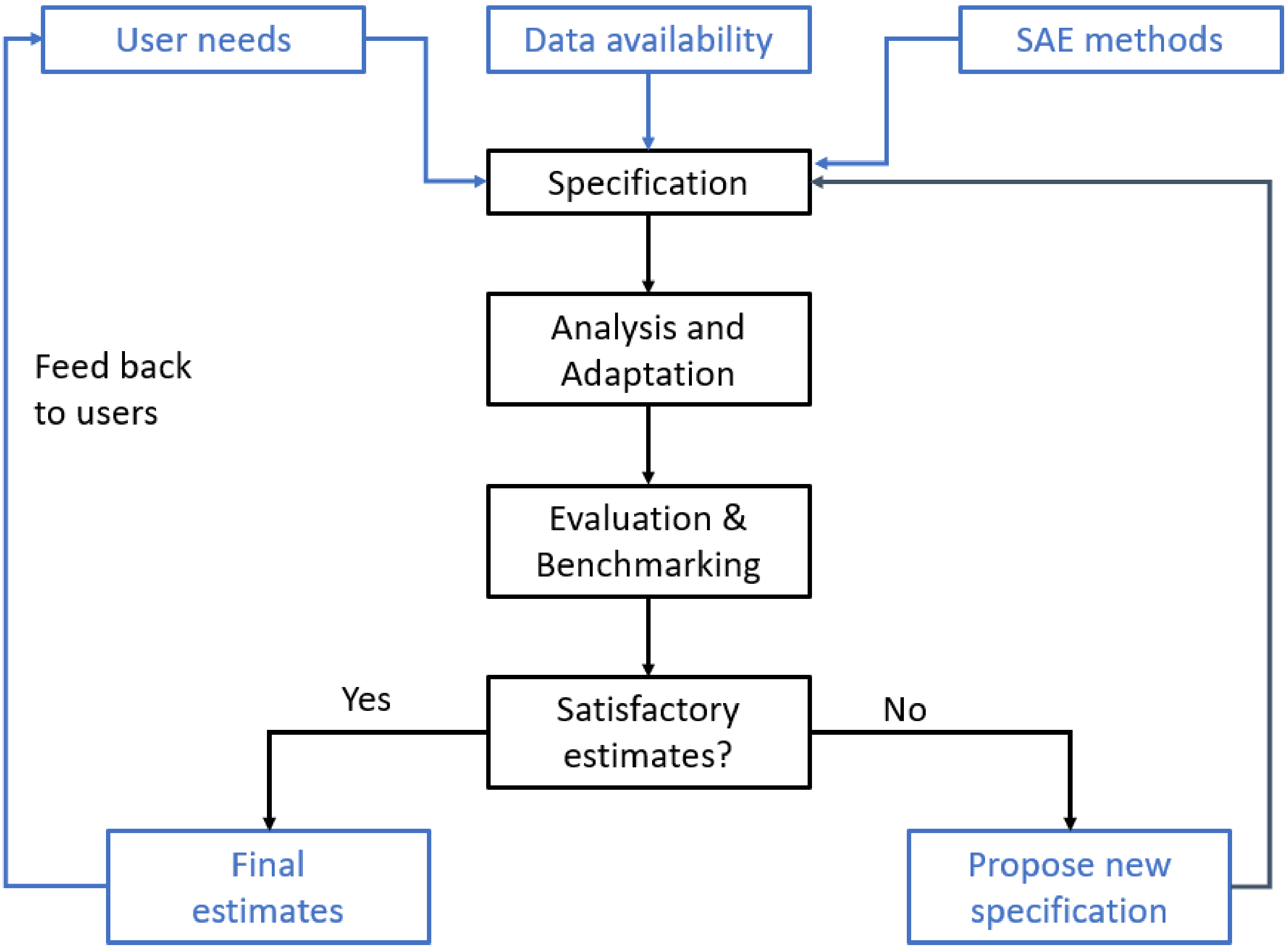
Small Area Estimation with Area Level Models

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*

Framework for SAE production



User needs

User needs can be identified by looking at the **purpose** of the estimation process:

- *What are the key policies or funding decisions?*
- *What are the questions that need to be answered?*

Important to identify **which indicator** can actually measure the information of interest:

- *What are you trying to measure?*
- *What type of indicator is the indicator of interest?*

Possible functional forms of indicators to be estimated:

Type	Example
Total	- Number of people moderately or severely food insecure according to the FIES. - Number of people with ownership or secure tenure rights over agricultural land
Mean/Average	Indicator 2.3.1 – Avrage labour productivity of small-scale food producers, Indicator 2.3.2 – Average income of small-scale food producers. Additional Other indicators – Average crop yield of smallscale food producers
Proportion	Indicator 2.1.2 – Prevalence of moderate or severe food insecurity in the population, based on the Food Insecurity Experience Scale
Rate	3.9.2 Mortality rate attributed to unsafe water.

Data availability

Having **good auxiliary information** is a crucial element for the implementation of model-based estimation. Some desirable characteristics are:

- *Good predictive power with respect to the indicator of interest;*
- *Availability for the considered estimation domain;*
- *Available for the same (or at least close) time period as the one to which the survey is referred.*

Possible sources of auxiliary data:

- Population and Agriculture Censuses;
- Administrative registers;
- Geospatial Information Systems;
- Other big data sources.

Problem with big data: one of the fundamental assumptions of basic SAE models is that auxiliary variables are measured without errors. In reality, variables retrieved from several big data sources are often affected by severe measurement errors and bias. Special SAE models allow to address this issue (only mentioned in this training)

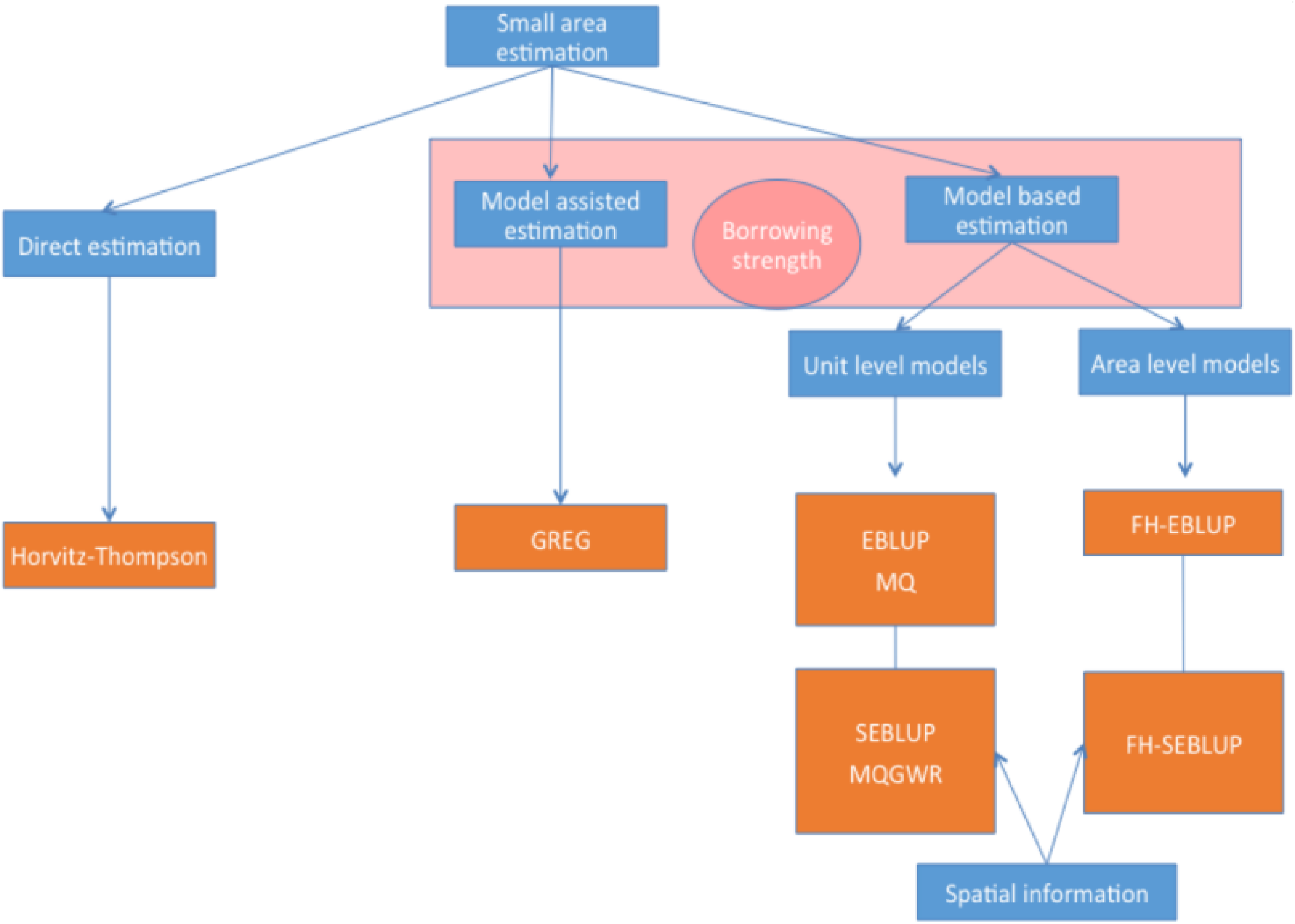
Data availability (2)

Auxiliary information can be available at **different levels of aggregation**:

- **Area-level data:** auxiliary data are aggregated at the level of the considered estimation domain (e.g. district, municipality, etc.) and not available at the unit level.

- **Unit-level data:** auxiliary data available for each unit in each domain. A fundamental requirement of Unit-level SAE approaches is that auxiliary variables share the same definition in the survey and additional considered data sources (e.g. census, admin register, etc.).

A schematic review of estimation approaches



Quiz time!

Question 1 – MULTI-SELECT

What are the main components of the specification phase of a SAE problem?

A. Data analysis and adaptation

B. Identification of user needs

C. Assessment of data availability

D. Estimates assessment and benchmarking

E. Selection of SAE method

Main advantages of model-based approaches

Adopting a model-based approach has the following main advantages:

- **Model diagnostics** can be used to find suitable models that fit the data well.

- **Area-specific measures of precision** can be associated with each small area estimate, solving the problem of instability that affect synthetic and composite estimators.

- Various types of models can be used: **linear mixed models** as well as **non-linear mixed models**

- Complex data structures, such as **spatial dependence** and **time series** structures, can also be handled

- Methodological developments for **random effects models** can be utilized to achieve accurate small area estimates.

AREA LEVEL SAE MODELS

Introduction to the Area-Level model

Area-level models:

- Popular models in research and official statistics due to their **easy application and interpretation**.
- Availability of area-level data is assumed: these models establish a relationship between an outcome (variable of interest) and a set of auxiliary variables available at the area level (e.g. district, municipality, etc.)

Relevant notation:

- $\boldsymbol{y_{ij}}$ = the value of the study variable (e.g. income, productivity, etc.) observed from survey data on unit j belonging to the $i-th$ small area.
- $\widehat{\boldsymbol{\theta}}_{i}^{dir}$ = design unbiased direct estimate for the parameter of interest $\boldsymbol{\theta_i}$ (e.g. total or average income, average productivity, etc.) in the i-th small area. Obtained using survey microdata and sampling weights
- $\boldsymbol{x_i^T}$ = vector of area-level auxiliary variables. The unit values of auxiliary variables can be aggregated at the area-level as totals, means, ratios, proportions, etc. Retrieved from a larger external data source (e.g. census, admin register, etc.)

Basic Area-level model: the Fay Herriot Model

Underlying Framework:

- 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 $\boldsymbol{N_i}$
- Sample data available on the target variable y
- Need to estimate D parameters of interest θ_i $(i = 1, \dots, D)$
- From a selected sample s we can obtain D direct estimates $\hat{\theta}_i^{dir}$ $(i = 1, \dots, D)$
- The sampling variance of direct estimates can be indicated with $\sigma_{e_i}^2$
- From a data source that is not affected by sampling errors, we can retrieve D p-vector of auxiliary variables aggregated at the area level, x_i^T $(i = 1, \dots, D)$

Basic Area-level model: the Fay Herriot Model (2)

The **basic area-level model**, also known as **FH model**, is given by the combination of a sampling model with an explicit linking model.

Sampling model:

$$\hat{\theta}_i^{dir} = \theta_i + e_i$$

Where $\hat{\theta}_i^{dir}$ is a design-unbiased direct estimator, and e_i are sampling errors with mean 0 and known variance $\sigma_{e_i}^2$. The basic FH model assumes the normality of sampling errors.

Linking model: Used to formalize the linear relationship between the population value θ_i and the domain level auxiliary information x_i.

$$\theta_i = x_i^T \beta + u_i$$

The assumption under the FH model is that the *u_i are iid following a normal distribution of mean 0 and variance σ_u^2*.

Basic Area-level model: the Fay Herriot Model (3)

The <u>combination of the sampling and the linking model</u> leads to a special **linear mixed model**

$$\hat{\theta}_i^{dir} = x_i^T\beta + u_i + e_i$$

Where β is a vector of regression parameters and u_i are random effects.

Under the above mentioned assumptions we have:

- $E_m\left[\hat{\theta}_i^{dir}\right] = E_m\left[x_i^T\beta + u_i + e_i\right] = x_i^T\beta$
- $MSE_m\left[\hat{\theta}_i^{dir}\right] = V_m\left[\hat{\theta}_i^{dir}\right] = E_m\left[\left(\hat{\theta}_i^{dir} - x_i^T\beta\right)^2\right] = E_m\left[\left(x_i^T\beta + u_i + e_i - x_i^T\beta\right)^2\right] = E_m\left[u_i^2 + e_i^2 + 2u_ie_i\right] = \sigma_u^2 + \sigma_{e_i}^2$

<u>*Under the assumption of normalty of u_i and e_i of the FH we have also*</u>

$$\hat{\theta}_i^{dir} \sim N(x_i\beta, \sigma_u^2 + \sigma_{e_i}^2)$$

Fay Herriot Model with Best Linear Unbiased Predictor

In this framework, the **Best Linear Unbiased Predictor (BLUP)** is obtained minimizing $MSE_m\left(\hat{\theta}_i^{dir}\right)$

- $\hat{\theta}_i^{dir} = x_i^T\beta + u_i + e_i = x_i^T a + b \longrightarrow \hat{\theta}_i^{dir}$ *is a linear estimator*
- $E_m\left[\hat{\theta}_i^{dir}\right] = x_i^T\beta = E_m[\theta_i] \longrightarrow \hat{\theta}_i^{dir}$ *is an unbiased estimator under the FH model*
- $Min_{\hat{\theta}_i^{dir}} MSE_m(\hat{\theta}_i^{dir}) \rightarrow$

$$\tilde{\theta}_i^{BLUP} = x_i^T\tilde{\beta} + \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{e_i}^2}\left(\hat{\theta}_i^{dir} - x_i^T\tilde{\beta}\right) = x_i^T\tilde{\beta} + u_i$$

Hence, $\tilde{\theta}_i^{BLUP}$ is the best linear unbiased predictor (<u>it's linear, it's unbiased, and it minimizes the MSE</u>)

Fay Herriot Model with Best Linear Unbiased Predictor (2)

The BLUP can be rewritten as follows:

$$\tilde{\theta}_i^{BLUP} = \gamma_i \hat{\theta}_i^{dir} + (1 - \gamma_i) x_i^T \tilde{\beta}$$

Which is a linear combination between a direct estimator and a synthetic estimator

- $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{e_i}^2}$
- σ_u^2 is unknown
- $\sigma_{e_i}^2$ is assumed to be known. However, in practical applications it usually needs to be estimated through the estimated variance of direct estimates
- $\tilde{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \hat{\theta}^{dir}$
- $\tilde{\theta}_i^{BLUP}$ can be considered as a special case of **COMPOSITE ESTIMATOR**

Fay Herriot Model with Best Linear Unbiased Predictor (3)

- Using the joint distribution $f\left(\hat{\theta}_i^{dir}, u_i\right)$ and under the normality assumption we can get the Restricted Maximum Likelihood (REML) estimates of σ_u^2, let's say $\hat{\sigma}_u^2$

- Plugging in $\hat{\beta}$ and $\hat{\sigma}_u^2$ we obtain the **Empirical BLUP (EBLUP)**:

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i \hat{\theta}_i^{dir} + (1 - \hat{\gamma}_i) x_i^T \hat{\beta}$$

- $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}$ is called shrinkage factor

MSE under the Fay Harriot Model

The uncertainty of the FH Empirical BLUP has been studied by many researchers. A review of possible approaches and formulas is provided in «Small Area Estimation» (2015) from Rao JNK and Molina I.

A correct estimation of the MSE of EBLUP is:

$$mse\left(\hat{\theta}_i^{EBLUP}\right) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2)$$

Where:

- $g_{1i}(\hat{\sigma}_u^2) = \hat{\gamma}_i * \sigma_{e_i}^2$ is the **leading term of the mse**
- $g_{2i}(\hat{\sigma}_u^2) = (1 - \hat{\gamma}_i)^2 x_i \left(\frac{\sum_{i=1}^{D} x_i x_i^T}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}\right)^{-1} * x_i$
- $g_{3i}(\hat{\sigma}_u^2) = \frac{\sigma_{e_i}^4}{\left(\hat{\sigma}_u^2 + \sigma_{e_i}^2\right)^3} * 2\left[\sum_{i=1}^{D} \frac{1}{\left(\hat{\sigma}_u^2 + \sigma_{e_i}^2\right)^2}\right]^{-1}$

Fay Harriot Model in out-of-sample areas

- Population divided into m **small areas**
- A sample is available in $m - k$ areas, meaning that **for k areas no sampling observation is available**
- The k areas are called **out-of-sample areas**
- For out-of-sample areas, the EBLUP estimator under the FH model corresponds to a synthetic estimator:

$$\hat{\theta}_c^{OUT} = x_c^T \hat{\beta} \quad c = 1, \dots, k$$

In order to compute the synthetic component of the estimator, auxiliary information needs to be available also for out-of-sample areas.

Fay Herriot model: recap

- The FH model has fewer data requirements compared to unit-level approaches
- In many applications, this area-level method can reduce the MSE of direct estimates
- Area-level models are among the simplest SAE approaches
- For out of sample areas, the method provides only model based synthetic estimates

$$\hat{\theta}_i^{EBLUP} = \boxed{\hat{\gamma}_i}\hat{\theta}_i^{dir} + (1 - \boxed{\hat{\gamma}_i})x_i^T\hat{\beta}$$

$$\boxed{\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}}$$

Linear combination of a direct estimate (based on sampling weights) and a synthetic estimator, where the two are balanced with the "shrinkage factor"

Fay Herriot model: recap (2)

Key question: is the **FH EBLUP** more accurate than traditional direct estimators (e.g. the HT estimator)? Yes it is, but…

- *Gains in terms of reduced variability are not certain. They depend on*
  - *The predictive power of selected auxiliary variables*
  - *The between area variability* $\hat{\sigma}_u^2$
  - *The initial variability of direct estimates* $\hat{\sigma}_{e_i}^2$

If auxiliary variables are not good predictors of the target variable, the EBLUP estimator tends to be similar to the direct estimator (large value of $\hat{\gamma}_i$) and the reduction of variability is negligible.

If $(\hat{\theta}_i^{dir} - x_i^T \hat{\beta}) \uparrow \Rightarrow \sigma_u^2 \uparrow \Rightarrow \hat{\gamma}_i \uparrow \Rightarrow mse(\hat{\theta}_i^{EBLUP}) \approx \hat{\gamma}_i * \sigma_{e_i}^2 \uparrow$

Fay Herriot model: recap (3)

Data needed:

- *Dataset containing direct estimates of the parameter of interest calculated from sample data for each small area*
- *Dataset containing a set of auxiliary variables aggregated at the small area level*
- *Sampling variance of the direct estimator (this has to be known or estimated before implementing the area level model)*

Output dataset:

- *Dataset with predicted parameters for each small area with quality indicators such as:*
  - *The estimated MSE for each small area*
  - *Diagnostics of model bias*
  - *Benchmarking*
  - *Model selection diagnostic (AIC, BIC,…)*

Fay Herriot Model: recap (4)

- **ADVANTAGES:**
  - Can be applied for estimation when few or even no sample data are available for one or more domains of interest.
  - Can be applied on macrodata available at the domain level.
  - Useful to improve direct estimators if a set of covariates with a strong relationship with the variable of interest is available.
  - The variances of the small area direct estimates has to be known. *[Usually a smoothed model for variance estimation is applied and variances are assumed to be known. This affects the MSE.]*
  - Covariates are needed only at domain level.

- **DISADVANTAGES:**
  - If the model is not correctly specified the estimator can be biased.
  - When adding up small domains estimates to a larger domain, the obtained aggregate could differ from direct broad area estimates. *[A simple way to ensure consistency is to suitably adjust the EBLUP area level estimator.]*
  - Symmetry of the distribution is required. However, this assumption is not always fulfilled
  - Benchmarking can be also set as a constraint to obtain small area estimates.
  - Assumptions of normality with known variance might be untenable at small sample sizes.
  - Model variance can be estimated to be zero (undesirable result).

Fay Herriot Model: recap (5)

Properties	Advantages	Disadvantages	Extensions
Model assumptions	Efficiency under the assumption of Normality of LMM	Linearity of the relation with fixed effects aux variables Lack of correlation between the random area effects	Non parametric extension NEBLUP (Giusti et al 2012) Spatial extension SEBLUP (Petrucci and Salvati, 2006; Pratesi and Salvati, 2008)
Design consistency	Design consistent		
Robustness to outliers		Not robust against outliers	
Out-of-sample predictions	Prediction available	Prediction not inclusive of spatial information	Spatial extension SEBLUP (Petrucci and Salvati, 2006; Pratesi and Salvati, 2008, 2009)

Quiz time!

Question 1 – MULTI-SELECT

The Fay-Herriot area-level model

A. Is the most common SAE area-level model

B. Assumes normality of random effects only

C. Assumes normality of both the error terms and the random effects

D. Is a linear mixed model combining two sampling models

E. Is a linear mixed model combining a sampling model with an explicit linking model

Question 2 – MULTI-SELECT

The EBLUP resulting from an area-level model is:

A. Linear combination of a direct estimator and a synthetic estimator

B. Can be seen as a model-based composite estimator

C. Is design-unbiased

D. The Empirical Broad Linear Unbiased Predictor

E. The Empirical Best Linear Unbalanced Predictor

F. The Empirical Best Linear Unbiased Predictor

Question 3 – SINGLE-SELECT

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

A. Yes

B. No

Question 4 – SINGLE-SELECT

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

A. 1

B. 10

C. Indefinite number

D. Maximum D-2, where D is the number of small areas

Question 5 – SINGLE-SELECT

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

A. Yes

B. No

Thank you!