Small Area Estimation

1 - Small area estimation problem

2 - Estimation for domains - **Direct estimators** – estimation for planned domains

3 – Coefficient of Variation and Minimum level of precision

4- Estimation for unplanned domains and/or where the sample size is not enough for the minimum level of precision – **Indirect estimators**

Recap

- Target parameters in the domain:

- Total of the study variable
- Mean of the study variable
- At risk of poverty rate
- Poverty gap

$$t = \sum_{k \in U} y_k$$

$$\bar{y} = \sum_{k \in U} y_k / N$$

$$P_0 = \frac{1}{N} \sum_{i=1}^{N} I(y_i < z).$$

$$P_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{G_i}{z}.$$

When direct estimates are not reliable?

From Statistics Canada (2010). Guide to the Labour Force Survey SAE methods:

Statistics Canada applies the following guidelines on LFS data reliability:
-if CV ≤ 16.5% then direct estimates are disseminated without restrictions (no release)
-if 16.5% < CV ≤ 33% then the estimates should be accompanied by warning (release with caveat)
-if CV > 33% then the estimates are not recommended for use (no release)

These threshold may be good for labour market variables.

What if we are dealing with estimates with CV > 33% ?

Not direct Domain Estimates – Indirect estimates

If a domain sample cannot support direct estimates of adequate precision, often indirect estimate is used to “borrow strength” by using values from related areas and/or time periods to increase effective sample size.

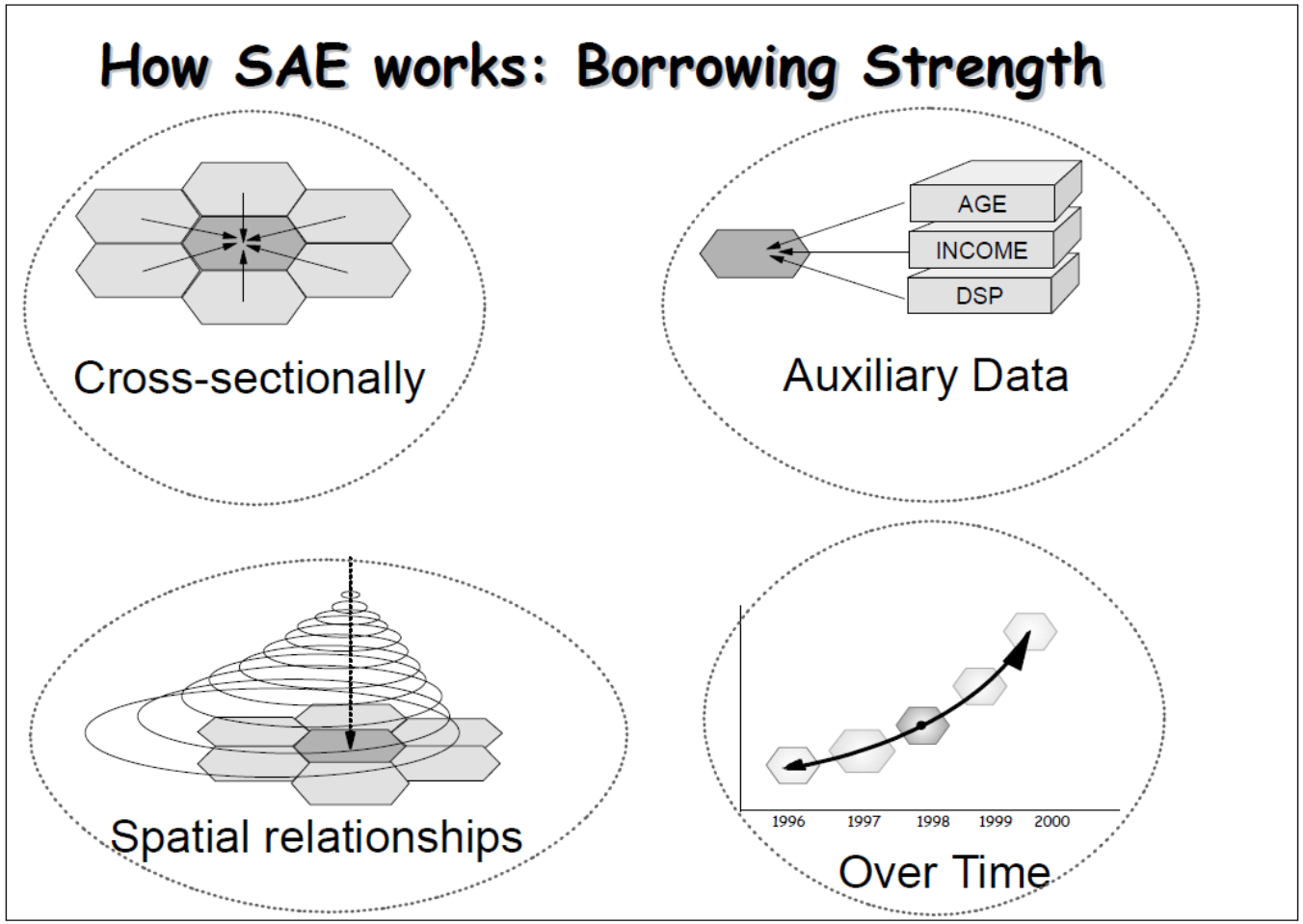
These values are used in the estimation through a model (implicitly or explicitly) that provides a link to related areas and/or time periods through supplementary information such as recent census counts or current administrative records related to the variable of interest.

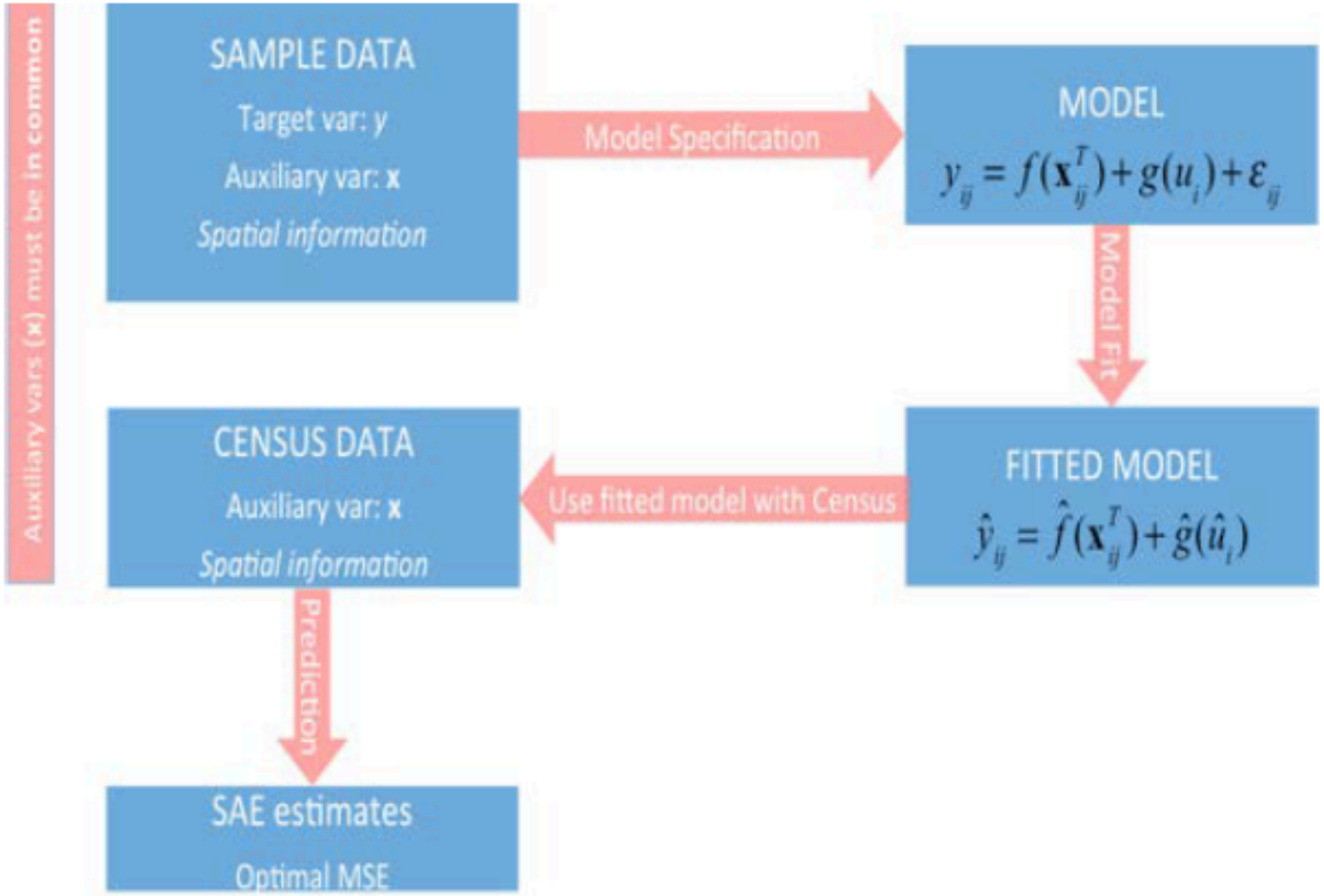
Indirect Domain Estimates

It is possible to divide the indirect estimates into:

— **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.

SAE: Borrowing Strength from?




Auxiliary vars (x) must be in common
SAMPLE DATA
Target var: y
Auxiliary var: x
Spatial information
Model Specification
MODEL
$y_{ij} = f(\mathbf{x}_{ij}^T) + g(u_i) + \varepsilon_{ij}$
Model Fit
FITTED MODEL
$\hat{y}_{ij} = \hat{f}(\mathbf{x}_{ij}^T) + \hat{g}(\hat{u}_i)$
Use fitted model with Census
CENSUS DATA
Auxiliary var: x
Spatial information
Prediction
SAE estimates
Optimal MSE

Indirect Domain Estimates

We can distinguish the indirect methods into:

— **Synthetic Estimator:** an estimator is called a synthetic estimator if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area.
As a consequence, the variance of a synthetic estimator is lower than the corresponding direct estimator but is biased if the model assumptions are not satisfied.

— **Composite Estimator:** it is a linear combination between a direct estimator and a synthetic estimator. For this reason it represents a good compromise in terms of efficiency between the characteristics of the two components.

Synthetic estimator

- Synthetic assumption: small areas have same characteristic as the large area (e.g. unemployment rates for different demographic groups for the Pisa Province is the same as that for Tuscany)

- Advantages of synthetic estimator:
  - Simple and intuitive
  - Applies to general sampling designs
  - Borrow strength from similar
  - Provides estimates for areas with no sample from the sample survey

Synthetic estimator

Broad Area Ratio Estimator (BARE)

this is one of the simplest types of small area models.

It is calculated by applying the rate for a broad area obtained from a survey (eg disability rate or unemployment rate or poverty rate) to the small area populations (available from say a population census or demographic estimates).

Synthetic estimator

Broad Area Ratio Estimator (BARE) with auxiliary data

this uses information that is correlated with the variable of interest and is available at the small area level to derive an estimate that adjusts for compositional differences in small areas.

Ratio estimator, Regression estimator

Composite estimator

A composite estimator is an estimator that combine direct and synthetic estimator:

$$\hat{Y}_{i,C} = \phi_i \hat{Y}_{i,D} + (1 - \phi_i)\hat{Y}_{i,S}$$

where

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

The aim of the composite estimator is to balance the potential bias of the synthetic estimator against the instability of the design-based estimator

$\hat{Y}_{i,D}$: HT estimator or simple ratio estimator

$\hat{Y}_{i,S}$: regression estimator

The choice of the weight ϕ_i

Optimal ϕ_i

a. Minimize the $MSE(\hat{Y}_{i,C})$ with respect to ϕ_i assuming $COR(\hat{Y}_{i,D}, \hat{Y}_{i,S}) \approx 0$

- the optimal solution is given by

$$\phi_i^* = \frac{MSE(\hat{Y}_{i,S})}{MSE(\hat{Y}_{i,S}) + V(\hat{Y}_{i,D})}$$

- the parameter ϕ_i can be estimated by

$$\hat{\phi}_i^* = \frac{\widehat{MSE}(\hat{Y}_{i,S})}{(\hat{Y}_{i,S} - \hat{Y}_{i,D})^2} = 1 - \frac{\hat{V}(\hat{Y}_{i,D})}{(\hat{Y}_{i,S} - \hat{Y}_{i,D})^2}$$

Note: very unstable $\hat{\phi}_i^*$

The choice of the weight ϕ_i

b. Minimize $m^{-1}\sum_{i=1}^{m} MSE(\hat{Y}_{i,C})$ with respect to a common weight $\phi_i = \phi$

- the optimal solution is given by

$$\phi^* = \frac{\sum_{i=1}^{m} MSE(\hat{Y}_{i,S})}{\sum_{i=1}^{m}(MSE(\hat{Y}_{i,S}) + V(\hat{Y}_{i,D}))}$$

- the parameter ϕ can be estimated by

$$\hat{\phi}^* = 1 - \frac{\sum_{i=1}^{m} \hat{v}(\hat{Y}_{i,D})}{\sum_{i=1}^{m}(\hat{Y}_{i,S} - \hat{Y}_{i,D})^2}$$

Comparison Between Direct, Synthetic and Composite Estimator

Empirical comparison of small area estimation methods for the Italian Labor Force Survey (LFS)

- Performance of small area estimators are studied by simulating sample selection from 1981 Population Census.

- Samples are 400 sample replicates (h), each of identical size of the LFS sample drawn following the LFS design (two stages with stratication)
- Design-based properties of the estimators: verification of its Bias and calculation of the CV of the estimator

Tzavidis, Salvati, Marchetti (MOOC 2020)

Comparison Between Direct, Synthetic and Composite Estimator

Empirical comparison of small area estimation methods for the Italian Labor Force Survey (LFS)

- **Active population, aged 15-64 - annual averages -** According to the definitions of the International Labour Organisation (ILO) for the purposes of the labour market statistics people are classified as employed, unemployed and economically inactive. The economically active population is the sum of employed and unemployed persons. Inactive persons are those who, during the reference week, were neither employed nor unemployed.

- Example - 14 Health Service Areas (HSA) of the Friuli Venezia Giulia Region are considered to be small areas
- Y_i – true mean of the *y* variable in the domain *i* (HSA) – annual average of active population (known from the Census)

Comparison Between Direct, Synthetic and Composite Estimator

Index used to evaluate the performances of the estimators

- Average Relative Bias

$$ARB = \frac{1}{14}\sum_{i=1}^{14}\left|\frac{1}{400}\sum_{h=1}^{400}\frac{\hat{Y}_i^{(h)} - Y_i}{Y_i}100\right|$$

- Relative Root MSE

$$RRMSE = \frac{1}{14}\sum_{i=1}^{14}\left(\frac{\sqrt{\frac{1}{400}\sum_{h=1}^{400}(\hat{Y}_i^{(h)} - Y_i)^2}}{Y_i}100\right)$$

Comparison Between Direct, Synthetic and Composite Estimator

ARB and RRMSE for Direct, Synthetic and Composite estimators

Table: Estimators performances

Estimator	ARB	RRMSE
Direct	2.39	31.08
Synthetic	8.97	23.80
Composite	6.00	23.57

Note: the RRMSE of Direct estimator is approximatively 30% higher than Synthetic and Composite estimator

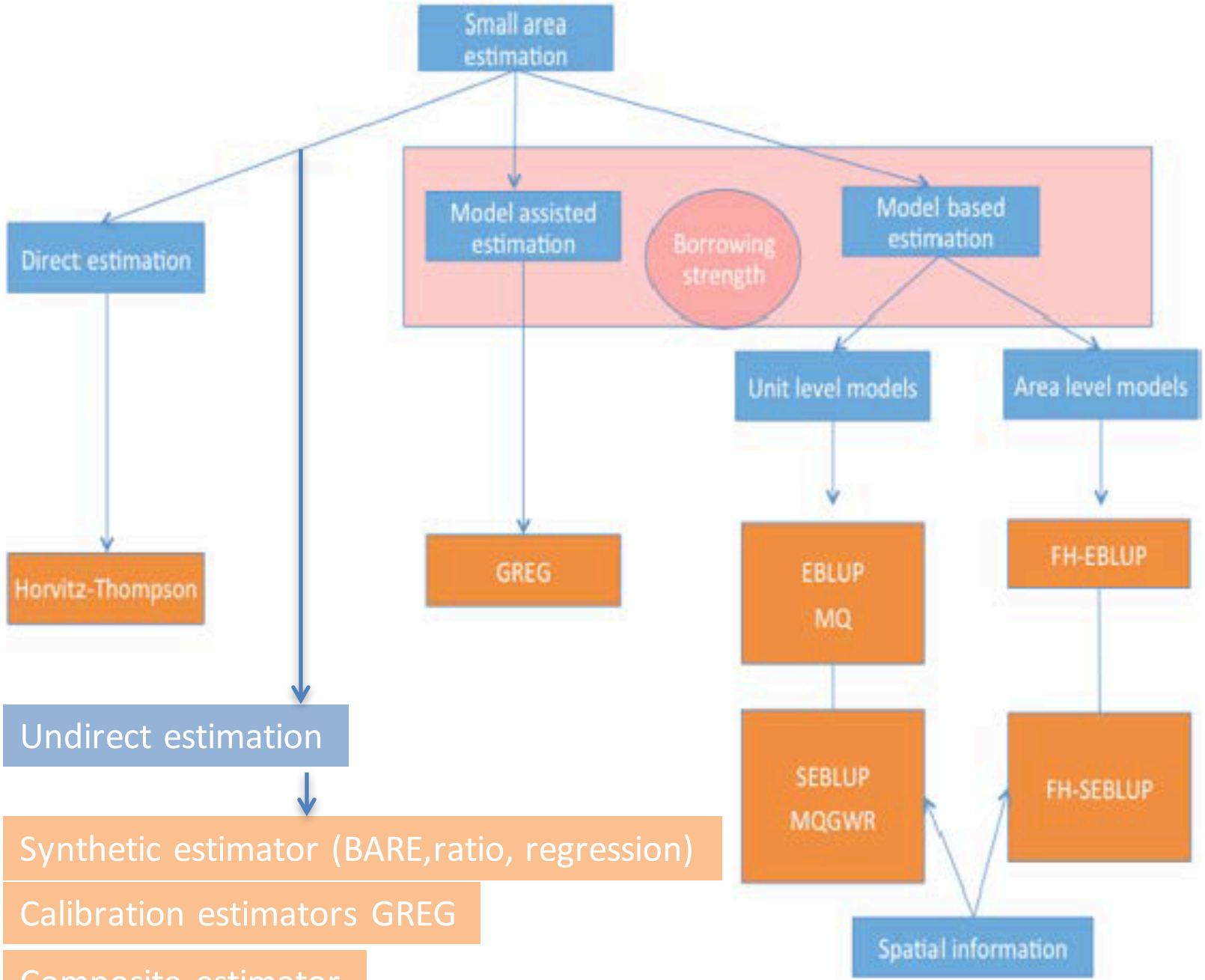
Inference framework

Further we have, depending on the "reference framework" for inference:

— **Design Based Approach:** Estimator properties are assessed with respect to the sampling design (see previous example). This framework is used for small area estimation, mainly because of its simplicity.

— **Model Assisted Approach:** In practice, the values of Y are typically defined by assuming a model for the distribution of Y given X. That is, practitioners have been willing to use models in order to identify optimal strategies for estimating T_Y. However, their assessment of these strategies remain design-based (Särndal, Swensson and Wretman, 1992).

— **Model Based Approach:** design-unbiasedness is no longer a requirement, the alternative property we require of the estimator under this approach is that it be model-unbiased $E\left(\hat{T}_Y - T_Y \middle| \mathbf{S}, \mathbf{X}\right) = 0$. given the sample S and aux info X.


Small area estimation
Direct estimation
Model assisted estimation
Borrowing strength
Model based estimation
Unit level models
Area level models
Horvitz-Thompson
GREG
EBLUP
MQ
FH-EBLUP
Undirect estimation
SEBLUP
MQGWR
FH-SEBLUP
Spatial information
Synthetic estimator (BARE,ratio, regression)
Calibration estimators GREG
Composite estimator

Model Based Approach

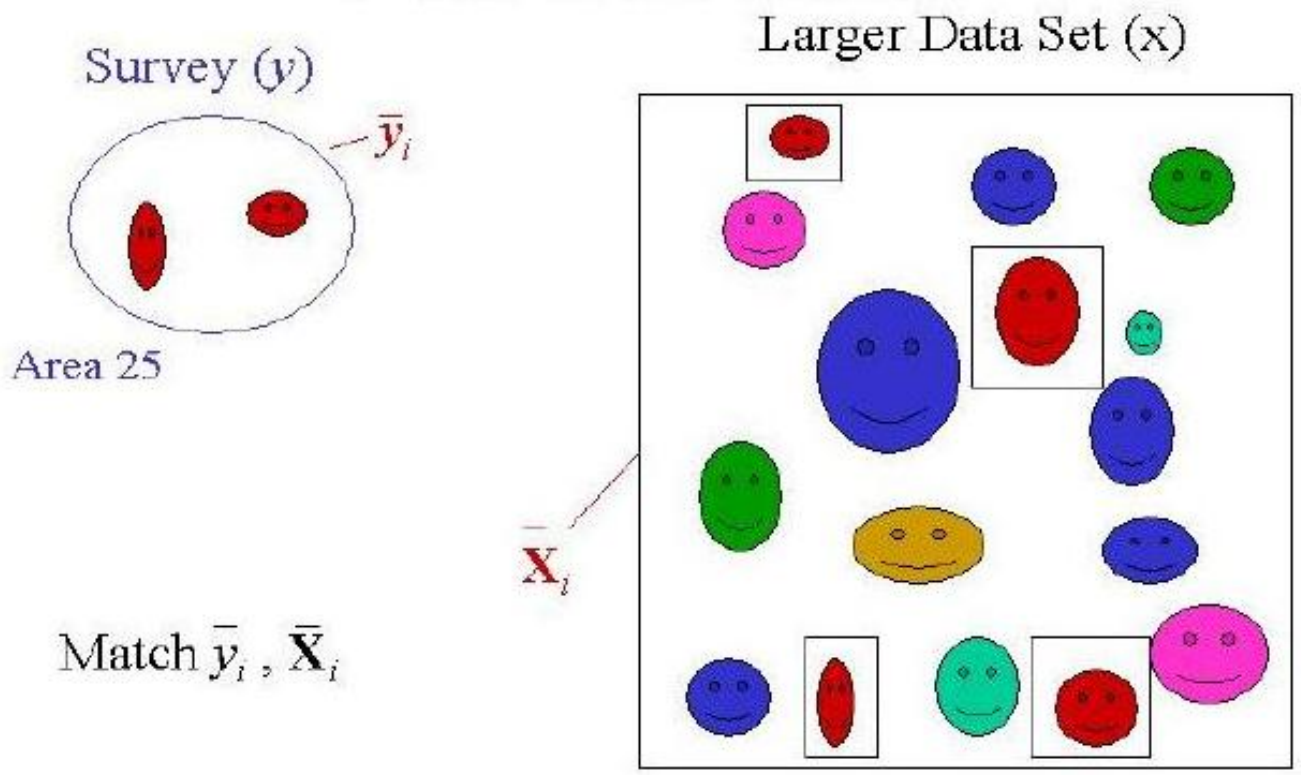
Two types of models depending on the data: area level model and unit level model.

— **Area Level Model:** area level models are appropriate if only area-level summary data available for the auxiliary and/or the response variables. It is possible to take into account the sampling weights into model.

— **Unit Level Model:** unit level models are generally to prefer if unit-specific information is available. They usually ignore the survey weights.

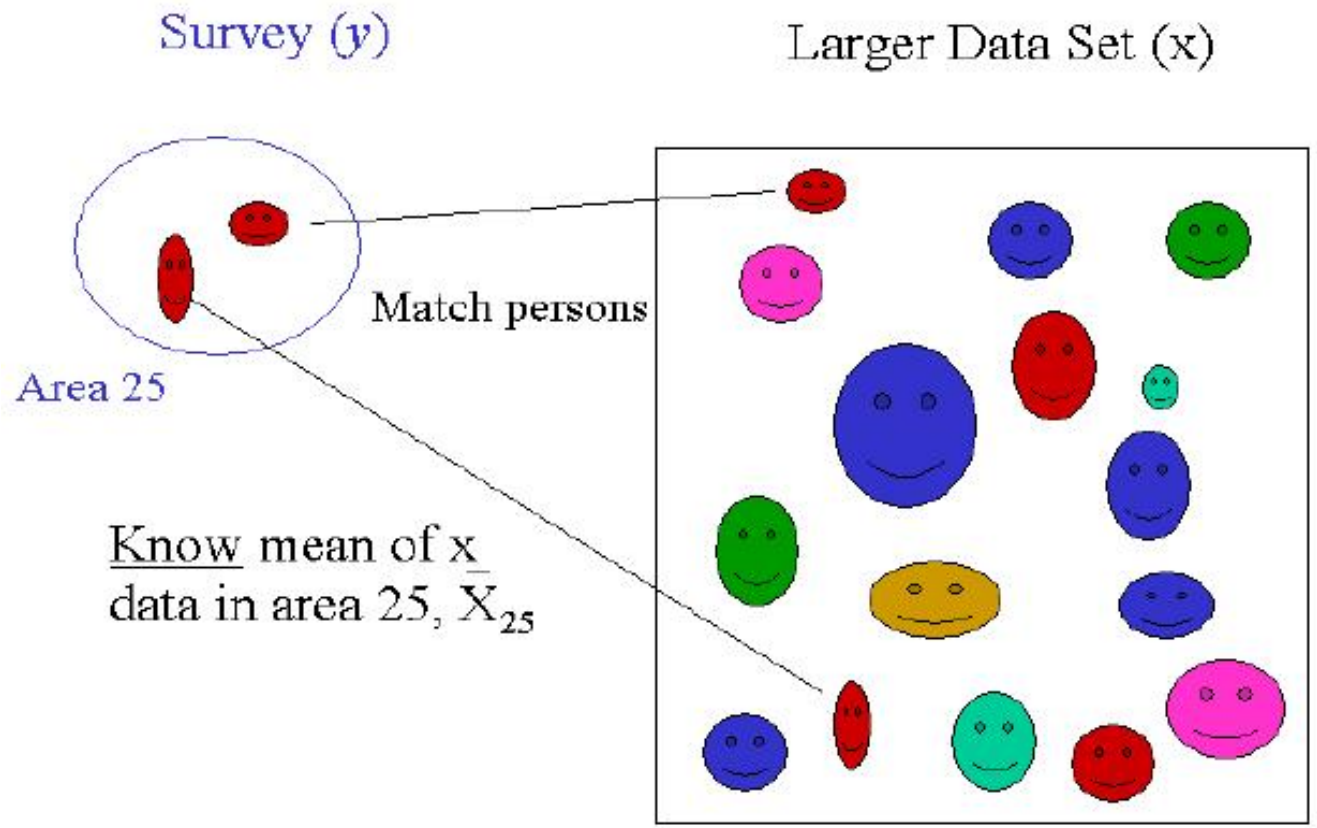
Auxiliary information in Area Level Approach

When only summary data (such as the traditional survey estimator) for the response variable is available at the small area level:



Auxiliary information in Unit Level Approach

When data for both the response variable and auxiliary variables are available at the unit level:



Small area estimation techniques

ABS suggestion 1

"The choice of small area method depends on the availability of auxiliary data and the relationship between these data and the variables of interest at the small area level. In essence, we are looking to "borrow strength" from these auxiliary data to increase the accuracy of the estimates. Small area models range from the simple to the more complex, the latter requiring considerably more time, effort, technical skill and available data. A range of quantitative and qualitative diagnostics should be used to choose the best model for the given data."

- Examples of application of the Synthetic and composite estimator during the R lab