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

$$\boxed{t = \sum\nolimits_{k \in U} y_k}$$

- Mean of the study variable

$$\boxed{\bar{y} = \sum\nolimits_{k \in U} y_k \,/\, N}$$

- At risk of poverty rate

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

- Poverty gap

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

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 $E\left(\hat{T}_Y - T_Y \,|\, \mathbf{S},\mathbf{X}\right) = 0$. given the sample S and aux info X.

What are we modelling?

We are modelling the relationship between an outcome and the auxiliary variables

note that:

- unplanned domains=geographical domains= areas

- notation:

Y_{ij} outcome= the value of the study variable (*income survey data unit j, individual or household, in area i*)

$\hat{\theta}_i^{dir}$ outcome= survey direct estimator (*per capita income in area i, total income in area i*)

What are we modelling?

The models are classied into two broad types:

1 Aggregate level (or area-level) models that relate the small area outcome (means, totals) to area-specific auxiliary variables. Such models are essential if unit level data are not available

2 Unit level models that relate the outcome (unit values of the study variable) to unit-specific auxiliary variables

The use of *explicit models* offers several advantages

What are we modelling?

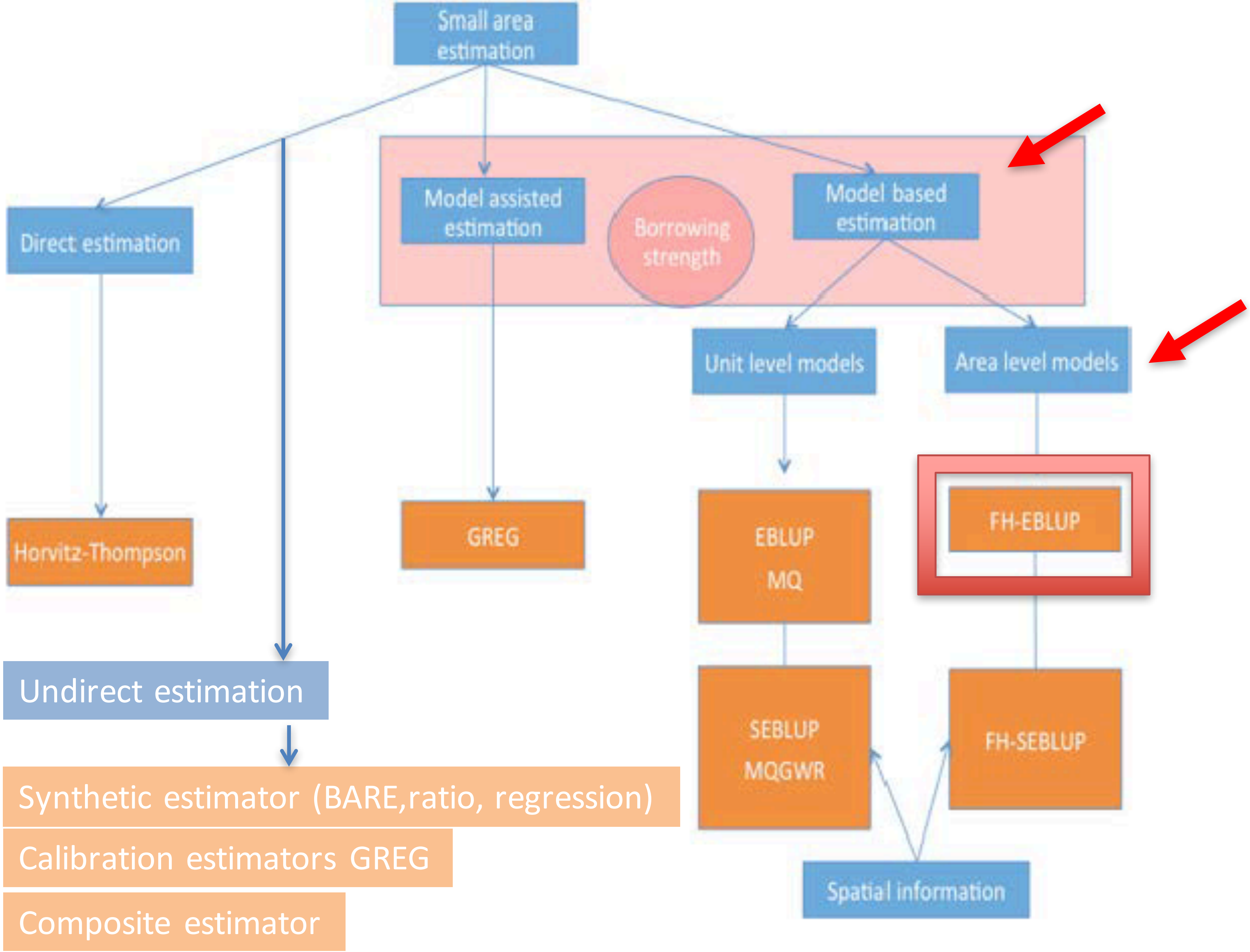
Advantages both for are-level and unit-level models:

1 Model diagnostics can be used to find suitable models that fit the data well

2 Area-specic measures of precision can be associated with each small area estimate, solving the problem of instability seen for synthetic and composite estimators

3 Linear mixed models as well as nonlinear models can be used.
4 Complex data structures, such as spatial dependence and time series structures, can also be handled

5 Methodological developments for random effects models can be utilized to achieve accurate small area inferences



Undirect estimation

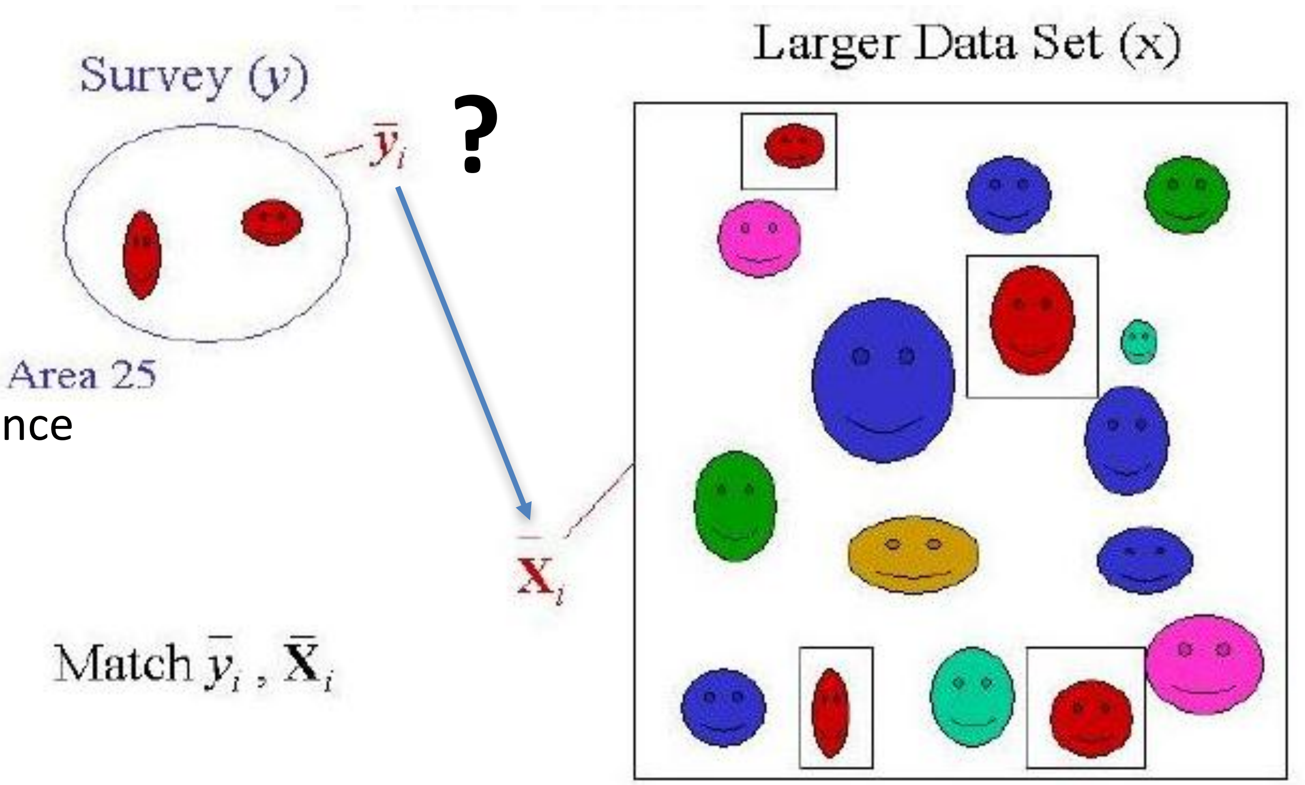
Synthetic estimator (BARE,ratio, regression)

Calibration estimators GREG

Composite estimator

Area Level Approach

1 - outcome $\hat{\theta}_i^{dir}$, auxiliary variable *X* available at unit level (j) synthetized at area level from a larger ta set
Ex: unit=household, area=province: Po direct estimates at province-level, X household size at hhs level

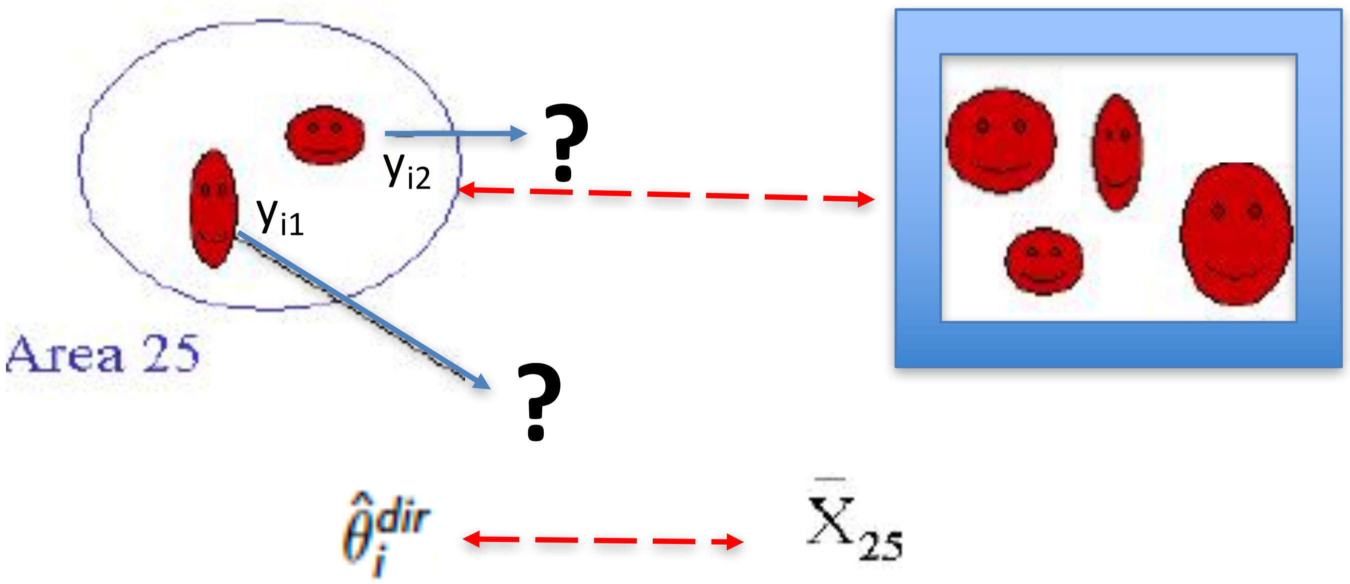


Pisa province

Area Level Approach

2 – outcome y_{ij} , only a synthesis of auxiliary variable X is available a larger data set

Ex: unit=household, area=province: y_{ij} income available at hhs level in area *i*, X population density available at area level



3 – outcome y_{ij} , X_{ji} is available …but you have no access to linkage codes of the microdata (Fiscal code – privacy issues) in the X data set
Ex: unit=household, area=province: y_{ij} income available at hhs level in area *i*, X declared income of the hhs

Fay Herriot Model

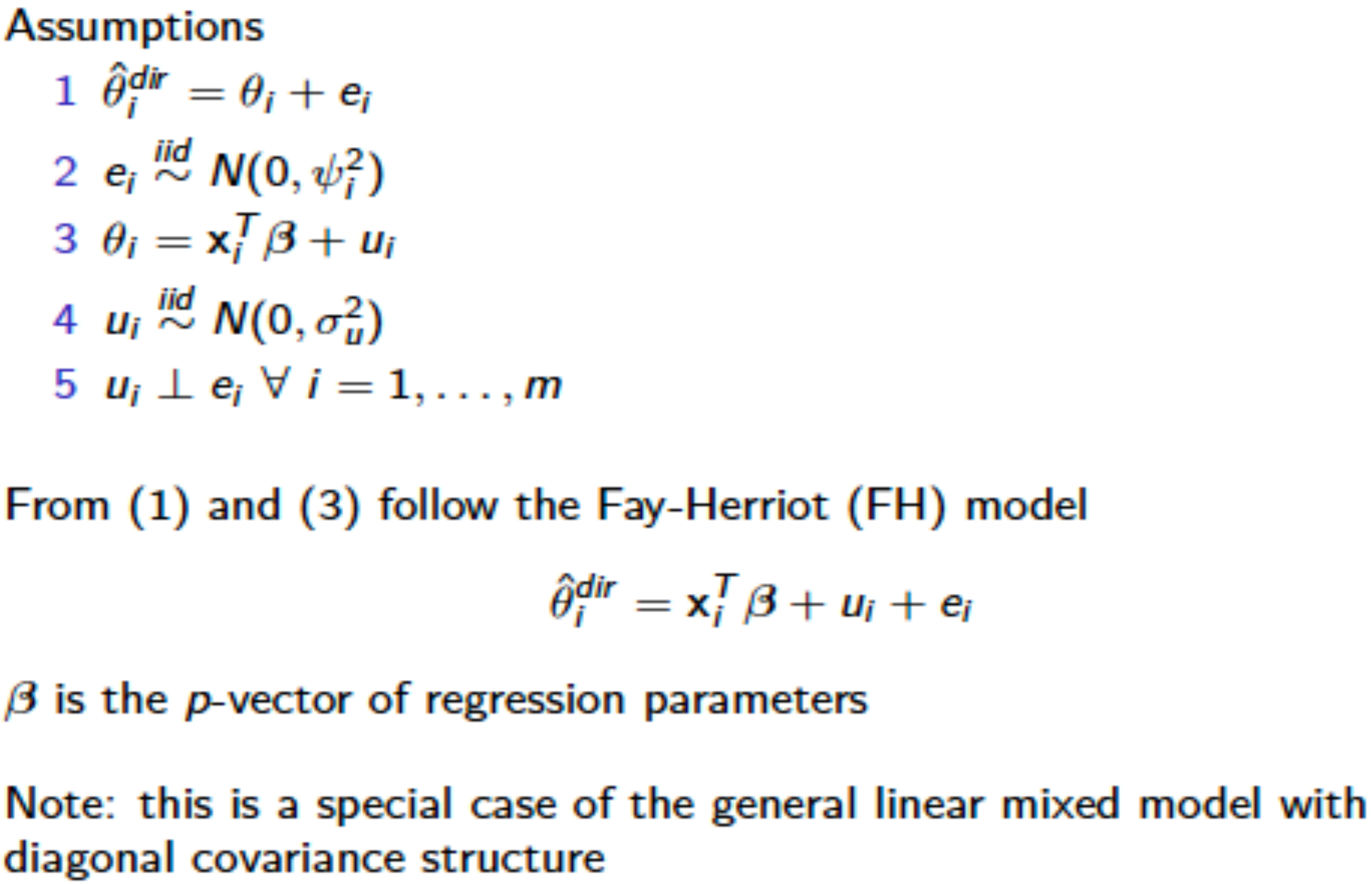
Framework

- Population Ω divided into m (small) areas
- Availability of sample data on target variable, y
- m parameter of interest (e.g. mean), θ_i, $i = 1, \ldots, m$
- From sample → m direct estimates, $\hat{\theta}_i^{dir}$, $i = 1, \ldots, m$
- and m *MSE* estimates, $mse(\hat{\theta}_i^{dir}) \approx \psi_i^2$, $i = 1, \ldots, m$
    - Here ψ_i^2 is the $MSE(\hat{\theta}_i^{dir})$ and is often considered known
    - In real application we know only its estimates $mse(\hat{\theta}_i^{dir})$
- From other data sources → m p-vector of auxiliary variables, $\mathbf{x}_i$ $\forall$ $i = 1, \ldots, m$

Fay Herriot Model

FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data,

Journal of the American Statistical Association, 74, 269–277.



Assumptions

1 $\hat{\theta}_i^{dir} = \theta_i + e_i$

2 $e_i \overset{iid}{\sim} N(0, \psi_i^2)$

3 $\theta_i = \mathbf{x}_i^T \beta + u_i$

4 $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$

5 $u_i \perp e_i \ \forall \ i = 1, \ldots, m$

From (1) and (3) follow the Fay-Herriot (FH) model

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

β is the p-vector of regression parameters

Note: this is a special case of the general linear mixed model with diagonal covariance structure

Fay Herriot Model

- Under above mentioned assumptions

$$\begin{aligned} E_m[\hat{\theta}_i^{dir}] &= E_m[\mathbf{x}_i^T\beta + u_i + e_i] = \mathbf{x}_i^T\beta \\ MSE_m(\hat{\theta}_i^{dir}) &= V_m(\hat{\theta}_i^{dir}) = E_m[(\hat{\theta}_i^{dir} - \mathbf{x}_i^T\beta)^2] \\ &= E_m[(\mathbf{x}_i^T\beta + u_i + e_i - \mathbf{x}_i^T\beta)^2] \\ &= E_m[u_i^2 + e_i^2 + 2u_ie_i] = \sigma_u^2 + \psi_i^2 \end{aligned}$$

- Under Normality of u_is and e_is and under FH model

$$\hat{\theta}_i^{dir} \sim N(\mathbf{x}_i\beta, \sigma_u^2 + \psi_i^2)$$

Fay Herriot Model

The Best Linear Unbiased Predictor (BLUP) is obtained minimizing $MSE_m(\hat{\theta}_i^{dir})$

- $\hat{\theta}_i^{dir} = \mathbf{x}_i^T \beta + u_i + e_i = \mathbf{x}_i^T \mathbf{a} + \mathbf{b}$, $\hat{\theta}_i^{dir}$ is a linear estimator
- $E_m[\hat{\theta}_i^{dir}] = \mathbf{x}_i^T \beta = E_m[\theta_i]$, $\hat{\theta}_i^{dir}$ is an unbiased estimator (under FH model)
- $\min_{\hat{\theta}_i^{dir}} MSE_m(\hat{\theta}_i^{dir}) \rightarrow$

$$\tilde{\theta}_i^{BLUP} = \mathbf{x}_i^T \tilde{\beta} + \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2}(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \tilde{\beta}) = \mathbf{x}_i^T \tilde{\beta} + u_i \qquad (2)$$

so $\tilde{\theta}_i^{BLUP}$ is the *best* linear unbiased predictor

Fay Herriot Model

The BLUP can be rewritten as follows

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

Combination of a direct estimator and a synthetic estimator

- $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2}$
- σ_u is unknown
- $\tilde{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \hat{\theta}^{dir}$
- ψ_i^2 is assumed known (actually is the estimated MSE of direct estimate)
- $\tilde{\theta}_i^{BLUP}$ is a composite estimator

Fay Herriot Model

Using the joint distribution $f(\hat{\theta}_i^{dir}, u_i)$ under the Normality assumption we can get the Restricted Maximum Likelihood (REML) estimates of σ_u, say $\hat{\sigma}_u$

- Pluggin in $\hat{\beta}$ and $\hat{u}_i$ into (2) we get the Empirical BLUP (EBLUP)

$$\hat{\theta}_i^{EBLUP} = \mathbf{x}_i^T\hat{\beta} + \hat{u}_i = \mathbf{x}_i^T\hat{\beta} + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}(\hat{\theta}_i^{dir} - \mathbf{x}_i^T\hat{\beta}) \qquad (3)$$

- The EBLUP in equation 3 can be rewritten as follows

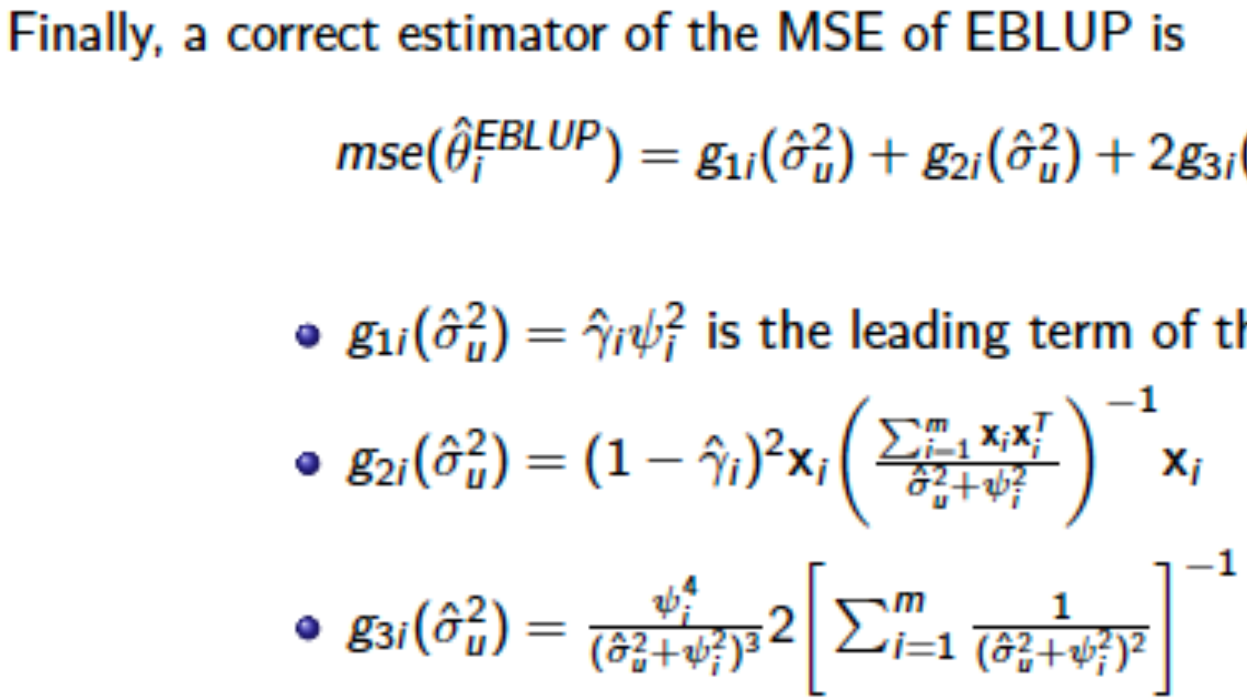
$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i\hat{\theta}_i^{dir} + (1 - \hat{\gamma}_i)\mathbf{x}_i^T\hat{\beta}$$

- $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}$
- $\hat{\gamma}_i$s are known as *shrinkage* factors

MSE under the Fay Herriot Model

The uncertainty of the FH Empirical BLUP has been studied by many researchers
A review of the formulas and approaches is in

RAO, J. N. K. and MOLINA, I. (2015). Small Area Estimation, Wiley, New Yersey, US.

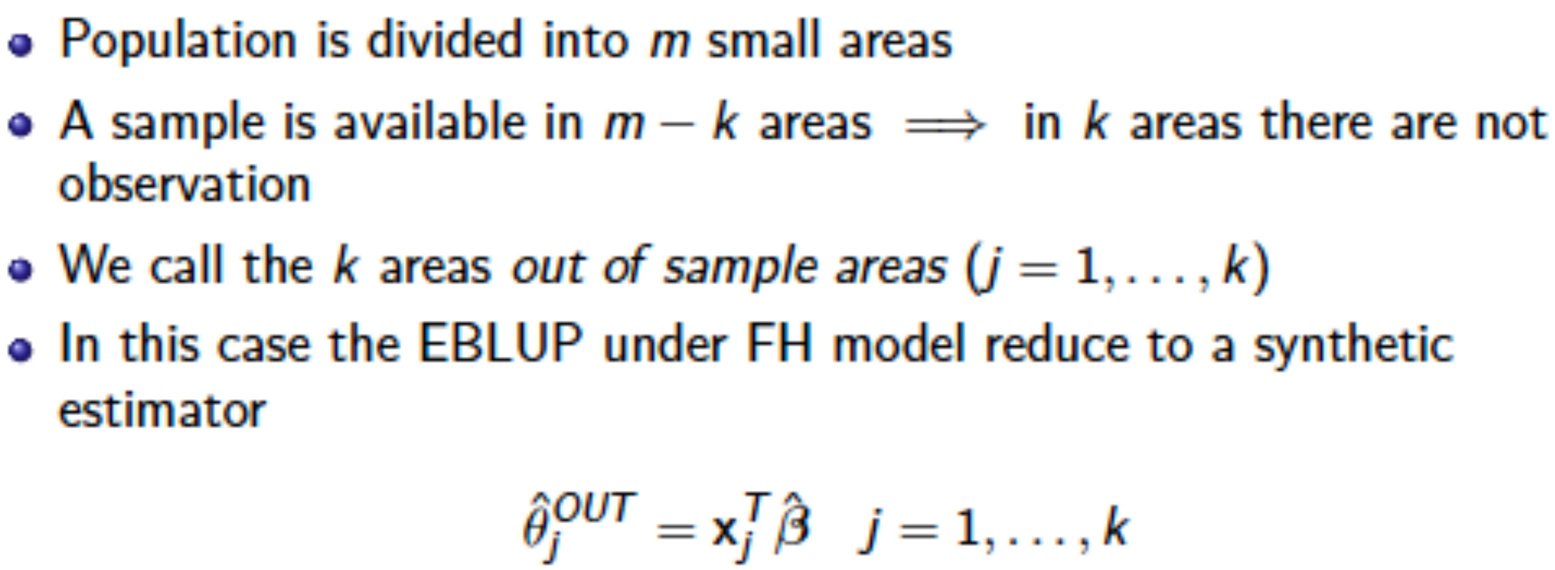


Finally, a correct estimator of the MSE of EBLUP is

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

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

Fay Herriot Model: out-of-sample areas

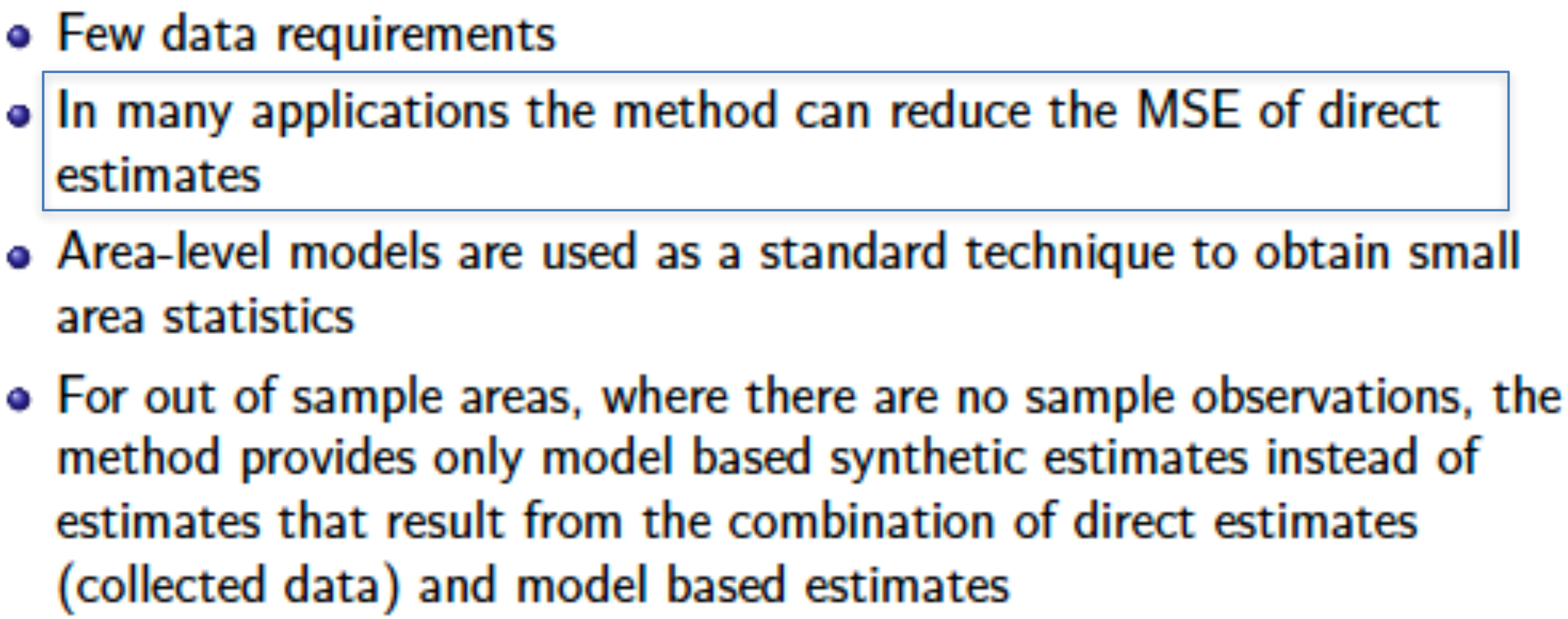


- Population is divided into m small areas
- A sample is available in $m-k$ areas $\Longrightarrow$ in k areas there are not observation
- We call the k areas *out of sample areas* $(j=1,\ldots,k)$
- In this case the EBLUP under FH model reduce to a synthetic estimator

$$\hat{\theta}_j^{OUT}=\mathbf{x}_j^T\hat{\beta}\quad j=1,\ldots,k$$

The synthetic estimator estimation is applicable when auxiliary information X is available for the k out-of-sample areas c

Fay Herriot Model: recap



- Few data requirements
- In many applications the method can reduce the MSE of direct estimates
- Area-level models are used as a standard technique to obtain small area statistics
- For out of sample areas, where there are no sample observations, the method provides only model based synthetic estimates instead of estimates that result from the combination of direct estimates (collected data) and model based estimates

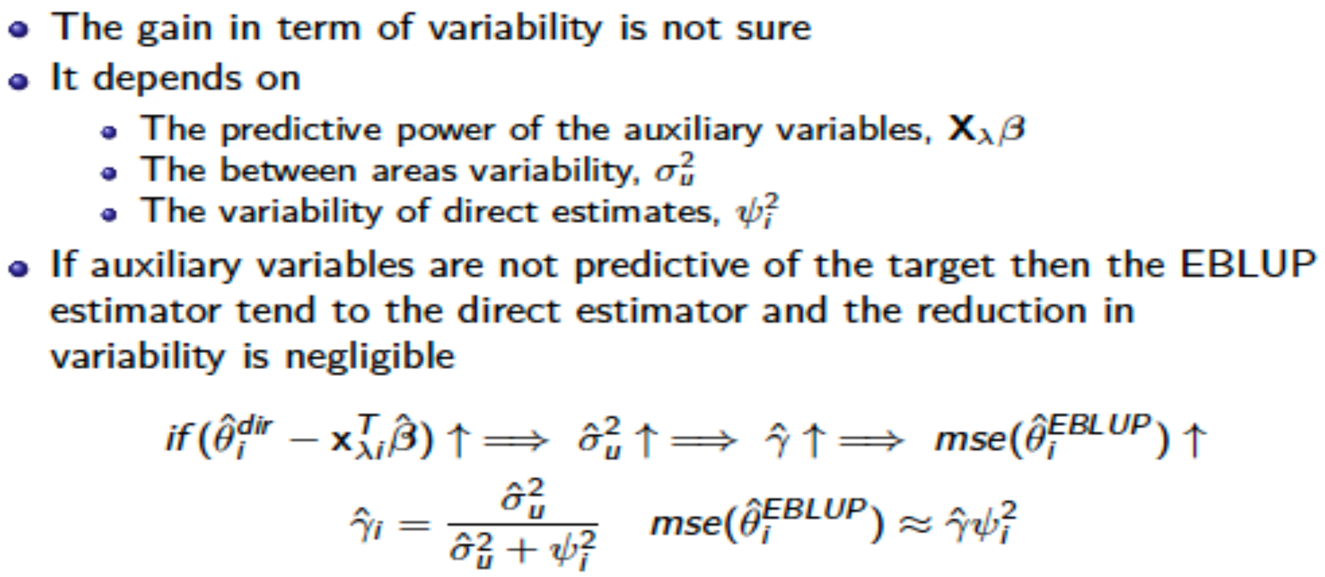
Fay Herriot Model: recap

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

Combination of a direct estimator (with sampling weights inside) and a synthetic estimator, balancing the two via the “shrinkage factor”

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}$$

Is the FH EBLUP more accurate than the direct estimator?
Yes, it is…but



- The gain in term of variability is not sure
- It depends on
  - The predictive power of the auxiliary variables, $\mathbf{X}_\lambda \beta$
  - The between areas variability, σ_u^2
  - The variability of direct estimates, ψ_i^2
- If auxiliary variables are not predictive of the target then the EBLUP estimator tend to the direct estimator and the reduction in variability is negligible

$$if(\hat{\theta}_i^{dir} - \mathbf{x}_{\lambda i}^T \hat{\beta}) \uparrow \Longrightarrow \hat{\sigma}_u^2 \uparrow \Longrightarrow \hat{\gamma} \uparrow \Longrightarrow mse(\hat{\theta}_i^{EBLUP}) \uparrow$$

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2} \qquad mse(\hat{\theta}_i^{EBLUP}) \approx \hat{\gamma}\psi_i^2$$

- Examples of application of the Fay-Herriot estimator during the R lab

Pros and cons of the model will be discussed after examples and applications to real data