



# Small Area Estimation, An Introduction

EMOS Learning Materials

Service Contract n. 2019.0249 between Eurostat and the University of  
Pisa, Italy

# Small Area Estimation, An Introduction

Service Contract n. 2019.0249 between Eurostat and the University of  
Pisa, Italy

# Outline

**Introduction to Small Area Estimation**

**Definitions**

**Design-Based Estimators**

**Synthetic Estimators**

**Composite Estimators**

**Model-based approach**

Area level models

Unit level models

**Concluding Remarks**

# Part I

## Introduction to Small Area Estimation

# Introduction to Small Area Estimation

- ▶ Problem: demand from official and private institutions of statistics referred to a given population of interest
- ▶ Possible solutions:
  - ▶ Census
  - ▶ Sample survey

Sample surveys have been recognized as cost-effectiveness means of obtaining information on wide-ranging topics of interest at frequent interval over time

# Introduction to Small Area Estimation

- ▶ Population of interest (or target population): population for which the survey is designed
    - *direct estimators* should be reliable for the target population
  - ▶ Domain: sub-population of the population of interest, they could be planned or not in the survey design
    - ▶ Geographic areas (e.g. Regions, Provinces, Municipalities, Health Service Area)
    - ▶ Socio-demographic groups (e.g. Sex, Age, Race within a large geographic area)
    - ▶ Other sub-populations (e.g. the set of firms belonging to an industry subdivision)
- we don't know the reliability of *direct estimators* for the domains that have not been planned in the survey design

# Introduction to Small Area Estimation

- ▶ Often *direct estimators* are not reliable for some domains of interest
- ▶ In these cases we have two choices:
  - ▶ oversampling over that domains
  - ▶ applying statistical techniques that allow for reliable estimates in that domains

## Small Domain or Small Area

Geographical area or domain where direct estimators do not reach a minimum level of precision

## Small Area Estimator (SAE)

An estimator created to obtain reliable estimate in a Small Area

# Introduction to Small Area Estimation: Example 1

- ▶ Target population: households who live in an Italian Region
- ▶ Variable of interest: Income
- ▶ Survey sample: EUSILC (European Union Statistics on Income and Living Conditions), designed to obtain reliable estimate at Regional level in Italy
  - ▶ planned design domains: Regions
  - ▶ unplanned design domains: e.g. Provinces, Municipalities
  
- ▶ EUSILC sample size in Tuscany: 1751 households
  - ▶ Pisa province 158 households → need SAE (or oversampling)
  - ▶ Grosseto province 70 households → need SAE (or oversampling)

## Introduction to Small Area Estimation: Example 2

- ▶ US sample sizes with an equal probability of selection method. Sample of 10,000 persons

State	1994 Population (thousands)	Sample size
California	31,431	1207
Texas	18,378	706
New York	18,169	698
⋮	⋮	⋮
DC	570	22
Wyoming	476	18

- ▶ Suppose to measure customer satisfaction for a government service:
- ▶ California 24.86% → leads to a confidence interval of 22.4%-27.3% (reliable); Wyoming 33.33% → leads to a confidence interval of 10.9%-55.7% (unreliable)

## Part II

# Classical Inference Approach

# Definitions

- ▶ Design-based estimation: the main focus is on the design unbiasedness. Estimators are unbiased with respect to the randomization that generates survey data
- ▶ Finite population  $\Omega = 1, \dots, N$
- ▶  $y$ : variable of interest, with  $y_i$  value of the  $i$ -th unit of the finite population
- ▶ Statistics of interest: e.g. total,  $Y = \sum_{\Omega} y_i$  or mean,  $\bar{Y} = Y/N$
- ▶ Sample  $s = 1, \dots, n$
- ▶  $p(s)$ : probability of selecting the sample  $s$  from population  $\Omega$ .  $p(s)$  depends on sampling design (variables such as stratum indicator and size measures of clusters)

# Definitions

- ▶ Bias of an estimator  $\hat{\theta}$  is defined as  $E[\hat{\theta} - \theta]$
- ▶ Variance of an estimator  $\hat{\theta}$  is defined as  $E[(\hat{\theta} - E[\hat{\theta}])^2]$
- ▶ Mean Squared Error of an estimator  $\hat{\theta}$  is  $E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + B[\hat{\theta}]^2$
- ▶ Design bias:  $Bias(\hat{Y}) = E_p[\hat{Y}] - Y$
- ▶ Design variance:  $V(\hat{Y}) = E_p[(\hat{Y} - y)^2]$

## Design-based properties

1. Design-unbiasedness:  $E_p[\hat{Y}] = \sum p(s) \hat{Y}_s = Y$
2. Design-consistency:  $\hat{Y} \rightarrow Y$  in probability

# Estimation of Means: Direct Estimator

- ▶ Data  $\{y_i\}, i \in s$
- ▶ Direct estimator for the mean (also known as expansion estimator):

$$\hat{Y} = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$$

- ▶  $w_i = \pi_i^{-1}$ , the basic design weight
- ▶  $\pi_i$  is the probability of selecting the unit  $i$  in sample  $s$

Remark: weights  $w_i$  are independent from  $y_i$

# Domain Estimation

- ▶ Let partitioning population  $\Omega$  into  $m$  partitions or domains:

$$\Omega = \cup_{i=1}^m \Omega_i$$

- ▶  $\Omega_i = 1, \dots, N_i$ , population of the domain  $i$
- ▶  $s_i = 1, \dots, n_i$ , sample of the domain  $i$
- ▶ Statistics of interest for the variable  $y$ :
  - ▶  $Y_i = \sum_{\Omega_i} y_j$ , the total of domain  $i$
  - ▶  $\bar{Y}_i = Y_i/N_i$ , the mean of domain  $i$

# Domain Estimation: Direct Estimator

- ▶ Data  $\{y_{ij}\}, j \in s_i, i = 1, \dots, m$
- ▶ Direct estimator of the mean for the domain  $i$ :

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

- ▶ where  $y_{ij}$  is the observation value and  $w_{ij}$  is the weight for unit  $j$  in area  $i$
- ▶ The case of the simple random sampling:

- ▶  $\pi_{ij} = \frac{\binom{1}{1} \binom{N_i-1}{n_i-1}}{\binom{N_i}{n_i}} = \frac{n_i}{N_i} \rightarrow w_{ij} = \pi_{ij}^{-1} = \frac{N_i}{n_i}$

- ▶  $\hat{Y}_i = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \frac{\sum_{j=1}^{n_i} \frac{N_i}{n_i} y_{ij}}{\sum_{j=1}^{n_i} \frac{N_i}{n_i}} = \frac{N_i}{n_i} \frac{\sum_{j=1}^{n_i} y_{ij}}{N_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$  (that is the sample mean)

# Domain Estimation: Direct Estimator

- ▶  $\hat{Y}_i$  is design unbiased
- ▶  $\hat{V}(\hat{Y}_i) = (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}$ , where  $S_i^2 = \frac{\sum_{j \in s_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1}$  is the sample variance
- ▶ The magnitude of the variance depends on 3 factors:  $n_i/N_i$ ,  $S_i^2$  and  $n_i$
- ▶ If  $n_i$  is small the design variance is likely to be large
- ▶ In such a situation, estimation of variance is even more problematic

# Synthetic Estimators

- ▶ 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 Estimation with no auxiliary variable (dummy estimator)

- ▶ Implicit model assumed:

$$y_j = \beta + \varepsilon_j, \quad j \in \Omega$$

- ▶ Synthetic estimator for the mean:

$$\hat{Y}_{i,S} = \frac{\sum_{j \in S} w_j y_j}{\sum_{j \in S} w_j} = \hat{Y}$$

- ▶  $E_p[\hat{Y}_{i,S} - \bar{Y}_i] \approx \bar{Y} - \bar{Y}_i$ , the bias relative to  $\bar{Y}_i$  is small if  $\bar{Y}_i \approx \bar{Y}$

# Synthetic Estimation with auxiliary variables

- ▶ Implicit model assumed:

$$y_j = \mathbf{X}'_j \boldsymbol{\beta} + \varepsilon_j, j \in \Omega_i$$

- ▶ Synthetic estimator:

$$\hat{Y}_{i,GRS} = \bar{\mathbf{X}}'_i \hat{\boldsymbol{\beta}}$$

- ▶  $\hat{\boldsymbol{\beta}} = (\sum_{j \in \mathcal{S}} w_j \mathbf{x}_j \mathbf{x}'_j / c_j)^{-1} (\sum_{j \in \mathcal{S}} w_j \mathbf{x}_j y_j / c_j)$

# Synthetic Estimation with auxiliary variables

- ▶  $E_p[\hat{Y}_{i,GRS} - \bar{Y}_i] \approx \bar{\mathbf{X}}_i' \boldsymbol{\beta} - \bar{Y}_i$ , expected bias
- ▶  $\boldsymbol{\beta} = (\sum_{j \in \Omega} \mathbf{x}_j \mathbf{x}_j' / c_j)^{-1} (\sum_{j \in \Omega} \mathbf{x}_j y_j / c_j)$
- ▶ The relative bias is small if both of the following conditions are satisfied
  - $\beta_i = \beta$ , where  $\beta_i = (\sum_{j \in \Omega_i} \mathbf{x}_j \mathbf{x}_j' / c_j)^{-1} (\sum_{j \in \Omega_i} \mathbf{x}_j y_j / c_j)$
  - $Y_i = \mathbf{X}_i' \boldsymbol{\beta}_i$

# Composite Estimators

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
- ▶  $\phi_i$  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

# The Choice of $\phi_i$

Optimal  $\phi_i$

- a. Minimize the  $MSE(\hat{Y}_{i,C})$  with respect to  $\phi_i$  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  $\phi_i$  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 $\phi_i$

- b.** Minimize  $m^{-1} \sum_{i=1}^m \text{MSE}(\hat{Y}_{i,C})$  with respect to a common weight  $\phi_i = \phi$
- ▶ the optimal solution is given by

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

- ▶ the parameter  $\phi$  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 from 1981 Population Census. Samples are drawn following the LFS design (two stages with stratification)
- ▶ 400 sample replicates, each of identical size of the LFS sample
- ▶ 14 Health Service Areas (HSA) of the Friuli Region are considered to be small areas

# 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 approximately 30% higher than Synthetic and Composite estimator

## Part III

# Introduction to Small Area Predictors based on Small Area models

# Recap

- ▶ Domain: sub-population of the population of interest, they could be planned or not in the survey design
    - ▶ Geographic areas (e.g. Regions, Provinces, Municipalities, Health Service Area)
    - ▶ Socio-demographic groups (e.g. Sex, Age, Race within a large geographic area)
    - ▶ Other sub-populations (e.g. the set of firms belonging to a industry subdivision)
- we don't know the reliability of *direct estimators* for the domains that have not been planned in the survey design

# Types of models & Data requirements

- ▶ Unit-level models
  - ▶ Use unit-level data (e.g. from surveys) for model fit
  - ▶ Area level covariates (predictor variables) are sufficient for estimating small area averages/proportions
  - ▶ Access to unit-level data → possible confidentiality issues
- ▶ Area-level models
  - ▶ Use only area-level data for model fit and SAE
  - ▶ Model specified at the area-level
  - ▶ Data access possibly less complex than access to unit-level data

# Area Level Approach, the Fay-Harriot Model

- ▶ The area level model includes random area-specific effects and area specific covariates  $\mathbf{x}_i$

$$\theta_i = \mathbf{x}_i\boldsymbol{\beta} + z_i u_i, \quad i = 1, \dots, m$$

- ▶  $\theta_i$  is the parameter of interest (e.g. totals,  $Y_i$  or means,  $\bar{Y}_i$ )
- ▶  $z_i$  are known positive constant
- ▶  $u_i$  are independent and identically distributed random variables with mean 0 and variance  $\sigma_u^2$  ( $u_i \sim N(0, \sigma_u^2)$ )
- ▶  $\boldsymbol{\beta}$  is the regression parameters vector

# Area Level Approach, the Fay-Harriot Model

Assumption

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

- ▶  $\hat{\theta}_i$  is a direct design-unbiased estimator
- ▶  $e_i$  are independent sampling error with mean 0 and known variance  $\psi_i$

The Fay-Harriot Model is obtained as

$$\hat{\theta}_i = \mathbf{x}_i\boldsymbol{\beta} + z_i v_i + e_i, \quad i = 1, \dots, m$$

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

# Area Level Approach, the Fay-Harriot Model

Under above mentioned assumptions

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

Let us to introduce matrix notation

- ▶  $\hat{\boldsymbol{\theta}}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$
- ▶  $\mathbf{u} \sim N(0, \mathbf{G})$  and  $\mathbf{e} \sim N(0, \mathbf{R})$
- ▶  $\hat{\boldsymbol{\theta}} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})$ , let  $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$

Given the estimates of  $\boldsymbol{\beta}$  and  $\mathbf{u}$  we obtain the Best Linear Unbiased Predictor (BLUP) for  $\boldsymbol{\theta}$

- ▶  $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$
- ▶  $\tilde{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\tilde{\boldsymbol{\beta}})$

Note: estimates for  $\boldsymbol{\beta}$  and  $\mathbf{u}$  can be obtained by penalized maximum likelihood ( $\mathbf{u}$  considered as fix).

# Area Level Approach, the Fay-Harriot Model

- ▶ In the “real world”  $\mathbf{G}$  (and  $\mathbf{R}$ ) are unknown and they must be estimated
- ▶ Using restricted likelihood optimized with scoring algorithm we obtain estimates for  $\sigma_u^2$  ( $\mathbf{G}$ )
- ▶ In the Fay-Herriot model  $\psi_i$  is considered as known (we use sampling variance)

Plugging in the estimated area-specific variance component  $\hat{\sigma}_u^2$  in the estimator for  $\beta$  and  $\mathbf{u}$  we obtain their estimates

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\theta} \\ \hat{\mathbf{u}} &= \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\hat{\theta} - \mathbf{X}\hat{\beta})\end{aligned}$$

# Area Level Approach, the Fay-Harriot Model

Finally using the obtained estimates in the Fay-Herriot model we have the Empirical BLUP (EBLUP) for the parameter of interest  $\theta$

$$\hat{\theta}_i^{FH} = \hat{\phi}_i \hat{\theta}_i + (1 - \hat{\phi}_i)(\mathbf{x}'_i \hat{\beta})$$

- ▶  $\hat{\phi}_i = \frac{z_i^2 \hat{\sigma}_u^2}{z_i^2 \hat{\sigma}_u^2 + \psi_i}$ , is the *shrinkage factor*
- ▶  $\hat{\theta}_i$  is the design estimator for  $\theta_i$

Note: using this procedure could happen that the estimate of  $\sigma_u^2$  is negative, in this case it must be truncated to 0

# Area Level Approach, the Fay-Harriot Model

The MSE of the Fay-Herriot small area estimator is

$$MSE(\hat{\theta}_i^{FH}) = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2)$$

- ▶  $g_{1i}(\sigma_u^2) = \phi_i \psi_i$  is due to the variability of random errors
- ▶  $g_{2i}(\sigma_u^2)$  is due to the variability of  $\beta$  estimate
- ▶  $g_{3i}(\sigma_u^2)$  is due to the variability of the estimate of  $\sigma_u^2$

An approximately correct estimate of the MSE is

$$\widehat{MSE}(\hat{\theta}_i^{FH}) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2)$$

Remark: Alternatively (for more complex models) use bootstrap or jackknife methods

# Example, Estimation of Mean Consumption Expenditure in Italian provinces

- ▶ Goal: obtain reliable estimate of the household consumption expenditure at provincial level (LAU 1/NUTS 3) in Italy
- ▶ Reliable consumption expenditure estimates in Italy are available at regional level (NUTS 2)
- ▶ Estimates at regional level do not capture (reflect) the heterogeneity of households' consumption behaviour and living conditions within each region (at provincial or at sub-provincial level)
- ▶ Depending on the availability of data, small area estimation methods can reach the goal

# Example, Estimation of Mean Consumption Expenditure in Italian provinces

Available data:

- ▶ Consumption expenditures 2012, unit level data from Household Budget Survey (HBS)
- ▶ Per capita taxable income in 2012, province level data from Italian Revenue Authority
- ▶ Home ownership (percentage of household who have the ownership of the house where they live), province level data from Population Census 2011

# Example, Estimation of Mean Consumption Expenditure in Italian provinces

- ▶ We use the EBLUP based on the Fay-Herriot model to estimate the mean of the equivalised consumption expenditures for the 110 Italian provinces
- ▶ We use the modified-OECD equivalence scale which assigns a value of 1 to the first adult in the household, 0.5 to each other adult and 0.3 to each child under 14
- ▶ Direct estimates of the mean equivalised consumption expenditure and their estimated variances are obtained from the HBS

# Example, Estimation of Mean Consumption Expenditure in Italian provinces

## Fay-Herriot model

- ▶ Direct estimate of the mean of equivalised consumption expenditures for each Italian province and its estimated variance is obtained from the HBS
- ▶ Small sample size at province level of the HBS reveal unreliable direct estimates, as expected. Sample sizes ranges between 4 and 1037, with quartiles equal to 85, 147 and 303
- ▶ Auxiliary variables: per capita taxable income, home ownership

# Example, Estimation of Mean Consumption Expenditure in Italian provinces

**Table:** FH-model regression parameters estimation

Aux Var	$\hat{\beta}$	Std. Err.	t-value	p-value
Intercept	-1966.75	504.99	-3.89	0.000
Taxable income	0.16	0.01	15.49	0.000
Home ownership	19.18	6.40	3.00	0.003

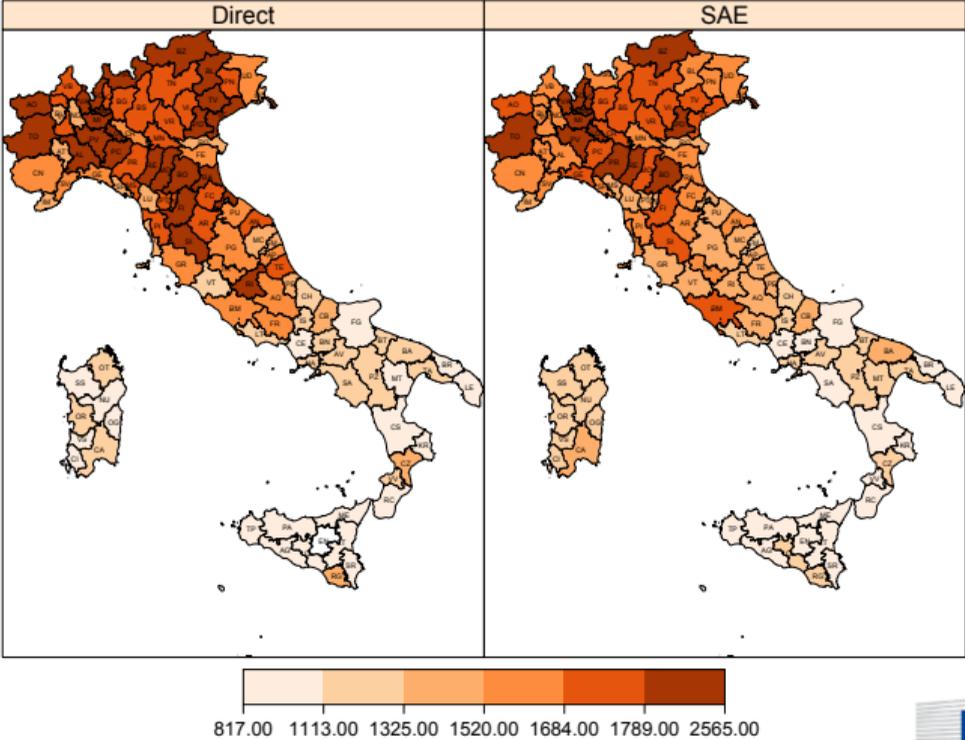
$\hat{\sigma}_u = 184$

# Example, Estimation of Mean Consumption Expenditure in Italian provinces

**Table:** Summary across Italian provinces of direct and small area estimates of the mean consumptions expenditure

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{\theta}_i^{dir}$	1242	1867	2395	2339	2705	3780
$\hat{\theta}_i^{FH}$	1375	1812	2307	2245	2602	3332

# Example, Estimation of Mean Consumption Expenditure in Italian provinces



# Example, Estimation of Mean Consumption Expenditure in Italian provinces

Precision gain of the SAE estimates

**Table:** Summary among provinces of the  $rmse(\hat{\theta}_i^{FH})$  and  $rmse(\hat{\theta}_i^{dir})$  ratio and the CV of direct and small area estimates

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$rmse(\hat{\theta}_i^{FH})/rmse(\hat{\theta}_i^{dir})$	0.10	0.53	0.64	0.64	0.80	0.97
$CV^{dir}$	0.04	0.07	0.11	0.12	0.14	0.56
$CV^{FH}$	0.04	0.06	0.07	0.07	0.08	0.11

# Example, Estimation of Mean Consumption Expenditure in Italian provinces

- ▶ The small area estimates of the mean consumption expenditure are sound
- ▶ Indeed, the reduction of the variability of the estimates is  $1 - 0.64 = 0.36 = 36\%$  in mean
- ▶ Direct and SAE point estimates are quite similar, but SAE fixes some extreme unrealistic values of the direct estimate
- ▶ The success of the exercise depend mainly on the predictive power of the auxiliary variables, that is high in this application

# Small Area Estimation by Borrowing Strength over Space

- ▶ In applications involving economic, environmental and epidemiological data observations that are spatially close may be more alike than observations that are further apart
- ▶ This creates a type of spatial dependency or spatial association in the data that invalidates the assumption of independent and identically distributed (iid) observations used by conventional regression models
- ▶ One approach to accounting for spatial correlation in the data is offered by specifying models with spatially correlated errors (Anselin 1992; Cressie 1993)
- ▶ Small area literature suggests that prediction of small area parameters may be improved by borrowing strength over space (Saei and Chambers 2003; Singh *et al.* 2005; Petrucci and Salvati 2006; Pratesi and Salvati 2007, 2009)

# Extension of the basic FH model: Spatial FH model

Under spatial relationship the FH model becomes:

$$\hat{\theta}_i = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

where

- ▶  $\mathbf{v} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}$  and  $\mathbf{v} \sim N(0, \mathbf{G})$  with  $\mathbf{G} = \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})^T(\mathbf{I} - \rho\mathbf{W})]^{-1}$ ;
- ▶  $\mathbf{W}$  is a  $m \times m$  spatial interaction matrix which indicates whether the areas are neighbour or not (on way to define  $\mathbf{W}$  is to set  $w_{ij} = 1$  if small area  $i$  and  $j$  are neighbour or 0 otherwise, however there are other ways to define  $\mathbf{W}$ );
- ▶  $\rho$  is the spatial autoregressive coefficient which defines the strength the spatial relationship among the random effects associated with the neighbouring areas.

# Spatial FH model

The Spatial EBLUP (SEBLUP) is obtained as

$$\tilde{\theta}_i^S = \mathbf{x}_i \hat{\beta} + \hat{v}_i$$

- ▶ A second order approximation of the MSE of the SEBLUP has been proposed by Singh et al. 2005 and Petrucci and Salvati (2006).
- ▶ Analytical approximations may require strong model assumptions and many small areas to approximate well the true values, therefore Molina et al. (2009) proposed parametric and non-parametric bootstrap procedures for estimation of the MSE under the SFH model.

# Other Spatial models for SAE

Other approaches to incorporate the spatial structure in the data are:

- ▶ Spatially non-stationary Fay-Herriot model (Chandra et al., 2015).
- ▶ Non-parametric spatial P-spline model for small area estimation (Opsomer et al., 2008; Giusti et al. 2012).

# EBLUP: Unit Level Approach

Unit level approach to small area estimation

- ▶  $\mathbf{y}$  the vector for the  $y$  variable for the population  $\Omega$
- ▶  $\mathbf{y} = [\mathbf{y}'_s, \mathbf{y}'_r]'$ , where  $\mathbf{y}_s$  is the vector of the observed units (the sampled ones) and  $\mathbf{y}_r$  is the vector of the non observed units ( $N - n$ ,  $r = 1, \dots, N - n$ )
- ▶  $\mathbf{X}$  is the covariates matrix and is considered known for all the population units
- ▶ Subscript  $i$  refers to small areas (e.g.  $\mathbf{y}_{s_i}$  is the vector of observed variables in area  $i$ )
- ▶ Model for the  $y$  variable (known as superpopulation model)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- ▶ that can be alternatively write as

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i + e_{ij}$$

# EBLUP: Unit Level Approach

Given that

- ▶  $\mathbf{u} \sim N(0, \mathbf{G})$ ,  $\mathbf{e} \sim N(0, \mathbf{R})$  and  $\mathbf{u} \perp \mathbf{e}$
- ▶  $\mathbf{R} = \sigma_e^2 \Sigma_e$ ,  $\mathbf{G} = \sigma_u^2 \Sigma_u$
- ▶  $\mathbf{X}$  is a full rank matrix (say rank equal  $q$ ) and rank of  $(\mathbf{X} : \mathbf{Z}_i) > q$
- ▶  $n \geq q + m + 1$

it can be shown that

- ▶  $V(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R} = \mathbf{V}$
- ▶  $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}_s$  is the BLUE for  $\beta$
- ▶  $\tilde{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta})$  is the BLUP for  $\mathbf{u}$
- ▶  $\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{V})$

# EBLUP: Unit Level Approach

Let  $\theta$  be a statistic of interest (e.g. the mean or the total of  $y$  in a given area) that we want to estimate. It is possible to express  $\theta$  in terms of linear combination between observed and unobserved units

$$\theta = \alpha'_s \mathbf{y}_s + \alpha'_r \mathbf{y}_r$$

- ▶  $\alpha = (\alpha'_s, \alpha'_r)$  is a vector of known constants of dimension  $N$

The estimate for  $\theta$  is easily obtained substituting the unknown vector  $\mathbf{y}_r$  with its prediction  $\hat{\mathbf{y}}_r$

$$\hat{\theta} = \alpha'_s \mathbf{y}_s + \alpha'_r \hat{\mathbf{y}}_r$$

- ▶  $\hat{\mathbf{y}}_r = \mathbf{X}_r \hat{\beta} + \hat{\mathbf{u}}$

# EBLUP: Unit Level Approach

For example the estimator of the mean for the  $i$ -th area is obtained as follow

$$\hat{Y}_i = \frac{1}{N_i} \left\{ \sum_{j \in S_i} y_{ij} + \sum_{k \in r_i} (\mathbf{x}'_{ik} \hat{\beta} + \hat{u}_i) \right\}$$

we can use the composite estimator form

$$\hat{Y}_i = \hat{\phi}_i [\hat{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i) \hat{\beta}] + (1 - \hat{\phi}_i) [\bar{\mathbf{X}}_i \hat{\beta}]$$

$$\blacktriangleright \hat{\phi}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$$

# EBLUP: Unit Level Approach

Next step is to derive an MSE estimator

- ▶  $MSE(\hat{\theta}_i) \approx g_{1i}(\sigma) + g_{2i}(\sigma) + g_{3i}(\sigma)$
- ▶  $g_{1i}(\sigma) = \alpha_r' \mathbf{Z}_r \mathbf{T}_s \mathbf{Z}_r' \alpha_r$
- ▶  $g_{2i}(\sigma) = [\alpha_r' b \mathbf{X}_r - \alpha_r' \mathbf{Z}_r \mathbf{T}_s \mathbf{Z}_s' \mathbf{R}_s' \mathbf{X}_s] (\mathbf{X}_s' \mathbf{V}^{-1} \mathbf{X}_s)^{-1} [\mathbf{X}_r' \alpha_r - \mathbf{X}_s' \mathbf{R}_s^{-1} \mathbf{Z}_s \mathbf{T}_s \mathbf{Z}_r' \alpha_r]$
- ▶  $g_{3i}(\sigma) = tr\{(\nabla(\alpha_r' \mathbf{Z}_r \Sigma_u \mathbf{Z}_s' \mathbf{V}_s^{-1})')' \mathbf{V}_s (\nabla(\alpha_r' \mathbf{Z}_r \Sigma_u \mathbf{Z}_s' \mathbf{V}_s^{-1})')' E[(\hat{\sigma} - \sigma)(\hat{\sigma} - \sigma)']\}$
- ▶  $\mathbf{T} = \Sigma_u - \Sigma_u \mathbf{Z}_s' (\Sigma_{es} + \mathbf{Z}_s \Sigma_u \mathbf{Z}_s')^{-1} \mathbf{Z}_s \Sigma_u$
- ▶  $\sigma = (\sigma_e^2, \sigma_u^2 / \sigma_e^2)'$

# EBLUP: Unit Level Approach

Finally, the estimator for the MSE of  $\hat{\theta}_i$  is

$$\widehat{MSE}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}) + g_{2i}(\hat{\sigma}) + 2g_{3i}(\hat{\sigma})$$

- ▶  $\hat{\sigma}$  is an unbiased estimator for  $\sigma$

Remark: it is possible to obtain an estimate of the MSE using alternative techniques, such as bootstrap and jackknife

# Example, Estimate of Mean Income in Tuscany Provinces

- ▶ Data on the equivalised income in 2005 for 1525 households in the 10 Tuscany Provinces are available from the EUSILC survey 2006
- ▶ A set of explanatory variables is available for each unit in the population from the Population Census 2001
- ▶ We employ the unit level small area model to estimate the mean of the household equivalised income
- ▶ The Municipality of Florence, with 125 units out of 457 in the Province, is considered as a stand-alone area

# Example, Estimate of Mean Income in Tuscany Provinces

- ▶ The selection of covariates to fit the small area model relies on prior studies on poverty assessment
- ▶ The following covariates have been selected:
  - ▶ household size
  - ▶ ownership of dwelling (owner/tenant)
  - ▶ age of the head of the household
  - ▶ years of education of the head of the household
  - ▶ working position of the head of the household (employed/unemployed in the previous week)
- ▶ Design-based estimates of the mean income have been carried out in order to show the gain in efficiency of the EBLUP

# Example, Estimate of Mean Income in Tuscany Provinces

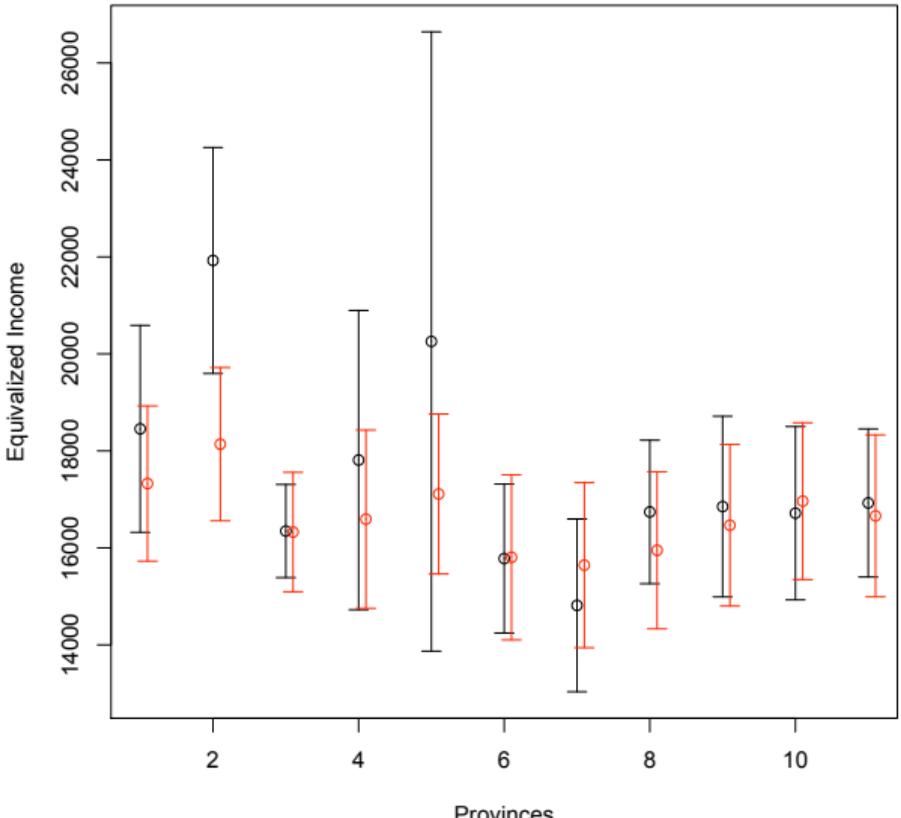
**Table:** Mean Income Estimate. Tuscany Provinces

Provinces	<i>EBLUP</i>	$\widehat{RMSE}_{EBLUP}$	<i>Design-Based</i>	$\widehat{RMSE}_{DB}$
Arezzo	17328	816.6	18455	1088.9
Florence M.	18139	806.4	21927	1188.8
Florence	16327	628.2	16347	490.1
Grosseto	16593	937.8	17811	1574.5
Livorno	17111	841.5	20257	3258.4
Lucca	15805	868.7	15780	783.5
Massa	15644	868.0	14814	909.0
Pisa	15950	826.5	16741	755.0
Pistoia	16467	850.1	16852	950.4
Prato	16964	824.9	16715	911.5
Siena	16660	852.0	16926	779.4

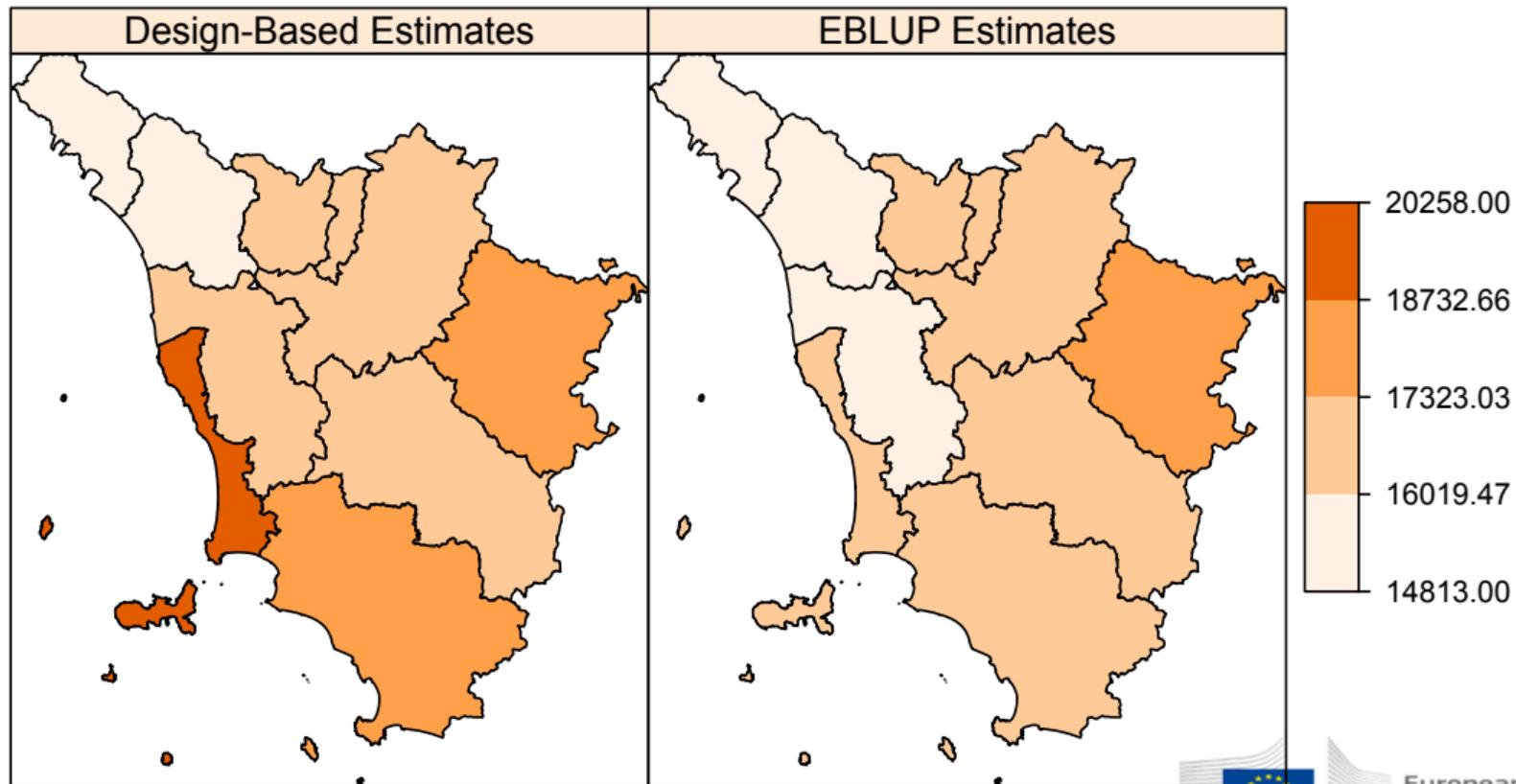


# Example, Estimate of Mean Income in Tuscany Provinces

Error bar for mean income estimates



# Example, Estimate of Mean Income in Tuscany Provinces



# Conclusions on EBLUP

- ▶ The EBLUPs are either a *composite estimators* and a *model-based estimators*
- ▶ EBLUPs can also be used when we know only the average of the auxiliary variables
- ▶ In many applications the EBLUPs perform better than the design based estimators in terms of relative root MSE (smaller confidence intervals)
- ▶ Actually EBLUPs are used as a standard technique to derive small area statistics

*The Empirical Best Linear Unbiased Predictor is the Industry Standard for Small Area Estimation*

# Conclusions on EBLUP

## Drawbacks

- ▶ Assumption of normality is needed for area effects and individual effects (but sensitivity analysis shows that the model is robust against non normality if symmetry of the distributions hold)
- ▶ It is not design-unbiased, in the sense that under complex survey design the estimates could be biased
- ▶ Parameters of interest in out of samples areas (areas with 0 observations) cannot be estimated (EBLUP needs minimum two observations per area)
- ▶ Extensions to the model are not easily implementable (complex derivation of the MSE estimator)

# Conclusions on EBLUP

- ▶ Improvements for the EBLUP
  - ▶ Spatial process (CAR and SAR models)
  - ▶ Time process
  - ▶ Spatiotemporal process
  - ▶ Robust estimation
  - ▶ Binary and count data models
  
- ▶ Alternative approaches
  - ▶ Quantile/M-Quantile approach
  - ▶ Bayesian approach

## Part IV

# Concluding Remarks

# On Going Research on Small Area Estimation

- ▶ Robustness in small area estimation
- ▶ Small area quantiles estimators (distribution function estimator)
- ▶ New models for small area problem
- ▶ Inclusion of the design weights in model-based (composite) estimators
- ▶ Models for non continuous data
- ▶ Big data and SAE

# Essential bibliography

- ▶ Battese, G., Harter, R. and Fuller, W. (1988) An error-components model for prediction of county crop areas using survey and satellite data. *J. Am. Statist. Ass.*, **83**, 28–36.
- ▶ Chambers, R. and Clark, R. (2012). *An Introduction to Model-based Survey Sampling with Applications*. Oxford: Oxford University Press.
- ▶ Fay RA, Herriot RE (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.
- ▶ Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68**, 2, 221–238.
- ▶ Lombardia M.J., Gonzalez-Manteiga W. and Prada-Sanchez J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of finite population distribution function. *Journal of Statistical Planning and Inference*, **116**, 367–388.
- ▶ Prasad N, Rao J (1990). The estimation of the mean squared error of small-area estimators. *J Am Stat Assoc.* **85**, 163–171.
- ▶ Pratesi M, Salvati N (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat Methods Appl*, **17**, 113–141.
- ▶ Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*. New York: Wiley.
- ▶ Royall, R. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, **73**, 351–358.
- ▶ Salvati, N., Chandra, H., Ranalli, M. G. and Chambers, R. (2010) Small area estimation using a nonparametric model-based direct estimator. *Computnl Statist. Data Anal.*, **54**, 2159–2171.
- ▶ Salvati, N., Chandra, H., Ranalli, M.G. and Chambers, R. (2010). Small Area Estimation Using a Nonparametric Model Based Direct Estimator. *Journal of Computational Statistics and Data Analysis*, **54**, 2159–2171.