



INSTITUTO NACIONAL DE ESTATÍSTICA  
STATISTICS PORTUGAL

# IWSM

International Workshop on Statistical Modelling



## » Small Area Estimation for Land Use and Land Cover

Pedro Campos, Suelma Pina, A. Manuela Gonçalves

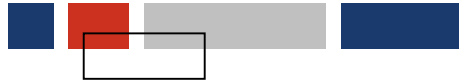
Statistics Portugal, Universidade do Minho

*Pedro.campos@ine.pt*



Guimarães, 11th July, 2019

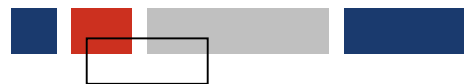




# Outline



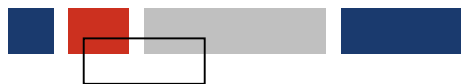
- » Small Area Estimation
- » Estimators
- » Data and Software
- » Results
- » Conclusions



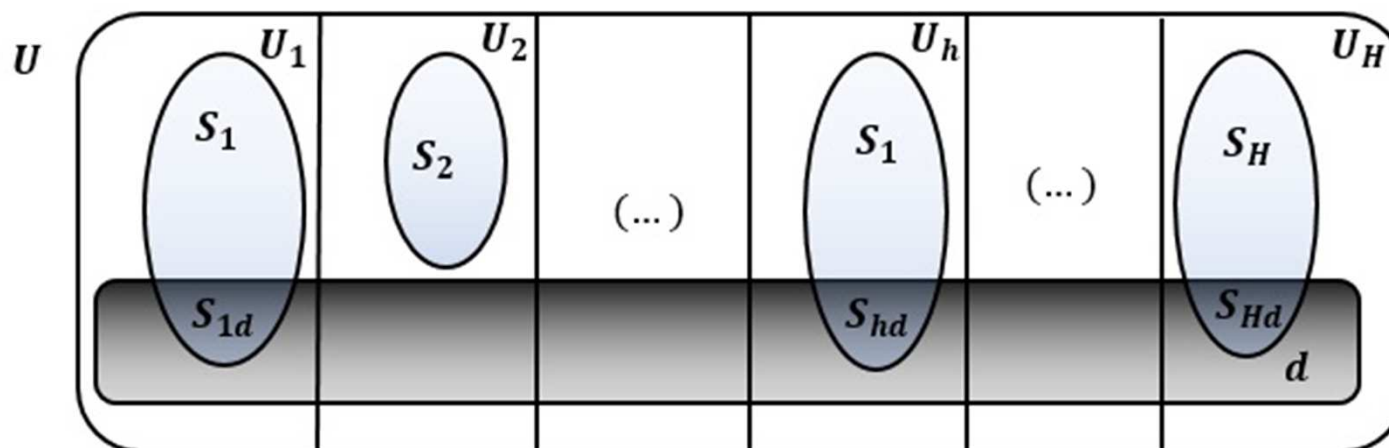
# Small Area Estimation



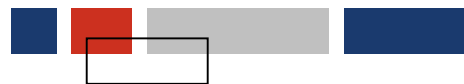
- Nowadays, public and private institutions are increasingly seeking more detailed information
- The need to produce reliable estimates for the total of variables of interest in small domains is fundamental.
- However, estimates cannot always be obtained through direct estimators because often:
  - **there are no samples for these domains,**
  - **or they are too small to obtain sufficient quality estimates.**



# Small Area Estimation



Dimension of stratum  $U_h$  is  $N_h$ ,  $h = 1, \dots, H$ , where  $N = \sum_{h=1}^H N_h$



# Small Area Estimation



- SAE models can be categorized in direct and indirect estimators.
  - **Direct estimators** only consider the observations of the variable of interest belonging to the study domain for the time period under analysis,
  - **Indirect estimators** take observations of the variable of interest as well as auxiliary sources outside the study domain for the considered period of time.



# Small Area Estimation

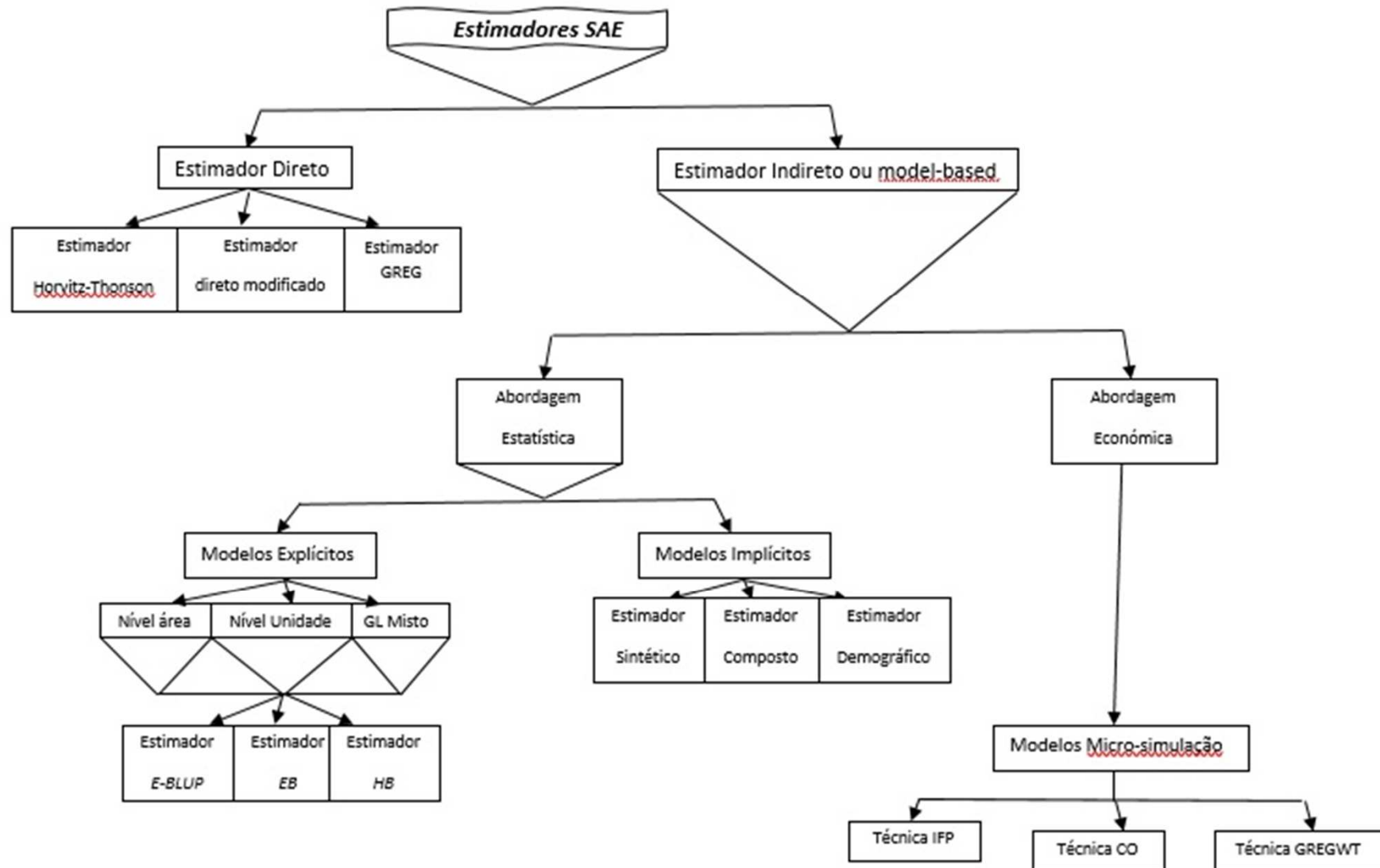


- The Model-based approach belongs to the class of indirect estimators
- Regression models use auxiliary variables from other data sources, such as census and administrative records to "lend" information from similar areas (Rao & Molina, 2015).



# Small Area Estimation

A broader view...





# Estimators



## » *Direct estimators*

- Horvitz-Thompson

$$D_1 = \hat{\tau}_{D1} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{i \in s_{hd}} y_i$$

$$D_2 = \hat{\tau}_{D2} = \sum_{h=1}^H \frac{N_{hd}}{n_{hd}} \sum_{i \in s_{hd}} y_i$$

$$Var(\hat{\tau}_{D1}) = \sum_{h=1}^H \frac{N_{hd}(N_h - n_h)}{n_h} (s_{hd}^2 + (1 - \frac{N_{hd}}{N_h}) \bar{y}_{hd}^2)$$







# Estimators

» *Direct estimator modified by regression*



$$\hat{\tau}_{d,reg} = \hat{\tau}_d + (\tau_{xd} - \hat{\tau}_{xd})' \hat{\beta}_g$$

where  $\hat{\beta}_g$  is the estimator of regression parameters  $\beta_g = (\beta_{g1}, \beta_{g2}, \dots, \beta_{gp})'$ .  
In this case there is an implicit link model:  $y_i = x' \beta_g + \epsilon_i$ , with  $i \in U_g$



# Estimators



## » EBLUP and SEBLUP

The EBLUP is a combined estimator.

Considering a finite population divided into  $D$  small domains, the Fay-Herriot base model (Rao & Molina, 2015) linearly relates the value of the  $d$ -th domain of the variable of interest  $\theta_d$  to a vector of  $p$  auxiliary variables aggregated at the  $x_d$  **area level** and includes an associated random  $v_d$  effect.

$$\hat{\theta}_{SEBLUP} = x'_d\beta + v_d + e_d = x'_d\beta + (I_D - \rho W)^{-1}u + e_d$$

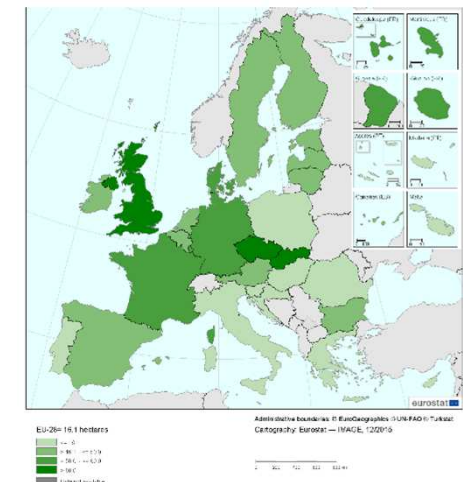
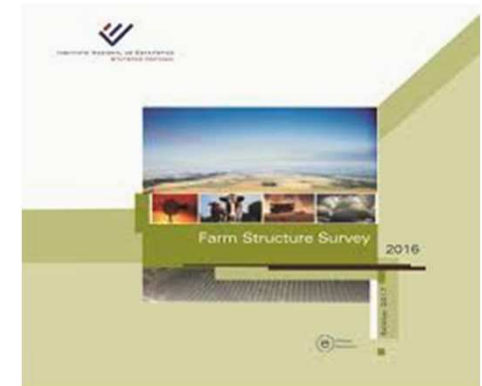
The SEBLUP estimator considers a spatial component. The main difference between the two models (EBLUP and SEBLUP) lies in the fact that SEBLUP uses the information of the distances between the domains through a proximity matrix (Pfeffermann, 2013).



# Data and Software

## » Farm Structure Survey

- The Farm Structure Survey (FSS), also known as the Survey on the structure of agricultural holdings, is carried out by all European Union (EU) Member States and provides comparable statistics across countries and time, at regional levels (down to NUTS 3 level).
- The edition of 2013 considers more than 650 variables. In this study several strata has been considered, based on size class, area status, legal status of the holding, objective zone and farm type (INE, 2013).
- .





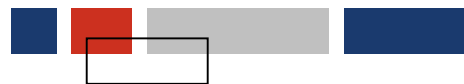
# Data and Software



## » Farm Structure Survey: goals

- To know the structure of agricultural holdings;
- To allow analysis of the evolution of agricultural production systems;
- To characterize the family farm population and the salaried labor force;
- To provide information on the origin of the producer's income;
- To provide a set of information related to rural development;
- To know some cultural practices



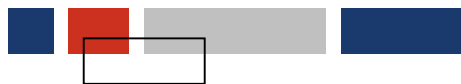


# Data and Software

## » Farm Structure Survey



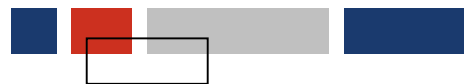
Population has been divided in 765 strata, ( $h=1, \dots, 765$ ) and 23 domains or small areas, corresponding to NUTS III, ( $d=1, \dots, 23$ ). The overall population size ( $N$ ) is 236696 agricultural holdings and the sample size ( $n$ ) is 23108, representing about 9,76 % of the population



# Data and Software

## » Farm Structure Survey



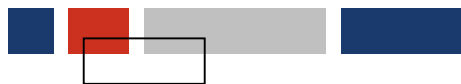


# Data and Software



## » Farm Structure Survey

- Algorithms to calculate the estimates, with the exception of the EBLUP estimator, were all programmed in R by the authors.
- The SEBLUP algorithm was obtained through the eblupSFH function of the R package sae (Molina Marhuenda, 2013)
- In order to measure and compare the quality of the estimators, the coefficients of variation (CV) are computed and shown in percentage.
- To see if the spatial information introduced by the SEBLUP provided some improvement in the CV estimates, in the analysis of the results we also consider the results of the EBLUP estimator computed through the Fay-Herriot method (EBLUPFH).



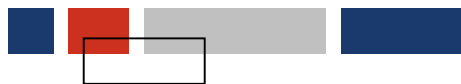
# Results



## » Comparison of Estimators

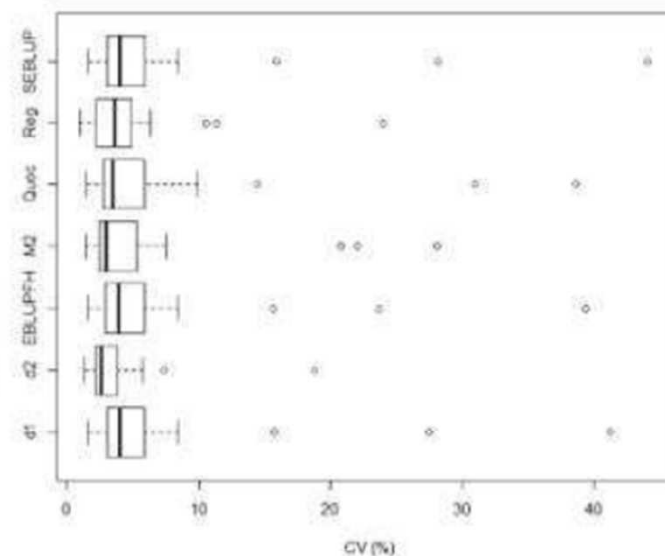
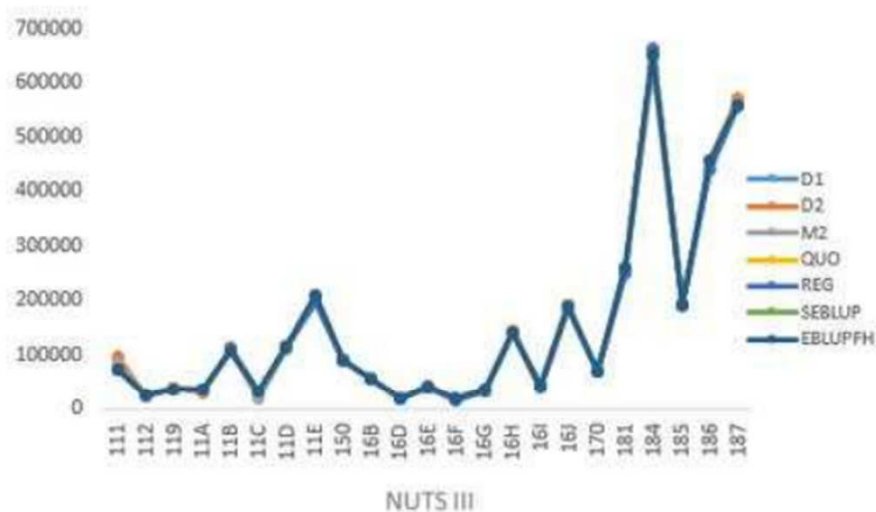
Estimator	CV range (%)	1st Quartile	Median	Mean	3rd Quartile	Quartile
$\hat{\tau}_{D1}(Direct_1 or D_1)$	1.63-41.21	2.99	3.99	7.14	5.83	9.32
$\hat{\tau}_{D2}(Direct_2 or D_2)$	1.29-18-82	2.12	2.57	3.72	3.84	3.61
$\hat{\tau}_{d,reg}(Reg)$	0.93-24.00	2.23	3.64	4.87	4.88	4.93
$\hat{\theta}_{SEBLUP}$	1.64-44.09	3.04	3.99	7.33	5.89	9.86
$\hat{\theta}_{EBLUP_{FH}}$	1.63-39.37	2.86	3.93	6.83	5.84	8.66





# Results

## » Comparison of Estimators



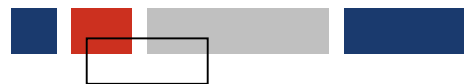


# Conclusions



## » Comparison of Estimators

- Mod. Reg, SEBLUP and EBLUP present greater gains in precision when the sample size is larger and when the correlation between the dependent and independent variables is greater.
- When analyzing the CV estimates of the different estimators studied by NUTS III for one of the most important variables, UAA (Utilized Agricultural Area), the regions of Baixo Alentejo (184) and Alentejo Central (187) are the ones with the highest CV values when compared with those of the other NUTS III regions.



# Conclusions

## » Comparison of Estimators



- This result ends up harming the interpretation of the mean CV values of the estimators, since in general the CV estimates for the other regions are much lower.