Small Area Estimation for Land Use and Land Cover

Pedro Campos[1], Suelma Pina[2], and A. Manuela Gonçalves[2]

[1]Statistics Portugal, Portugal
[2]Universidade do Minho, Portugal

E-mail for correspondence: pedro.campos@ine.pt

Abstract: Small Area Estimation (SAE) is a part of statistical science that combines survey sampling and inference of finite populations with statistical modeling. The main objective of this paper is to analyze and test the implementation of different types of estimators of small domains in order to improve the quality of the estimates produced within the framework of the Farm Structure Survey (FSS) at NUTS III level. Under the EUROSTAT Land Use and Cover Area Statistical Survey (LUCAS) project, this is a fundamental tool for environmental studies, forestry and agricultural resource planning.

Keywords: Small Area Estimation; Regression Estimator; EBLUP; SEBLUP; Farm Structure Survey

1. Introduction

Nowadays, public and private institutions are increasingly seeking more detailed information to aid their decision-making process, and the National Statistical Offices do fall into this new paradigm. 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 (that use only the observations of the variable of interest belonging to the domain for the time period under analysis), because often there are no samples for these domains, or they are too small to obtain sufficient quality estimates. In order to solve this problem, several types of estimators for small domains have been proposed: some of them combine the auxiliary information of the variable of interest of the domain of study in different periods of time, or even consider variable sources of other domains (the so-called indirect estimators). The main objective of this project is to develop, analyze and test the implementation of different types of small area estimators in order to improve the quality of the estimates produced within the framework of the Farm Structure Survey (FSS) at regional (NUTS III) level. Currently, Statistics Portugal publishes these estimates at National (NUTS I) and Regional (NUTS II) levels. Under the EUROSTAT Land Use and Cover Area Statistical Survey (LUCAS) project, Statistics Portugal intends to use this information to detail the agriculture class, thus providing information on agricultural land use up to the third level of patent nomenclature in the Land Use and Land Cover mapping (LULC), a fundamental tool for environmental studies, forestry and agricultural resource planning (EUROSTAT, 2013).

In this work, five different estimators (direct, modified and combined) are used to estimate 44 variables by NUTS III in mainland Portugal: the direct estimator (1 and 2), the estimator modified by the Regression, the EBLUP estimator using the Fay-Herriot method and the EBLUP estimator by the spatial level of the area (SEBLUP). Based on the results, we may conclude that when auxiliary variables are available, the estimator modified by the Regression performs better when compared to other estimators.

2. Small Area Estimators

In this section we introduce Small Area Estimation (SAE) and shortly describe the main estimators used in this work. In a stratified random sampling design, let U be a finite population of N distinct elements, U = {1, .., N}, the subpopulations (in this case, strata), U$_h$, U$_h$ ⊂U, h = 1, ..., H, for which certain parameters have to be estimated according to the domain d (see Figure 1).

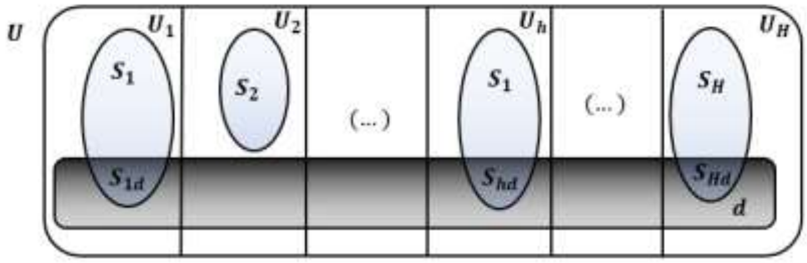

Figure 1: Representation of domains, under the SAE perspective

The population dimension of each stratum U_h is denoted by N_h, $h = 1, \dots, H$, where $N = \sum_{h=1}^{H} N_h$, and the subpopulation dimension U_{hd} is denoted by N_{hd}, where $N_d = \sum_{h=1}^{H} N_{hd}$; we consider s as a sample of size n collected from U that may be decomposed in s $= \bigcup_{h=1}^{H} s_h$ and $s_d = \bigcup_{h=1}^{H} s_{hd}$, which are sampling units of size n_d and n_{hd} randomly selected, where $n = \sum_{h=1}^{H} n_h$ and $n_d = \sum_{h=1}^{H} n_{hd}$.

We usually denote population U as being composed by two quantities. Y (the explained variable, or variable of interest) and $X = (X_1, \dots, X_j) \epsilon \mathbb{R}^j$, the values of the covariates or auxiliary variables. Auxiliary variables are always assumed to be known, whereas the variable of interest may be unknown for some areas if individuals in these areas were not sampled. Assuming that we want to obtain estimates of the total, τ$_d$, the total of the variable of interest for the population of the domain of interest d is given by: $\tau_d = \sum_{i \in U_d} y_i$.

In general, 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, whereas indirect estimators take observations of the variable of interest as well as auxiliary sources outside the study domain for the considered period of time. The Model-based approach belongs to the class of indirect estimators and regression models are used here between data from the sample and auxiliary variables from other data sources, such as census and administrative records, to "lend" information from similar areas (Rao and Molina, 2015). Indirect estimators can also be divided in synthetic and combined estimators, which can be derived under a design-based approach or taking into account the fact that an explicit area level or unit level model exists. Combined estimators are basically weighted averages of a direct estimator and an indirect estimator (Rao and Molina, 2015, Pfeffermann, 2013)

2.1. Direct Estimators (D1 and D2)

We start with the fundamental Horvitz-Thompson estimator, defined in Rao and Molina (2015):

$$D1 = \hat{\tau}_{d1} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{hd}} y_i$$

$$Var(\hat{\tau}_{d1}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(s_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right) \bar{y}_{hd}^2 \right).$$

A second estimator is used, where we assume to know the dimension of each population defined by the intersection of NUTS III with the strata defined a priori in the sampling plan (N_{hd} e n_{hd}):

$$D2 = \hat{\tau}_{d2} = \sum_{h=1}^{H} \frac{N_{hd}}{n_{hd}} \sum_{i \in s_{hd}} y_i,$$

$$Var(\hat{\tau}_{d2}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} s_{hd}^2$$

Where s_{hd}^2 is the sampling variance in the subsample defined by the intersection of stratum h with domain d.

2.2 Direct Estimator modified by Regression (Reg)

For the application of this estimator it is necessary to know the values of the auxiliary variables for all units of the population at individual level, the vector of the totals of the auxiliary variables in domain $\boldsymbol{\tau}_{xd}$ and their observed values in the sample units of the subpopulation g, $\boldsymbol{x}_i$, $i \in s_g$. The regression estimator for the total estimate is given by

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

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

2.3. EBLUP and SEBLUP

The EBLUP is a combined estimator. Considering a finite population divided into small D domains, the Fay-Herriot base model (Rao and Molina, 2015) linearly relates the value of the d-th domain of the variable of interest $\boldsymbol{\theta}_d$ to a vector of p auxiliary variables aggregated at the $\boldsymbol{x}_d$ area level and includes an associated random $\boldsymbol{v}_d$effect. The model is given by $\boldsymbol{\theta}_d = \boldsymbol{x}_d' \boldsymbol{\beta} + \boldsymbol{v}_d$, $d = 1, \dots, D$, where $\boldsymbol{\beta}$ is a vector of regression parameters, $\boldsymbol{v}_d$ are the random effects. Then, the combined estimator SEBLUP, $\hat{\theta}_{SEBLUP}$, of parameter $\boldsymbol{\theta}_d$ may be written as

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

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 the proximity matrix (Pfeffermann, 2013).

3. Data, Software, and Results

3.1. Data and Software

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).

Thereforeyears the FSS is carried out as a sample survey, and once in ten Therefore, the population has been divided into 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. 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 and 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 (FH-EBLUP).

3.2. Results

Results of the coefficient of variation (CV) of the five estimators are presented in Table 1.

Table 1 – Results of the coefficient of variation (CV) of the five estimators

Estimator	CV Range (%)	1st Quartile	Median	Mean	3rd Quartile	Stand. Dev.
$\hat{\boldsymbol{\tau}}_{\boldsymbol{d1}}$ (Direct 1 or D1)	1.63 – 41.21	2.99	3.99	7.14	5.83	9.32
$\hat{\boldsymbol{\tau}}_{\boldsymbol{d2}}$ (Direct 2 or D2)	1.29 – 18.82	2.12	2.57	3.72	3.84	3.61
$\hat{t}_{d,Reg}$ (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
$\hat{\theta}_{EBLUPFH}$	1.63– 39.37	2.86	3.93	6.83	5.84	8.66

The wide variation of the CV range is due to the fact that different small areas (the NUTS III regions) differ much in terms of sample sizes. We can see (Figure 2) that lowest values of CV were provided by Reg (the Direct Estimator modified by Regression), although Direct 2 (Direct Estimator 2) also performed well.

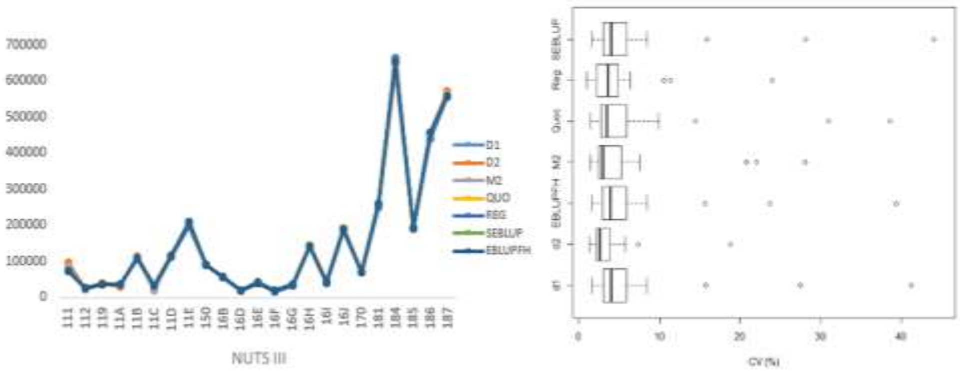

Figure 2 – Graphical comparison of the estimates and boxplots of CV for the five estimators under analysis. (Note: we introduced two extra estimators: M2, the modified estimator and Quo, the Quotient estimator).

4. Conclusions

With regard to modified and indirect estimators Reg, SEBLUP and EBLUP, we found out that they 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 Central Alentejo (187) are the ones with the highest CV values when compared with those of the other NUTS III regions. 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.

References

EUROSTAT (2005). LUCAS 2009 (Lande Use/Cover Area Frame Survey), *Quality report*. Luxembourg: Eurostat.

Instituto Nacional de Estatística (INE) (2013). Inquérito à estrutura das explorações agrícolas, *Documento Metodológico*. Lisboa: INE.

Rao, J.N.K. (2003). Small Area Estimation. Hoboken. New Jersey: John Wiley & Sons, Inc.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, 28(1), 40 – 68.

Rao, J.N.K., Molina, I. (2015). Small Area Estimation. 2nd Edition *Wiley Series in Survey Methodology*. John Wiley & Sons, Inc., Hoboken, New Jersey.