

# Small Area Estimation, an Introduction

Stefano Marchetti

2020

## 1 Introduction to the course

This is a short module on a general overview of small area estimation methods. The aim is to provide the basic instruments to understand how small area estimation works. The flow of the course starts from the definition of small area (or small domain), the problem of estimation in such areas or domains and the possible solutions.

Examples are presented in the course to show the potentiality of the method. Moreover, the course is completed with two practical exercises on the R software.

The following comments can support the lecture. Each paragraph is enumerated and it corresponds to a slide. Some slides are self-explained and there is no comment.

## 2 Outline of the course

(3) The course is organised as follows. We start by introducing the small area estimation problem and the definition of a small area. After some definitions, we present small area estimation under the classical inference approach: the expansion estimator, the generalized regression estimator, the synthetic estimator and the composite estimator. Next, we present the model-based approach to small area estimation that is mainly classified into area level approach and unit level approach. Finally, we finish with ongoing research topics.

## 3 Introduction to Small Area Estimation

(5) National statistical offices and other public and private institutions are asked to provide sound estimates. The cost-effective solution is to use sample surveys to obtain information on a wide range of topics of interest.

(6) Usually, a survey is designed to infer on a target population (population of interest). Therefore, estimators based on sample observations - direct estimators - should be reliable for the target population. Often the

population can be divided into areas or domains, which are sub-set of the population that can be planned or not in the survey design.

(7) Usually, direct estimators for unplanned domains or areas are not reliable. To obtain reliable estimates for such areas we can increase the number of observations (oversampling), or apply statistical methods that allow for reliable estimates in that areas. This problem is faced by the small area methods. The definition of small area or small domain is a geographical area or domain where direct estimators do not reach a minimum level of precision. A small area estimator in an estimator created to obtain sound estimates in a small area.

(8) Here an example of small area based on Italian data. Let the household who live in an Italian region be the target population, and let the income be the variable of interest. In Italy, income information are collected by the EU-SILC survey that is designed to obtain reliable estimate at the regional level (NUTS 2). The planned domains (areas) are the regions, while smaller subdivisions are unplanned domains, like provinces (NUTS 3/ LAU 1) and municipalities (LAU 2). For example, the EUSILC sample size in Tuscany (an Italian region) in 2008 was 1751 households, usually enough to obtain sound direct estimates. If we focus on provinces, the sample size is smaller. For example in the province of Pisa there was 158 sampled households and 70 in the province of Grosseto. Usually direct estimation is not reliable with such sample sizes (very large confidence intervals).

(9) Another example, from the book of Rao (2003), is about the sample size at the state level in the USA for an equal probability sample of 10000 persons. We can see that we have large sample size in California, Texas and so on, and small sample size in state like D.C. and Wyoming. In California the sample size was 1207 persons, while in Wyoming was 18 persons. Suppose to measure the customer satisfaction of a government service and obtain a fraction of satisfied persons equal to 0.248 and 0.333 in California and Wyoming respectively. A 95% confidence interval for the satisfied in California is 22.4% – 27.3%, and it can be judged reliable, while for Wyoming we have an interval 10.9% – 55.7% that is judged unreliable.

## 4 Definitions

(10) – (12) Some definitions about basic concepts in survey methodology are provided here. Slides 10 to 12 don't need further comments.

## 5 Design-based estimators

(13) The most used design-based estimators are the expansion estimator and the general regression estimator. In this course we call the expansion estimator “Direct Estimators”. The direct estimator for the mean of a finite

population use inclusion probability to "expand" the sample to the entire population. To be clear in the small area literature a direct estimator is an estimator based only on specific area data.

(15) After partitioning a finite population into  $m$  areas (or domains), the direct estimator of the mean of area  $i$  is defined. We can show that under simple random sampling the direct estimator is the sample mean of the target in area  $i$ .

(16) The direct estimator is design unbiased and in the sample case of the simple random sampling its variance depends on sampling fraction ( $n_i/N_i$ ), sampling variance ( $S_i^2$ ) and sample size ( $n_i$ ). When in an area the sample size is small the variance of the direct estimator is likely to be very large. Remark: the direct estimator showed here is known as expansion estimator (as already said), however, in the SAE literature direct estimators are those estimators based only on the area specific observations of the target variable (and inclusion probabilities).

## 6 Synthetic Estimators

(17) If the small areas have the same characteristic as the large areas with respect to the target variable then we can define a synthetic estimator, which as several advantages: it is simple, it applies to general sampling designs, it borrows strength from similar and it provides estimates for areas with no sample from the sample survey that are usually defined as out of sample areas.

(18) Here we show that the simplest synthetic estimator we can define is a simple constant regression model that corresponds to the expansion estimator of the mean.

(19) Using a set of auxiliary variables available at unit level, we can build a synthetic estimator that make use of all the sample to estimate the  $\beta$  regression coefficients and then obtain an area  $i$  estimator using area related covariates.

(20) The synthetic estimator with auxiliary variables is biased, however the bias can be small if the area specific regression coefficients  $\beta_i$  are the same as the coefficients  $\beta$  (that is a vector of  $p$  coefficients, with  $p$  the number of auxiliary variables).

## 7 Composite Estimator

(21) A composite estimator is a weighted mean between a direct and a synthetic estimator, with weight  $\phi_i$  between 0 and 1.

(22) The basic idea of a composite estimator is to find a weight  $\phi_i^*$  that minimize the estimator mean squared error. The optimal weight is the ratio between the MSE of the synthetic estimator and the sum of MSE of synthetic

estimator and variance of direct estimator. Unfortunately, the estimate of  $\phi_i$  is very unstable, and particularly when area sample size are small.

(23) As an alternative, a common weight for all areas can be obtained minimizing the average of the MSE of the composite estimator among the areas. By this way it is possible to estimate the common weight at the price of a lower efficiency with respect to have an area specific weights.

(24) The estimator presented are those of the classical inferential statistics. It can be of interest to show a comparison between them using a simulation experiment. We use the 1981 Italian Population Census, and we draw repeatedly a two-stage stratified sample. The design and the sample size are those of the Italian Labour Force Survey. The small areas are 14.

(25) The performance of the estimators are evaluated using the average relative bias and the relative root MSE.

(26) In terms of bias the direct estimator is the best, the synthetic is the worst and the composite is in the middle. This is an expected result, because the direct estimator (i.e. expansion estimator) is design unbiased, while the synthetic estimator is biased. The composite estimator is a weighted average of the two, then it is expected to be less biased than the synthetic estimator. In terms of relative root MSE the composite is the best, the synthetic performs similarly and the direct estimator is the most variable. Paying a small price on the bias the composite estimator obtained a great gain in efficiency (i.e. reduction of variability).

## 8 Model-based Approach to Small Area Estimation

(28) We have seen that composite estimators can improve the efficiency respect to a direct estimator, at the price of some biases, moreover, the optimal weight is not area-specific. By introducing small area predictors based on small area models we can obtain best predictors.

(29) The family of small area models can be divided into two groups: unit-level models and area-level models. The first uses unit-level survey data (also called micro-data) and area-level (or aggregate) covariates for estimating the average, while to estimate other target statistics, such as quantiles, inequality indexes, etc., there is need of unit-level covariates for all the population units. The latter uses only area-level (aggregate) data for model fit and for estimating target statistics. Data availability plays a crucial role in the choice of the models. We will start by the basic area-level models, or Fay and Herriot (1979) model.

## 8.1 Area level models

(32) If the above mentioned assumptions are reasonable, then  $\hat{\theta}_i$  (the direct estimator of area  $i$ ) is normally distributed with mean  $x_i\beta$  and variance  $z_i^2\sigma_u^2 + \psi_i$ , where  $\sigma_u^2$  is the variance of the area random effects  $u_i$  and  $\psi_i$  is the variance of the direct estimator. The matrix notation is introduced because it is compact and can be easily implemented on computers. Assuming the variance matrix  $G$  and  $R$  are known we can define the Best Linear Unbiased Predictor (BLUP) for the average of area  $i$ .

(33) Of course, matrix  $G$  and  $R$  are unknown in real applications and must be estimated. There is a variety of methods to obtain such estimates, here we show that of Prasad and Rao (1990) that use the restricted maximum likelihood (REML) to estimate  $G$  - i.e.  $\sigma_u^2$ . Indeed, in the Fay-Herriot model  $\psi_i$  is considered known (matrix  $R$  known), but a smoothed estimator of the variance of the direct estimator is used in real applications. In literature there are works that account for the uncertainty due to the estimation of the variance of the direct estimator.

(34) Plugging-in the estimated  $G$  and  $R$  matrix in the estimator of  $\beta$  and  $u$  we obtain the Empirical BLUP (EBLUP),  $\hat{\theta}_i^{FH}$ , that can be written as a composite estimator of direct estimator and synthetic estimator, with area-specific weights  $\hat{\phi}_i$ . The weights are the ratio between the estimated variance of the random area effect  $\hat{\sigma}_u^2$  and the sum of  $\hat{\sigma}_u^2$  and the variance of the direct estimator  $\psi_i$ . If the covariates have a good prediction power than the variance of the random area effects is small, and then  $\hat{\phi}_i$  is small and more weight is done to the synthetic predictor. Viceversa, if the covariates are not good predictors then  $\hat{\sigma}_u^2$  is large and  $\hat{\phi}_i$  is big (tends to be close to one), then more weight is given to the direct, and this is ok because the synthetic predictor in such a case is not good.

(35) The derivation of the MSE of the EBLUP is not immediate. Here, we show only its definition and one possible estimator. The MSE of the EBLUP can be decomposed into three components, which depends on  $\sigma_u^2$ . The MSE is leaded by the first component, while the second and third components are of smaller order. The first term is equal  $\phi_i\psi_i$ , then the weights  $\phi_i$  is how much the MSE of EBLUP shrink the variance of the direct estimator, which is  $\psi_i$ . Indeed, if the covariates has a good prediction power we have seen that  $\phi_i$  is small. In this case we have a great reduction of the variability of the EBLUP with respect to the direct. On the contrary, if the covariates have not prediction power then  $\phi_i$  is big (say close to one) and the MSE of the EBLUP is then more or less equal to the variance of the direct estimator ( $\psi_i$ ). Different analytical forms of the MSE of the EBLUP exists in literature. The estimation of the MSE of the EBUP is not straightforward. An approximately unbiased estimator is given by the sum of the estimated first, second and twice the third components. Details can be found in Molina and Rao (2015). Alternative estimators based on

bootstrap and jackknife are available.

(36) In what follow we show an application of the basic area level model to Italian data. Our goal is to obtain reliable estimates of the mean consumption expenditure at provincial level (LAU 1/NUTS 3) in Italy, taking into account that consumption expenditure data come from a survey designed to ensure reliable estimates at regional level (NUTS 2).

(37) In the application we use data from the household budget survey (2012) from which we get the direct estimates of the mean consumption expenditure at provincial level, data from the Italian Revenue Authority from which we compute the per capita taxable income in 2012, obtained as the ratio between the total income tax in the province and the population size in the province. Finally, we use also the percentage of households who have the ownership of the house where they usually live, available from the Population Census 2011 at provincial level. Data about ownership of the house are one year before HBS and tax data, however, this is not a problem because ownership is stable over time and we expect an irrelevant change from one year to another.

(38) In Italy there are 110 provinces, and we will use the Fay-Herriot model to obtain the EBLUP of the mean equivalised consumption expenditure. Consumption expenditure has to be equivalised to account for the economy of scale present at the different dimension of the households. We use the OECD scale as equivalence scale. Once we equivalised the consumption expenditure we obtain the direct estimate of the mean equivalised consumption expenditure at provincial level and also the estimate of its variance. The direct estimates are obtained using the expansion estimator, where survey weights have been calibrated to population total at provincial level. Estimation of the variance of the direct estimates has been obtained assuming simple random sampling within the provinces since the effect of the design is unknown at the province level.

(39) As expected, since the sample size at provincial level is small, direct estimates are unreliable. Indeed it ranges between 4 and 1037, with quartiles equal to 85, 147 and 303 sampled households. The Fay-Herriot model uses the direct estimates of the equivalised mean consumption expenditure as response variable and per capita taxable income and home ownership as auxiliary variables.

(40) Here, the model parameters estimates. First of all we have to say that we are not interesting in interpreting the effects of auxiliary variables on the response variable. We are interested only in the prediction power of the model, in particular we want to get the estimate of  $\sigma_u^2$  as small as possible. However, we also have to check if the model coefficients make sense and if the model assumptions are reasonable. Higher per capita taxable income in the province and higher fraction of home ownership increase the mean equiv consumption expenditures, and that's ok. Moreover, coefficients are significantly different from zero. We test the normality assumption of

the area random effects using the Shapiro-Wilk normality test which reject the null hypothesis. However, estimated random effects are symmetric (not shown here) and Fay-Herriot model is robust against departure from normality assumptions (Lahiri and Rao, 1995). Very often consumption data are modelled using log transformation. However, we stress that here we are modelling direct estimates - i.e. means - so they are asymptotically normal distributed. Even if the sample size in each area is relatively small, we can expect small departure from normality assumption. Moreover, both raw-scale and log-scale direct estimates show departures from normality according to a Quantile-Quantile plot (which is not showed here).

(41) The distribution of direct estimates and EBLUPs (obtained under the Fay-Herriot model) across the 110 Italian provinces are reported in this table. We can see that the distribution is quite similar, and this is a frequent results because the EBLUPs are driven from the direct estimates, in particular those who have smaller variance (smaller  $\psi_i$ ). However, the range of the EBLUPs is smaller than that of the direct estimates. This is a typical behaviour of the EBLUPs obtained under the Fay-Herriot model. We usually say that EBLUPs shrink the direct estimates. This behaviour is not always negative, since extreme values of direct estimates may be unrealistic and due only to the small sample size. Consider as another example the estimation of the poverty rate (ratio of poor households in an area). It may happen that in some small areas the poverty rate is zero because of the small sample size (e.g. 5 or 10 households). Nevertheless, it is unrealistic to think that in that areas there are no poor households. In such a case the EBLUPs won't be zero giving a more realistic prediction of the poverty rate.

(42) When small areas refer to geographical territories it is useful to map the results. In the figure the direct estimates and the EBLUPs at provincial level in Italy are mapped. The two maps are quite similar, however, there are some differences, e.g. the province of Roma (RM, in the middle of the "boot"), the province of Belluno (BL, in the north-east) and others.

(43-44) The most important result in a small area application is to increase the efficiency of the small area estimates (the EBLUPs) with respect to the direct estimates. This is the reason why we use the EBLUP. One way to compare the efficiency of the EBLUP is to compute for each area the ratio between the estimated root MSE of the EBLUP and that of the direct estimator (which correspond to the estimated standard error of the direct estimates). A Ratio smaller than one means an increase in the efficiency. In our application we obtained an increased efficiency of the EBLUP in all the provinces: the ratio  $rmse(EBLUP)/rmse(Direct)$  ranges from 10% to 97%. A ratio of 10% means that in a given province the estimated root MSE of the EBLUP is *one tenth* (1/10) of the estimated standard error of the direct (in the same province), this results in an increased efficiency of  $1 - 0.1 = 0.9 = 90\%$ . Across the 110 Italian provinces the variability of the EBLUP is on average 36% (100-64) smaller that that of the direct. Looking

at the distribution across the provinces of the Coefficient of Variation (CV), we can see that all the EBLUPs can be considered reliable - CV lesser than about 10% - while a lot of direct estimates aren't - 55 provinces with a CV bigger than 10%.

(45) We now show one of the possible extension to the basic area level model (or Fay-Herriot model). When there is spatial dependency between data some assumptions of the basic area-level model are violated, then there is need to account for that spatial correlation/dependency. Moreover, if accounted spatial dependency may improve the efficiency of small area estimators. Other extensions are available in literature.

(46) The spatial effects are incorporated in the random area effect  $v$ , which follow a Simultaneously Autoregressive Process (SAR).

(47) An EBLUP for the spatial area level model can be obtained using REML or ML, we refer to this as Spatial EBLUP (SEBLUP). A second order approximation of the MSE of the SEBLUP has been proposed by Singh et al. (2005) and Petrucci and Salvati (2006). Alternatives make use of bootstrap techniques.

(48) Other approaches are discussed in the small area literature to account for the spatial dependency. Chandra et. al (2015) discussed the problem of non stationary spatial dependency, and Giusti et al. (2012) use p-spline to incorporate spatial information in the area-level model, while Opsomer et al. (2008) did the same for the unit level models.

## 8.2 Unit-level models

(49) Under the basic unit level approach the target variable is modelled using a random intercept model. Random intercept accounts for the between area variation that is not explained by the auxiliary variables. Model predictions for the non sampled units of the population are used to obtain an EBLUP for small areas.

(50) Under these assumptions the target variable is normally distributed with (overall) mean  $X'\beta$  and variance matrix  $V = ZGZ' + R$ . The BLUE of  $\beta$  and the *BLUP* of  $u$  depends on the parameters  $\sigma_u^2$  and  $\sigma_e^2$  that are unknown and have to be estimated. Usually, they can be estimated using REML, see Rao and Molina (2015) for further details. Plugging-in the estimates  $\hat{\sigma}_u^2$  and  $\hat{\sigma}_e^2$  into the BLUE of  $\beta$  and the BLUP of  $u$  we obtain the empirical estimates  $\hat{\beta}$  and  $\hat{u}$ , and then the EBLUP.

(51) The basic idea of the unit level approach is to divide the population into sampled (set  $s$ ) and non-sampled (set  $r$ ) units. Then, the vector of the target variable can be split into sampled and non-sampled units, here denoted by  $y_s$  and  $y_r$  respectively. Statistics of interest can be defined as a linear combination between  $y_s$  and  $y_r$ . Under the random intercept model the predicted value for the target variable is  $\hat{y}_r = X_r'\hat{\beta} + \hat{u}$ . Predicted values can be plugged into the linear combination to estimate the statistics

of interest, usually means or totals.

(52) Let see for example the estimation of the small area mean  $\bar{Y}_i$ . This statistic of interest is simply defined as the sum of the target variables values in area  $i$  - i.e.  $\sum_{j=1}^{N_i} y_{ij}$  - divided by the population size in area  $i$  - i.e.  $N_i$ . As said in the previous slide, we divide the numerator of the small area  $i$  mean -  $\bar{Y}_i$  - between the sum of the sampled units in area  $i$  -i.e.  $\sum_{j \in s_i} y_{ij}$  - and the sum of the non sampled units in area  $i$  - i.e.  $\sum_{k \in r_i} y_{ik}$ . The latter sum is unknown and can be estimated using the predicted values under the random intercept model,  $\hat{y}_{ik} = x'_{ik}\hat{\beta} + \hat{u}_i, k \in r_i$ . Then, an estimator of the mean of area  $i$  is  $\hat{Y}_i$ , which is an EBLUP (Empirical Best Linear Unbiased Predictor). Looking at the expression of  $\hat{Y}_i$  we can note that there is need to know the auxiliary variables  $x_{ik}$  for all the units in the non sampled part of the population  $k \in r_i, i = 1, \dots, m$ . Having these data is usually a problem. Nevertheless,  $\hat{Y}_i$  can be easily obtained using the small area mean of the auxiliary variables - i.e.  $\sum_{k \in r_i} x'_{ik}\hat{\beta} + \hat{u}_i = (N_i\bar{X}_i - n_i\bar{x}_i)\hat{\beta} + (N_i - n_i)\hat{u}_i$ , where  $n_i$  is the sample size in area  $i$ ,  $\bar{X}_i$  and  $\bar{x}_i$  are the vector of population means and sample means for the auxiliary variables in area  $i$ . We can also show that the EBLUP can be rewritten as a composite estimator with area-specific weight  $\phi_i$ .

(53-54) The derivation of an analytic form of the MSE of the unit-level EBLUP is quite complicated as it is its estimator. Here we show the three main components of the MSE of the EBLUP of  $\hat{\theta}_i$  (e.g.  $\hat{\theta}_i = \hat{Y}_i$ ) obtained using a Taylor linearisation. Details can be found in the Rao and Molina (2015) book.

(55) In what follows we show an application of the unit-level small area model to get reliable estimates of the mean of the equivalised income at the provincial level in Tuscany. In Italy data on households income are collected using the EU-SILC survey, which is designed to obtain sound estimates at the regional level (NUTS 2). Usually, within regions there are differences in the level of mean income and a picture of the mean income at provincial level (NUTS 3/LAU 1) can help policy makers to get data driven decisions. A set of auxiliary variables is available at unit-level data (micro data) from the Italian Population Census 2001. Income data refers to year 2005 and come from the EU-SILC survey 2006.

(56) Using previous studies we select these variables from the Census: 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 of the census survey). Design-based estimates at provincial level of the mean equivalised income have been obtained using the expansion estimator as well as estimates of their standard errors. These estimates are used as benchmark to make comparison with the small area estimates.

(57) In Tuscany there are 10 provinces, however, the province of Florence

has been divided into two areas, the municipality of Florence (Florence M.) and the rest of the province (Florence). The estimated root MSE of the EBLUP is smaller than the estimated standard error of the expansion estimator (design-based estimator) for the provinces of Arezzo, Florence M., Grosseto, Livorno, Massa, Pistoia and Prato, while in the provinces of Florence, Lucca, Pisa and Siena the expansion estimator perform better than the EBLUP.

(58) Even if point estimates between the two estimator are different in this figure we can see that 95% confidence interval overlap for all the areas. The red interval refers to the EBLUP and the black to the expansion estimator.

(59) Mapping the results can help to understand the phenomena and also to make a comparison between estimators. Using quartiles of the expansion estimator we draw a choropleth map, where we can see that EBLUPs shrink the mean income, in particular in the province of Livorno and Grosseto, the ones who show the biggest estimated standard error of the expansion estimator.

(60) A lot of literature is available on unit-level small area model to improve the efficiency of the EBLUP in different situations: include spatial information in the unit level model, account for time series data, account both time and spatial information, handle outliers (robust estimation), define models for binary and count target variable. Other extension are available. Moreover, different approaches exists in literature, such as the M-quantile (see Chambers and Tzavidis, 2006) and the Bayesian approach to small area. There are also contributions about the benchmarking problem of the EBLUP (and many other model-based estimators) - i.e. mean of the EBLUPs doesn't match with the design-based population mean estimate.

## 9 Concluding Remarks

(64) Giving the importance to get best from available data, the on going research on small area estimation is wide. It includes robustness, estimation of small area quantiles, definition of new models (Bayesian, non-parametric and semi-parametric), model-based estimator design-unbiased, model for non continuous data and in the last year also how to use Big Data in small area models. Also many applications of small area methods can be found in literature, and more and more statistical offices produce small area estimates (e.g. SAIPE program of the US Bureau of Statistics).