Small Area Estimation

Angelo Moretti

Department of Computing and Mathematics

Manchester Metropolitan University, UK

A.Moretti@mmu.ac.uk

The Small Area Estimation Problem

Many social phenomena such as poverty, well-being and social exclusion are spatially heterogeneous; therefore, policy makers are interested in small area estimates of those. Unfortunately, most large-scale social sample surveys are designed to provide reliable estimates of population target parameters (e.g. disposable income) at a national level or for large areas. According to Rao and Molina (2015), an area is defined as 'small', if the area is an unplanned domain and the specific sample size may not be large enough to provide reliable direct estimates. Interestingly, small areas can be also defined by the cross-classification of geographical areas by social, economic or demographic characteristics.

In this context, direct estimators, such as the well-known Horvitz-Thompson estimator (Horvitz and Thompson, 1952) provide large variability in the estimates and in small areas

with zero sample size, they cannot be computed. In this case, indirect model-based estimation methods, in particular small area estimation (SAE) approaches can be used to estimate population target parameters (e.g. means, totals, ratios) at small area level (Rao and Molina, 2015).

SAE is defined as a set of statistical techniques with the aim of producing accurate and precise estimates for small areas, as well as for domains with zero sample size. SAE approaches 'borrow strength' from related small areas through the use of auxiliary variables available at population level from the Census or administrative data.

SAE under the indirect model-based approach can be classified into two approaches: unit-level and area-level approach. The unit-level approach is used when auxiliary variables are available for each observed unit, while the area-level approach is used when auxiliary variables are known only at the area-level (e.g. Census means or totals for each area). This work is organised as follows. First, the direct estimation approach and the notation used are described. Second, the traditional SAE indirect model-based approaches are presented. Third, the importance of model diagnostics is discussed. Fourth, some applications of SAE relevant to the Social Sciences are presented and discussed. This work concludes with some final remarks.

Direct Estimation

It is assumed that a finite target population Ω, e.g. the population of Italy, of size N is partitioned into D non-overlapping small area areas $d = 1, \dots, D$. These can be for example the Italian provinces. A random sample $s \subset \Omega$ of size n is drawn from the population. $N - n$ are the non-sampled units and these are denoted by r; thus, $s_d = s \cap \Omega_d$ is the sub-sample from the small area d of size n_d, $n = \sum_{d=1}^{D} n_d$, and $s = \cup_d s_d$. r_d denotes the non-sampled units for small area d of $N_d - n_d$ dimension.

In this work, the target parameter (e.g. the equivalized disposable income) is the population mean $\bar{Y}_d$ of a variable Y for area d (e.g. the mean of the equivalized disposable income for province d). This needs to be estimated and is given as follows:

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}, \tag{1}$$

Where Y_{di} denotes the value of variable Y for unit i in area d.

A method to produce estimates of (1) for $d = 1, \dots, D$ is the direct estimator, which uses the sample information only. This is given by:

$$\hat{\bar{Y}}_d^{Direct} = \frac{\sum_{i \in s_d} y_{di} w_{di}}{\sum_{i \in s_d} w_{di}}, \tag{2}$$

Where w_{di} denotes the design-weight for unit i in area d. There are different types of direct estimators that can be used in survey sampling. In the following section, we consider an

estimator widely used in SAE also in the model-based approach.

We now consider the following example: the population of Italy, is stratified by regions and households are selected, within each region, according to a simple random sampling without replacement design. Statistical inference is further carried out at sub-regional level, e.g. province level. The direct estimates are therefore computed for each province.

Horvitz-Thompson Estimator

An unbiased direct estimator for (1) is the Horvitz–Thompson (HT) estimator developed by Horvitz and Thompson (1952). This can be used to estimate a finite population parameter, e.g. the equivalized disposable income for area d, when a sample is selected with unequal probabilities without replacement.

π_{di} denotes the first-order inclusion probability of unit i from area d in s_d. Thus, $w_{di} = \pi_{di}^{-1}$ is the corresponding sampling weight. . The HT estimator is given by:

$$\hat{\bar{Y}}_d^{HT} = \sum_{i \in s_d} y_{di} w_{di} / N_d \, . \tag{3}$$

The first-order inclusion probability refers to the probability that unit i is included in the sample s_d.

An unbiased estimator of the variance of (3) requires the second order inclusion probabilities denoted by $\pi_{d,ik}$. If it is assumed that $\pi_{d,ik} \approx \pi_{di}\pi_{dk}$, an approximation of the variance of (3) can be written as follows:

$$\hat{V}(\hat{\bar{Y}}_d^{HT}) = \frac{1}{N_d^2} \sum_{i \in s_d} w_{di}(w_{di} - 1) y_{di}^2. \quad (4)$$

The reader may want to refer to Rao and Molina (2015) for more details on variance estimation of small area direct estimators.

Unfortunately, when the small area estimation problem arises, i.e. the sample size in area d is "small" or even zero, estimator (3) may return large variability in the estimates or it cannot be computed in case of zero sample sizes. This problem is also known as "unplanned domains" issue in survey statistics.

Indirect Estimation

Due to the small area estimation problem, auxiliary variables for every small area from the Census, administrative data or other reliable and available data sources can be used to improve the small area estimates. In particular, indirect estimators that borrow strength from related small areas are constructed. This approach is the indirect estimation approach in small area estimation. It is called "indirect", because variables from both the survey sample and auxiliary data e.g. Census are used in the estimation.

Unit-level Approach

Since the population target parameter given in (1) is a linear quantity, it can be decomposed into two components: one related to the sample elements s_d, and one related to the out-of-

sample elements r_d:

$$\bar{Y}_d = N_d^{-1}\left(\sum_{i\in s_d} y_{di} + \sum_{i\in r_d} y_{di}\right), \qquad i = 1, \dots, N_d. \tag{5}$$

Notice that the quantity $\sum_{i\in r_d} y_{di}$ is not observed, since it refers to out-of-sample units; thus, it needs to be *predicted*. In the unit-level approach, it is assumed that auxiliary variables are available for all the units in the sample. For example, if the response variable is the income, auxiliary variables may be variables related to education, gender and labour force.

In order to predict the out-of-sample units, the Battese, Harter and Fuller (BHF) model (Battese, et al., 1988) can be used. For example, if the population size in area d is 2000 households and in the sample 3 households were selected in area d, $2000 - 3 = 1997$ are the out-of-sample units. The quantity related to these needs to be predicted using a model. The model is defined as follows (Battese, et al., 1988):

$$\begin{gathered} y_{di} = \boldsymbol{x}_{di}^T\boldsymbol{\beta} + u_d + e_{di}, \\ u_d \overset{iid}{\sim} N(0, \sigma_u^2), e_{di} \overset{iid}{\sim} N(0, \sigma_e^2)\ u_d \text{ and } e_{di} \text{ are } \textit{independent}, \end{gathered} \tag{6}$$

Where u_d denotes the random effect for area d and e_{di} is the individual error term. u_d takes into account for the between-area variation, whereas e_{di} takes into account for the within-area variation. Note that (6) is a two-level model where unit i is nested in area d. Model parameters can be estimated via Restricted Maximum Likelihood (REML), Maximum Likelihood (ML) or other estimation techniques (see Rao and Molina, 2015). Once model (6) is estimated for $i = 1, \dots, n_d$, the Empirical Best Linear Unbiased Predictor (EBLUP) of (5)

is produced and it is given by:

$$\hat{\bar{Y}}_d^{EBLUP} = N_d^{-1}\left(\sum_{i\in s_d} y_{di} + \sum_{i\in r_d} \boldsymbol{x}_{di}^T\widehat{\boldsymbol{\beta}} + \hat{u}_d\right), \quad (7)$$

Where $\widehat{\boldsymbol{\beta}}$ and $\hat{u}_d$ denote the estimates of $\boldsymbol{\beta}$ and u_d, respectively.

In order to check that (7) is a good estimator for (5), a quality measure needs to be estimated. In particular, the estimator of the Mean Squared Error (MSE) of (7) denoted by $\widehat{MSE}(\hat{\bar{Y}}_d^{EBLUP})$ is usually obtained (Rao and Molina, 2015). The MSE is always positive, and values closer to zero indicate a higher reliability of the small area estimate. More details on this can be found in González-Manteiga et al. (2008) and Rao and Molina (2015).

Area-level approach: Fay-Herriot model

In many applications, auxiliary variables may not be available for the sample elements due to confidentiality and privacy restrictions. For example, data may be released only at area-level e.g. direct estimates at province level. Hence, unit-level models cannot be used in this context. In the area-level approach, it is assumed that $\hat{\bar{Y}}_d^{HT}$ relates to a set of auxiliary variables via the following model (Fay and Herriot, 1979):

$$\hat{\bar{Y}}_d^{HT} = \bar{\boldsymbol{x}}_d^T\boldsymbol{\beta} + b_d + \epsilon_d, \quad (8)$$

$$b_d \overset{iid}{\sim} N(0,\sigma_b^2), \epsilon_d \overset{iid}{\sim} N(0,\psi_d)\ b_d \text{ and } \epsilon_d \text{ are } \textit{independent},$$

Where $\bar{\boldsymbol{x}}_d$ denotes the means of the auxiliary variables for area d, b_d is the area-specific random effect for area d with unknown variance σ_b^2 which needs to be estimated in the

sample data, and ϵ_d is the sampling error with known variances ψ_d. The estimates of these variances are denoted by $\hat{\sigma}_b^2$ and $\hat{\psi}_d$. $\bar{\boldsymbol{x}}_d$ are usually taken from the Census or other reliable data sources. This model is known as Fay-Herriot model in the small area estimation literature (see Fay and Herriot, 1979). Note that model (8) is an area-level model, estimated for $d = 1, \dots, D$ small areas. Model parameters estimates $\boldsymbol{\beta}$ and σ_b^2 are obtained by estimating an area-level model on the sample data.

As it has been discussed for the unit-level case, a small area estimator for $\bar{Y}_d$ needs to be developed under model (8). The EBLUP of $\bar{Y}_d$ under model (8) is now given as follows:

$$\hat{\bar{Y}}_d^{EBLUP,FH} = (1 - \hat{\gamma}_d)\hat{\bar{Y}}_d^{Direct} + \hat{\gamma}_d \bar{\boldsymbol{x}}_d^T \hat{\boldsymbol{\beta}}, \quad (9)$$

Where $\hat{\boldsymbol{\beta}}$ is the estimator for $\boldsymbol{\beta}$ and $\hat{\gamma}_d = \frac{\hat{\psi}_d}{\hat{\sigma}_b^2 + \hat{\psi}_d}$, $0 \leq \hat{\gamma}_d \leq 1$. $\hat{\gamma}_d$ is a very important quantity, also known as "shrinkage factor". $\bar{\boldsymbol{x}}_d^T \hat{\boldsymbol{\beta}}$ is called "synthetic estimator". It can be seen that, when n_d is small and so the direct estimator $\hat{\bar{Y}}_d^{HT}$ becomes unreliable, more weight is attached to the model based component $\bar{\boldsymbol{x}}_d^T \hat{\boldsymbol{\beta}}$. On the contrary**,** when n_d is large, then more weight is attached to $\hat{\bar{Y}}_d^{HT}$. This is a form of composite estimators in SAE where the aim is to balance for the large variability in the direct estimates and the possible bias arising from synthetic estimates due to, for example, issues in model fitting (see also Rao and Molina (2015) for more details).

In order to improve the small area estimates, further extensions of the EBLUP under Fay-

Herriot model are proposed in the literature. In particular, Petrucci and Salvati (2006) propose an EBLUP estimator based on spatially correlated random area effects in Fay-Herriot model. The use of a spatial model is particularly helpful when the auxiliary variables do not fully take into account for the spatial variation in the sample. Thus, introducing a spatial predictor may improve substantially the reliability of the small area estimates (see Petrucci and Salvati, 2006).

The Importance of Model Diagnostics

The indirect small area estimators described in the paragraphs above rely on the model assumptions i.e. they are model-based. Therefore, it is extremely important to perform model diagnostic before producing small area estimates. This is crucial in order to evaluate that the small area estimates and their mean squared errors are not biased. First, since normality is required, it is important to check whether the estimated residuals and random effects follow a Normal distribution. Graphically, the Normal Q-Q plot is very helpful and can be used to check this assumption. Hypothesis tests, such as the Kolmogorov–Smirnov and Shapiro–Wilk tests can also be performed. Furthermore, it can be noted that constant variance of the error term is also assumed (homoscedasticity assumption). This needs also to be evaluated in the model diagnostic stage. Another important diagnostic is the bias diagnostic. Brown, et al (2001) suggest to compare the model-based estimates to the direct estimates. Since the direct estimates are unbiased, the model-based estimates can be plotted against the direct estimates.

This provides a graphical illustration of the bias of the small area estimates obtained under the model. Thus, a regression model between the model-based and direct estimates can be estimated.

Applications in the Social Sciences

In the last decade, there has been a growing attention to the development and application of small area estimation methods for poverty and well-being indicators. Particularly, Pratesi (2016) discusses many issues in poverty data and small area estimation, where a large variety of SAE approaches is discussed and evaluated in detail.

In this section, some applications and methodological advances of small area estimation in the field of poverty, well-being and crime measurement at small area level are presented.

Poverty and Well-being Indicators

An important work related to poverty indicators is proposed by Molina and Rao (2010). They provide a methodology based on the unit-level model described in the previous sections for estimating the non-linear Foster–Greer–Thorbecke poverty indicator. Their method can be extended to other non-linear indicators, which are widely diffused in poverty and well-being measurement. They also provide and evaluate a mean squared error bootstrap algorithm.

Regarding small area estimation of multidimensional well-being indicators, Moretti, Shlomo and Sakshaug (2019) develop multivariate predictors under factor analysis models. In these

paper, an application on economic well-being, in particular related to housing quality multidimensional indicators in Tuscany (an Italian region) are provided using the European Union Statistics for Income and Living Conditions data. In the application, two dimensions are considered and their composite estimates are provided: residential area deprivation and housing material deprivation. They show in the paper that by taking into account for correlations between well-being dimensions, the small area estimates can be improved. They also evaluate the methods in a simulation study where different scenarios are considered and multivariate SAE is compared to univariate SAE.

Crime Indicators

Crime indicators are also heterogeneous at small geographical level. Thus, small area estimates are also asked by policy makers in this field. More recently, Buil-Gil et al (2019) provide reliable model-based small area estimates of worry about crime at regional level from the European Social Survey data. They use area-level Fay-Herriot models and they find that worry about crime is higher in most South and East European regions, compared to Northern and Central Europe. Buil-Gil, Solymosi and Moretti (2019) develop a two-step method with the goal of producing reliable small area estimates from crowdsourced data, which suffer from different types of bias due to their non-random nature. An application to safety perceptions in Greater London using Place Pulse 2.0 data is presented in their article. Particularly, small area estimates of perceived safety in 1,368 LSOAs across Greater London

are produced and presented in a map available in the publication. They show that there are large differences within each Greater London borough. The lowest values of the estimates of perceived safety can be seen in Eastern neighbourhoods, in particular in some areas of Newham, Waltham Forest and Tower Hamlets. On the contrary, the highest values of perceived safety can be seen in areas of the central boroughs of City of London and Westminster.

Final Recommendations

There is a growing need of estimates at small area levels for many social phenomena such as poverty, crime and well-being. However, many large-scale national sample surveys are not designed to produce reliable estimates at those levels due to their sampling designs. Therefore, small area estimation methods can be used to improve the small area estimates.

In order to apply small area estimation methods there are some steps that must be followed carefully.

First, the small area model needs to be chosen according to the problem that the user is facing. For example, it is important to consider whether the auxiliary variables are available at unit-level or area-level. Also, in a model-based approach, it is crucial to investigate the distribution of the response variable. In this work, normality is assumed, but there are many

other small area models for other types of distributions (see Rao and Molina, 2015 and Pratesi, 2016).

Second, once the model parameters are estimated, it is fundamental to perform model diagnostic; this is to ensure that model assumptions are met.

Third, after the small area estimates are produced, validation of these estimates needs to be carried out e.g. bias diagnostic.

Finally, measures of uncertainty (e.g. mean squared error) need to be produced to check whether the model-based small area estimates return a gain in efficiency compared to the direct estimates.

Further Readings

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis* 58, 308-325.

Molina, I. and Marhuenda, Y. (2015). sae: An R package for Small Area Estimation. *R Journal* 7(1).

Molina, I., Salvati, N. and Pratesi, M. (2009). Bootstrap for estimating the MSE of the Spatial EBLUP. Computational Statistics 24, 441-458.

Moretti, A., Shlomo, N., & Sakshaug, J.W. (2019). Small Area Estimation of Latent

Economic Well-being. *Sociological Methods & Research.* Online First.

Rahamn, A. and Harding, A. (2019). Small Area Estimation and Microsimulation Modeling. CRC Press.

Tzavidis, N., Zhang, L-C., Luna, A., Schmid, T. and Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society Series A* 181, Part4, 927-979.

References

Battese, G. E., Harter, R. M. & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association, 83*(401), 28–36.

Brown, G., Chambers, R., Heady, P., & Heasman, D. (2001). Evaluation of small area estimation methods – an application to unemployment estimates from the UK LFS. In Statistics Canada (Ed.) *Symposium 2001 - Achieving data quality in a statistical agency: a methodological perspective*. Ottawa: Statistics Canada.

Buil-Gil, D., Moretti, A., Shlomo, N., & Medina, J. (2019). Worry about crime in Europe: A model-based small area estimation from the European Social Survey. *European Journal of*

Criminology. Online First.

Buil-Gil, D., Solymosi, R., & Moretti, A. (Forthcoming, 2020). Non-parametric bootstrap and small area estimation to mitigate bias in crowdsourced data. Simulation study and application to perceived safety. In C. A. Hill, P. P. Blemer, T. Buskirk, L. Japec, A. Kirchner & L. E. Lyberg (Eds.), *Big data meets survey science*. Wiley.

Fay, R. E., & 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.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. Journal of Statistical *Computation and Simulation, 78*(5), 443–462.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics, 38*(3), 369-385.

Moretti, A., Shlomo, N., & Sakshaug, J.W. (2019). Multivariate Small Area Estimation of Multidimensional Latent Economic Well-being Indicators. *International Statistical Review.* Online First.

Petrucci, A., & Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological, and Environmental Statistics, 11*(2),

169-182.

Pratesi, M. (Ed.) (2016). Analysis of poverty data by small area estimation. Chichester: Wiley.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation. Second edition*. Hoboken: Wiley.
