SMALL AREA ESTIMATION OF POVERTY INDICATORS

Michele D'Alò, Danila Filipponi, Stefano Gerosa, Francesco Isidori

Abstract. ISTAT has been carrying out extensive research to implement Small Area Estimation (SAE) methods for computing Sustainable Development Goals (SDGs) indicators related to health, occupational status, gender equality, and poverty. This work aims to present the main results obtained applying some SAE methods to estimate the "At Risk of Poverty" indicator for unplanned domains using EU-SILC data. The sub-domains of interest are the provinces (NUTS3) and metropolitan cities, while the survey is designed to provide estimates up to the NUTS2 level (regions). The Small Area Estimation (SAE) methods considered encompass both area and unit-level mixed models, and their results are compared against each other. Administrative data sourced from ISTAT's Integrated System of Registers (ISR), specifically from the Population Register and the Labour Register, integrated with income-related administrative data, are used to specify the models. Furthermore, with direct estimates and administrative auxiliary information available from 2017 to 2021, SAE methods can borrow strength not only from other areas but also from various survey cycles. A final step in the process of estimating small-area statistics through an inferential model-based approach is establishing coherence between estimations of the target indicator computed at various levels of granularity. It is performed to align SAEs with precise and unbiased direct estimates computed at higher planned domain levels. This final calibration is not merely cosmetic. It is essential to meet user requirements on coherence and also to enhance the overall accuracy and reliability of model-based SAEs. The application of Small Area Estimation (SAE) estimates allows gains of efficiency compared to direct estimates.

1. Introduction

Poverty indicators are receiving increasing attention from worldwide institutions searching for innovative approaches to contrast socio-economic inequalities.

Recently, EUROSTAT implemented new precision requirements for the At-Risk-of-Poverty-and-Social-Exclusion (AROPE) indicator, applicable at both national and regional levels (Regulation (EU) 2019/1700 (2019)), annually provided by National Statistical Institutes through the EU-SILC survey. In such context, ISTAT is engaged in an ongoing extensive study on AROPE index and its components (At-risk-of-poverty (ARP), low work intensity (LWI), and severe material deprivation (SMD) indicators) involving, among others, small area estimation techniques. The

final aim of the project is to provide stable and affordable estimates of all the above-mentioned indicators, for the period from 2017 up to the present, at NUTS3 level of aggregation. In the present paper, we will focus on ARP index for the year 2021, the last one for which auxiliary variables are currently available, as a starting point for exploring AROPE components. Hopefully, by collecting enough experiences and case studies, we will be able to add steps towards standardizing the production of small area estimates for indicators mentioned above, such as model selection and tuning, collecting appropriate auxiliary variables, and draw an effective process pipeline. The paper is structured as follows. In section 2 we briefly present the survey sampling design and techniques involved in the production of direct estimates of the ARP index. Sections 3 and 4 are dedicated to the description of applied SAE methodologies. Section 5 contains details about the applications and the obtained results. In section 6 some conclusions are drawn.

2. Target variable, sampling design and direct estimates

The at-risk-of-poverty-rate (ARPR) is a relative poverty index defined as the share of people with an equivalised disposable income below the at-risk-of-poverty threshold, set at 60 % of the national median equivalised disposable income. The modified OECD scale is applied to compare households with different size and composition, so that total household income is converted in equivalised disposable income and is attributed equally to each member of the household.

The IT-SILC sampling design is a two-stage design. Primary sampling units are municipalities, while secondary sampling units are households. Municipalities are stratified according to the number of residents and the stratification is carried out inside each administrative region. Moreover, they are selected in each stratum with probability proportional to their size. Households are not stratified, but are selected with equal probability by systematic sampling in each selected municipality from population register lists and no substitution of unit non-response is applied. More details about EU-SILC survey and sampling design can be found in ISTAT (2008). Direct estimates have been obtained using a calibration estimator (Deville-Särndal (1992)), where the final weights reproduce a set of known socio-demographic totals. Since ARPR is a non-linear indicator, its variance has been estimated by a generalized linearization method (Osier (2009)), taking into account all features of the sampling design and of the calibration estimator (strata, sampling stages, externally calibrated weights).

3. Area-level Models

Small area estimation based on an area-level mixed model, often referred to as the Fay-Herriot (FH) method, is a technique used to estimate the parameters of interest for specific sub-domains by combining survey data with available auxiliary information at the area level. Let d be the generic small area of interest ($d = 1, 2, \dots, D$), $\hat{\theta}_d$ the direct estimate of the target parameter θ_d related to area d, and $\boldsymbol{X}_d$ a set of auxiliary variables known for each area of interest. The area-level mixed model is given by the combination of the following sampling and linking models:

$$\hat{\theta}_d = \theta_d + e_d \; ; \qquad \theta_d = \boldsymbol{X}_d \beta + u_d .$$

The combination of these two models provides the area-level mixed-effects model,

$$\hat{\theta}_d = \boldsymbol{X}_d \beta + u_d + e_d . \tag{1}$$

In the model above, the random effects u_d are assumed to be independent of the sampling errors e_d, and both are normally distributed. The variance $\sigma_{e_d}^2$of the sampling errors is assumed to be known and the other model parameters are estimated by using restricted maximum likelihood method as described e.g. in Rao and Molina (2015, Chapter 5). The Empirical Best Linear Unbiased Predictor (EBLUP) for the target parameter θ_d is a linear combination of a direct estimator and a synthetic estimator.

$$\hat{\theta}_d^{sae} = \hat{\gamma}_d \, \hat{\theta}_d + (1 - \hat{\gamma}_d) \, X_d^T \, \hat{\beta} \; .$$

The weights assigned to the direct estimates are directly related to the variance of the area random effect and inversely related to the sampling variance of the direct estimates

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_{e_d}^2 + \hat{\sigma}_u^2},$$

Since direct estimates and administrative auxiliary information are available from 2017 to 2021, efficiency can be improved by borrowing strength not only from other areas but also from other survey occasions. Rao and Yu (1994) proposed an extension of the basic Fay-Herriot model to handle both time-series and cross-sectional data. This model still consists of a sampling model and a linking model.

Let d be the generic small area of interest ($d = 1, \ldots, D$) and *t* the generic period of time ($t = 1, \ldots, T$) the sampling model is:

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}$$

It deals with the errors associated with the sample data collected for various areas and each time period, considering the variability introduced by random sampling errors. The area linking model is given by:

$$\theta_{dt} = X_{dt}^{T}\beta + u_d + v_{dt}$$

This model focuses on how data, such as direct estimates and known area auxiliary information, from different areas are related over time. The final linear mixed model is given by:

$$\hat{\theta}_{dt} = X_{dt}^{T}\beta + u_d + v_{dt} + e_{dt} \quad (2)$$

where θ_{dt} is the true value corresponding to the estimate $\hat{\theta}_{dt}$ of interest, X_{dt}^{T} is a (DxP) - dimensional matrix of *P* covariates available for each area and time, and e_{dt} are the normal sampling errors. Given the true value θ_{dt}, each vector $e_d = (e_{d1}, \ldots, e_{dT})'$ has a multivariate normal distribution with zero mean and with known variance-covariance matrix Ψ_d. Moreover, $u_d \sim N(0, \sigma_d^2)$ is the area random effect and

$$v_{dt} = \rho v_{d,t-1} + \varepsilon_{dt}$$

with $|\rho| < 1$ and $\varepsilon_{dt} \sim N(0, \sigma_\varepsilon^2)$ is the area-by-time random effect. In this model, e_d, u_d and ε_{dt} are assumed independent of each other and in our application Ψ_d is diagonal, with elements ψ_{dt} , for $t = 1, \ldots, T$. By combining the direct and synthetic estimators, the final composite estimator efficiently borrows strength across small areas and time periods. For a small area d at time t, the composite estimator $\hat{\theta}_{dt}$ can be expressed as:

$$\hat{\theta}_{dt}^{sae} = \hat{\gamma}_{dt}\,\hat{\theta}_d + (1 - \hat{\gamma}_{dt})\,X_{dt}^{T}\,\hat{\beta}$$

in which:

$$\hat{\gamma}_{dt} = \frac{\hat{\sigma}_u^2}{\hat{\psi}_{dt} + \hat{\sigma}_u^2}.$$

4. Unit-level models

The small area estimation using unit-level models are based on the seminal paper of Battese, Harter, and Fuller model (Battese et al. (1988)). These models utilize auxiliary information for each statistical unit, integrating it with survey data to specify a model that borrows strength from similar areas. This approach typically employs a linear mixed model framework to predict parameters of interest in small sub-domains of interest. The basic unit-level mixed model can be formulated as follows:

$$y_{id} = X_{id}^{T}\beta + u_d + \epsilon_{di} \tag{3}$$

Where, denoted with *i* and *d* respectively the generic unit and area, y_{di} is the observed outcome; $X_{di}^{T}$is the vector of auxiliary variables; β is the vector of fixed effect coefficients; u_d is the random effect with $u_d \sim N(0, \sigma_d^2)$) and ϵ_{di} is the random error term for unit *i* with $\epsilon_{di} \sim N(0, \sigma_\varepsilon^2)$.

The variable of interest is a binary indicator which identifies individuals with an equivalised disposable income below the at-risk-of-poverty threshold. Due to the binary nature of the target variable, a logistic mixed model is a natural choice to consider, being specifically designed to handle binary outcomes. The logit of the probability of the outcomes associated with each unit i belonging to the domain can be expressed as follows:

$$logit(P(y_{id} = 1)) = X_{id}^{T}\beta + u_d + \epsilon_{id}, \tag{4}$$

The Empirical Best Linear Unbiased Predictor (EBLUP) estimates of the target parameter are derived after computing the Restricted Maximum Likelihood (REML) estimates of the model parameters β and σ_u^2 σ_ε^2. The small area estimator $\hat{\theta}_d$ is a combination of the direct estimate and the model-based estimate. It can be expressed as:

$$\hat{\theta}_d^{sae} = \hat{\gamma}_d \hat{\theta}_d^{dir} + (1 - \hat{\gamma}_d)\, \hat{\theta}_d^{model} \tag{5}$$

in which:

$$\hat{\theta}_d^{dir} = \frac{1}{n_d} \sum\nolimits_{i \in s_d} y_{id}$$

where n_d is the number of sampled units in area d, while

$$\hat{\theta}_d^{model} = \frac{1}{N_d} \sum\nolimits_{i \in U_d} \hat{y}_{id}$$

where N_d is the total number of units in area d and $\hat{y}_{id}$ is the predicted value for unit i in area d. In case of linear mixed model (3), the predicted values are given by:

$$\hat{y}_{id} = X_{id}^T \hat{\beta} + \hat{u}_d.$$

In case of logistic mixed model (4), the predicted values are instead given by:

$$\hat{y}_{id} = \frac{\exp(X_{id}^T \beta + u_d)}{1 + \exp(X_{id}^T \beta + u_d)}$$

Finally the weight of the area d of the composite estimator (5) is determined on the basis the variance components estimates and is given by:

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 / n_d}$$

The composed small area estimator based on unit-level mixed models effectively integrates direct estimates and model-based predictions, by leveraging individual-level data and area-specific random effects.

5. Application and analysis of the results

The objective of this case study is to estimate the At-Risk-of-Poverty (ARP) indicator at the provincial level (NUTS3) and for 14 Metropolitan Cities. These are considered unplanned domains for the EU-SILC survey since the finest planned domain is at the regional level (NUTS2). Overall, we have 121 unplanned domains, and the target parameter has been estimated using survey data from 2017 to 2021. The next two figures illustrate the distribution of the Coefficients of Variation (CVs) for the direct estimates of the ARP indicator. Figure 1 refers to the NUTS2 planned domains, while Figure 2 shows the distribution of CVs for the NUTS3 unplanned domains. Each box plot in both figures represents the observed distribution of CVs for each available reference period from 2017 to 2021.

Figure 1 – *Boxplot of direct estimates CV within NUTS2 domains*.

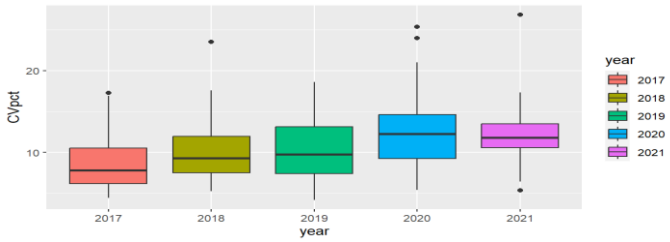
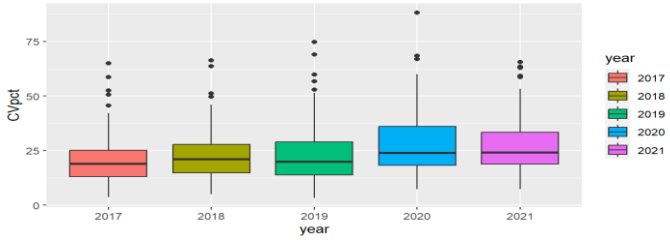


Figure 2 – *Boxplot of direct estimates CV within NUTS3 domains*.



These two plots provide insights into the variability and reliability of the direct estimates over time across different domain levels. It's evident that CVs of the direct estimates have increased over the last two years, largely due to a higher level of non-response rates during the COVID-19 pandemic. This trend appears to be slightly mitigated in Figure 2, which shows the trend of CVs for unplanned domains. For this level of granularity, the impact of missing responses may be masked by the larger variance of direct estimates in unplanned domains, due to the small sample sizes. Given that NUTS2 are planned domains, ARP estimates at the regional level tend to have relatively high CVs. However, since this application focuses on the implementation of SAE methods for unplanned survey domains any analysis of results pertaining to planned domains is set aside for now. In order to evaluate the computed estimates, we will utilize the criteria proposed by Statistics Canada (https://www150.statcan.gc.ca/n1/pub/71-543-g/2016001/part-partie7-eng.htm#archived), although other criteria determined consulting users and experts, may also be considered. According to such criterion, estimates having CV less than 16.6% can be released without any restriction. Estimates with CV between 16.6% and 33.3% can be released with caveats and should be always accompanied by a warning regarding their accuracy. Finally, if CV is greater than 33.3% the corresponding estimates should not be released. Table 1 shows the number of direct estimates in the described groups for the year 2021. As expected, along with the two out of sample areas, many direct estimates of the target parameter show high CVs.

Hence, implementing small area estimator methods is a crucial step to try to enhance the efficiency of estimates at the desired level of disaggregation.

Table 1 – *NUTS3 Estimates grouped by CV*.

CV%	Evaluation	Number of estimates
≤ 16.5	Publishable	20
(16.5; 33.3]	Publishable with caution	67
>33.3	Not recommended for publication	32
Not available	Not available	2

To compute small area estimates of ARP for the 121 sub-domains of interest, we consider the two estimators (eblup_lin and eblup_logit) based on the mixed unit-level model described in paragraph 4, and the two estimators (FH and YR) based on area-level mixed model described in paragraph 3. The fixed part of the models were specified using administrative information available in ISTAT's Integrated System of Registers (ISR), particularly from the Population Register and the Labour Register, integrated with administrative data on income (Baldi et al., 2018). Specifically, we considered:

- Population distribution for 7 age classes;
- Population distribution for 3 education level classes (Primary education, secondary education, university degree)
- At risk of poverty index administrative proxy;
- Quintiles of equivalent income at the national, regional, and provincial level;
- Population distribution for work income, pension income and capital income grouped in five 5 classes;
- Population distribution for 4 classes of the average number of working weeks, obtained by dividing the year into quarters.

This auxiliary information was integrated with the survey data to specify and fit the unit-level models (3) and (4). Aggregated mean values of the same information at the domain level were instead used to fit the mixed area-level models (1) and (2). The standard Fay-Herriot area-level estimator assumes normality and independence of the error terms. However, in this application, these assumptions appear to be violated. To address this issue, the area-level model has been specified on the log-transformed direct estimates. SAEs based on this model have been computed using the emdi package (see Harmening, et al. (2023)). The log-transformation ensures a better fit of the area-level model to the normality assumptions of the random error components. This transformation introduces a bias when converting back to the

original scale, which can be adjusted applying a so called 'crude' method, as described by Harmening et al. (2023).

Additionally, like the direct estimates, their corresponding variance estimates can also be very unstable. Therefore, smoothed estimated variances has been considered for computing both standard and log-transformed FH - SAEs. Assuming that the variance of estimates depends on the area sample size and the intensity of the target variable, a simple linear model has been used to smooth the estimated variances of the direct estimates. The applied model is:

$$\ln(var(\hat{\theta}_d)) = \beta_0 + \beta_1 \ln(n_d) + \beta_2 \ln(\hat{\theta}_d)$$

where n_d and $\hat{\theta}_d$ are respectively the realized sample size in the area d and the direct estimates.

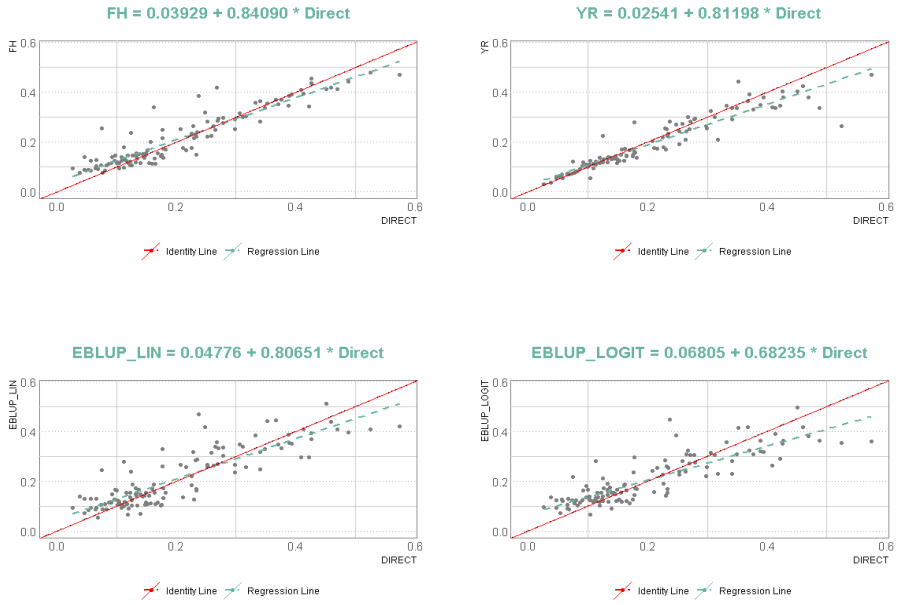
Both of the above adjustments lead to more satisfactory model fitting and improved properties of the small area estimates. Consequently, the following results and analyses related to FH estimates will refer just to the small area estimates computed on the basis of a standard area-level model specified considering the log-transformation of the direct estimates and the smoothing of their variance.

Figure 3 illustrates comparisons between direct estimates and model estimates, highlighting that SAEs based on area-level models align more closely with direct estimates compared to those computed using two EBLUP estimators based on unit-level models. ARP's FH estimates, as expected with SAE methods, allow to smooth both the lowest and highest direct estimates, while YR mainly reduces the intensity of the largest direct. The increasing trend of ARPR estimates from 2017 to 2021 (see Figures 1 and 2) can lead the YR estimator to produce lower SAEs, as the model is specified to borrow strength not only from other areas but also from the time occasions of the survey. The logistic unit-level model does not seem to yield better results compared to its linear counterpart, as expected given the binary nature of the response variable. Further in-depth analysis is needed to understand the reasons.

A benchmarking procedure, aimed at ensuring the consistency of target indicator estimates across different levels of disaggregation, is performed to align SAEs with precise and unbiased direct estimates computed at planned domain levels. This final calibration is not merely cosmetic. It is essential to meet user requirements on coherence and also to enhance the overall accuracy and reliability of model-based SAEs, by reducing the possible bias of SAEs. Those estimates should be aligned to the finer planned domains' direct estimates (NUTS2 level). However, as shown in Figure 1, these estimates are not sufficiently reliable, with some having a CV exceeding 20%. Consequently, the small area estimates were benchmarked against the more reliable ARP direct estimates computed at the NUTS1 level, corresponding to a division of the Italian territory into five groups of administrative regions: North

East, North West, Centre, South, and Island. The benchmarking adjustment introduces an extra variability that is added to the original MSE of SAEs.

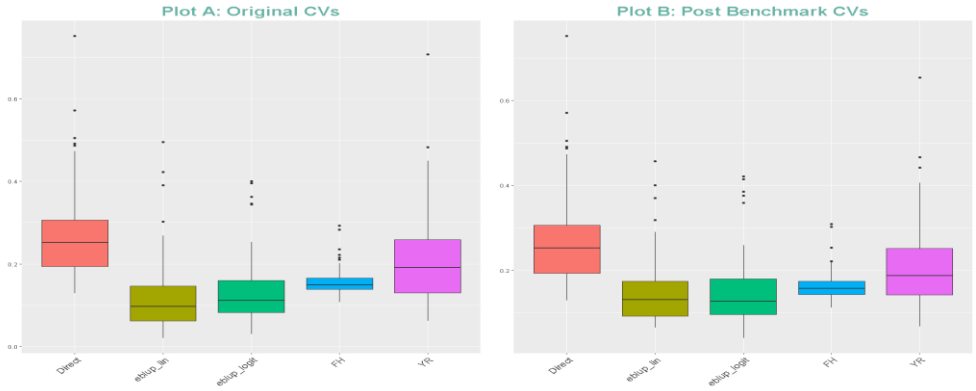
Figure 3 – *Comparisons between direct and models estimates.*



The set of estimates' coefficient of variation is reported in Figure 4. Plot A displays the original CVs of the SAEs, while plot B shows the distribution of CVs of correspondent post-benchmarked estimates. All SAE methods allow considerable efficiency gains over the direct estimator, with the FH estimator outperforming the other SAE methods. It is worth highlighting that the CV distribution for the two estimators based on unit-level models have a lower median compared to other estimators. This can be attributed to the significant correlation between the response variable and the set of unit-level administrative data used to specify these models. However, both SAE methods based on linear and logit unit-level models show a broader distribution of CVs, with higher maximum values compared to the FH estimator. Further model specifications and assumptions analysis are needed to better understand the reasons behind these outcomes. The distribution of CVs for the YR estimator is not good, with only slight improvement observed after the benchmarking, which mitigates the over-shrinkage of the estimates computed with this estimator. To improve results, like done for the FH methods, one might consider

to specify the You-Rao model on the log transformation of the direct estimates and after the smoothing of the estimated variances.

Figure 4 – *CV of direct and the original and post benchmarked SAEs.*



6. Conclusions

The results are encouraging, with the FH estimator outperforming the other SAE methods considered in this study. We are currently working on incorporating both variance smoothing and logarithmic transformation into the YR time series area-level model, as done with the basic FH model. Additionally, other SAE methods that utilize longitudinal information in unit-level models and account for spatial correlation in both area and unit-level models should be considered. Exploring the Empirical Bayes Predictor (EBP) proposed by Molina and Rao (2015) for estimating poverty indicators is also worthwhile, despite its high computational cost. Finally, it is essential to thoroughly evaluate the models' goodness of fit, validate the specified assumptions, and conduct both statistical and thematic assessments of all SAE estimates produced.

References

BALDI C., CECCARELLI C., GIGANTE S., PACINI S., ROSSETTI F. 2018. The Labour Register In Italy: The New Heart Of The System Of Labour Statistics, *Rivista Italiana di Economia, Demografia e Statistica*, Vol. LXXII, No. 2, pp. 95-105.

BATTESE G. E., HARTER R. M., FULLER W. A. 1988. An error components model for prediction of county crop area using survey and satellite data, *Journal of the American Statistical Association*, Vol. 83, pp. 28-36.

DEVILLE J. C., SÄRNDAL C. E. 1992. Calibration estimators in survey sampling, *Journal of the American Statistical Association*, Vol. 87, No. 418, pp. 376-382.

FAY R. E. AND 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*, Vol. 74, No. 366, pp. 269-277.

HARMENING S., KREUTZMANN Ann-K., Salvati N., Schmid S. 2023. A Framework for Producing Small Area Estimates Based on Area-Level Models in R, *The R Journal*, Vol. 15, pp. 316-341.

OSIER G. 2009. Variance estimation for complex indicators of poverty and inequality using linearization techniques, *Survey Research Methods*, Vol. 3, No. 3, pp. 167-195.

REGULATION (EU) 2019/1700 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing a common framework for European statistics relating to persons and household. 2019. *Official Journal of the European Union.* LI 261/1

RAO J.N.K., MOLINA I. 2015. *Small Area Estimation.* Wiley Series.

RAO J. N. K., YU M. 1994. Small-Area Estimation by Combining TimeSeries and Cross-Sectional Data, *The Canadian Journal of Statistics*, Vol. 22, No. 4, pp. 511–528.

Michele D'ALÒ, Istat, dalo@istat.it
Danila FILIPPONI, Istat, dafilipp@istat.it
Stefano GEROSA, Istat, gerosa@istat.it
Francesco ISIDORI, Istat, francesco.isidori@istat.it