SMALL AREA ESTIMATION OF SEVERE FUNCTIONAL LIMITATION FROM ITALIAN DATA OF THE EUROPEAN HEALTH INTERVIEW SURVEY

Michele D'Alò, Andrea Fasulo, Francesco Isidori, Maria Giovanna Ranalli

Abstract. This paper focuses on the methodology used to estimate the indicator of severe functional limitation (SFL) using data collected from the European Health Interview Survey (EHIS) in Italy. While direct estimates of SFL are reasonably accurate up to the regional level (NUTS2), there is a demand for more detailed estimates at the provincial level (NUTS3), disaggregated by sex and two age groups (15-64 years and 65 years and above). This requires the computation of estimates for 428 unplanned domains. To address this challenge, a small area estimation approach based on an area-level model has been applied, integrating auxiliary information known from administrative registers with EHIS data. To meet the assumptions of the model and ensure in this way a better accuracy of the final estimates, the model has been specified on a log-transformation of direct estimates. The case study presented here is one of the first attempts at obtaining small area estimates for unplanned domains within the EHIS survey and the results obtained are very promising.

1. Introduction

There is a growing demand for increasingly detailed statistical information, particularly regarding estimates of socio-economic indicators at a highly granular level. This demand arises from the need to support urban policies that effectively consider and incorporate specific local characteristics. Decision-makers and policymakers require comprehensive and accurate data to tailor policies that cater to the unique needs and challenges of specific places. Moreover, a considerable number of United Nations Sustainable Development Goals are pursued using survey indicators at a very detailed level.

In order to explore appropriate estimation methodologies for obtaining estimates at a level of granularity beyond that of planned domains by the main social surveys carried out by the Italian National Institute of Statistics (ISTAT), a working group was established, comprising experts in small area estimation from both ISTAT and the Academia. The primary objective of the group was to

delineate the process for computing small area estimates (SAEs) of relevant indicators for unplanned domains, drawing from the main social surveys conducted by ISTAT. In the latest edition of this working group, three subgroups were formed, each with a specific focus on defining the production process of SAE for relevant indicators, collected through three distinct surveys: the European Survey on Income and Living Conditions (EUSILC) for poverty indicators, the Aspect of Daily Life survey (AVQ) for ITC indicators, and the European Health Interview Survey (EHIS) for health indicators. In the previous edition of the working group, there was a dedicated subgroup focusing on SAE of indicators derived from the Labour Force Survey (LFS), particularly at the functional area level. However, since ISTAT has a well-established tradition of applying SAE techniques for LFS indicators, the decision was made to temporarily set aside this specific focus. Nonetheless, the expertise and methodologies developed to produce SAEs for LFS indicators have been valuable for implementing case studies and computing SAEs from other social surveys.

This paper aims to describe the methodology used to estimate the indicator of severe functional limitation (SFL) using data collected through the European Health Interview Survey (EHIS). EHIS gathers information on key aspects of the population's health conditions and the use of healthcare services for citizens aged 15 and above. The adopted sampling design is a two-stage stratified sampling: municipalities are first-stage units, while households are second-stage units. The final sample includes approximately 22,800 households, living in 835 municipalities of different sizes and distributed throughout the national territory. The areas considered in the survey include the five main geographical areas (North-West, North-East, Center, South, Islands) according to the NUTS 1 classification, Italian Regions (NUTS2), and the two autonomous provinces of Bolzano and Trento. Broad areas defined according to the national health program are also domains of interest.

Direct estimates of SFL have an acceptable level of error up to the regional level. Therefore, ISTAT internal request to provide estimates at the provincial level (NUTS3), disaggregated by sex and two age groups (15-64 years; 65 years and above), has required the need to compute estimates for 428 unplanned domains. There are no "a priori" guarantees about the validity of estimates for this level of granularity, as they may have high sampling errors. Therefore, to meet the request to produce an estimate for the indicator at such a level of disaggregation, specific estimation methods for small areas have been adopted.

The paper is structured as follows. In Section 2, a concise overview of the survey sampling design and the methodology employed to calculate the direct estimates of SFL is presented. Section 3 provides a brief description of the small area estimation method applied in the study. In section 4, the main outcomes of the

case study concerning the target parameter are examined in detail. In conclusion, Section 5 presents definitive insights drawn from the main findings and outlines the necessary future work required to further validate the proposed model-based estimates.

2. Sample design, direct estimates and sampling errors

The European Health Interview Survey (EHIS) is conducted in all European Union member states with the aim of computing comparable health indicators at the European level on key aspects of the population's health conditions, the use of healthcare services, and health determinants. The Eurostat methodological manual[1] provides all the recommendations and instructions to best implement the survey. Italy has selected modules on the social participation of people with disabilities (Disability module) and the evaluation of received healthcare services (Patient Experience module) among the additional modules.

The sampling design has the usual structure of most social surveys on households carried out by the ISTAT. This design is based on a two-stage sampling design, with stratification of municipalities based on their population size. The 2019 survey design was integrated with the one used for the Master Sample of the Permanent Census. The selected municipalities are a sub-sample of the 2850 municipalities present in the Master Sample selected for the 2018 Census round. The second-stage units are households, randomly selected (*i*) from the population registers for sample municipalities with fewer than 1000 inhabitants and (*ii*) from the list of households selected for the 2018 Permanent Census for sample municipalities with more than 1000 inhabitants. As stated in the Introduction, the final sample consists of approximately 22,800 households in 835 municipalities of various sizes and distributed throughout the national territory. For further details on the adopted sampling strategy, see ISTAT (2021).

The topics covered in the survey relate to three main areas: health status, health determinants, and access to and use of healthcare services. Most sections of the survey modules refer to the population aged 15 and above, as required by the European Regulation.

The estimates produced by the survey are absolute and relative frequencies, referring to households and individuals. The estimates are obtained using a calibrated estimator (Devaud and Tillé, 2019; Deville and Särndal,1992; Särndal, 2007), with benchmark constraints given by:

[1] https://ec.europa.eu/eurostat/documents/3859598/8762193/KS-02-18-240-EN-N.pdf/5fa53ed4-4367-41c4-b3f5-260ced9ff2f6?t=1521718236000

- distribution in the 21 Italian regions (19 regions plus the provinces of Trento and Bolzano) by sex and seven age groups (0-14, 15-24, 25-44, 45-54, 55-64, 65-74, 75+);
- distribution of the population in the 5 territorial divisions by sex and nine age groups (0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+);
- distribution of the population by citizenship (households of Italian citizens, households of foreign citizens, mixed households).

In particular, the target indicator, SFL, is given by the proportion of people above 15 years of age who have long-term physical, mental, intellectual, or sensory impairments that, when interacting with various barriers, may hinder their full and effective participation in society on an equal basis with others (ISTAT, 2015).

To assess the reliability of estimates, the classification based on the coefficient of variation (CV %) considered by Statistics Canada for the Labour Force Survey is applied. In particular, estimates are classified as follows:

1. estimates publishable without restrictions, CV % $\leq$ 16.5;
2. estimates publishable with caution, 16.5% < CV % $\leq$ 33.3%;
3. estimates not recommended for publication, CV % > 33.3%.

The estimates are based on 2019 EHIS survey data. Table 1 shows the number of estimates falling into the three categories, as well as the number of domains for which direct estimates are not available. The large number (168) of estimates with CV exceeding the threshold for release, added to the unavailability of 21 domains that are out of sample, highlights the need of employing small area estimation methods. These methods allow to enhance the precision of estimates for disaggregated domains by borrowing strength from other areas and exploiting the relationship between the variable of interest and a set of auxiliary variables available at the elementary unit or area level. Due to privacy concerns, integrating information from other sources, such as administrative data with EHIS survey data at the elementary unit level is not feasible. Consequently, only auxiliary information known at the area level can be employed. In this informative context, the applied method is an estimator based on a mixed-effects model defined at the area level, proposed by Fay and Herriot (1979).

Table 1 – *Estimates that can be released, that can be released with warning and estimates that are too unstable to be released, for the indicator SFL – year 2019.*

CV%	Evaluation	Number of estimates
$\leq$ 16.5]	Publishable	87
(16.5; 33.3]	Publishable with caution	152
>33.3	not recommended for publication	168
Not available	Not available	21

3. Small Area Estimation based on a Mixed Area Level Model

Small area estimation based on an area-level mixed model, often referred to as the Fay-Herriot method, is a technique used to estimate the parameters of interest for specific domains (areas) 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 two models:

$$\hat{\theta}_d = \theta_d + e_d$$

$$\theta_d = \boldsymbol{X}_d\beta + u_d$$

where the sampling errors e_d and the area-specific random effects u_d have zero mean. The combination of these two models provides the following mixed-effects model,

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

where the random effects u_d are assumed to be independent of the sampling errors e_d, and both are normally distributed. The variance σ_e^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). Denoted with $\hat{\beta}$ and $\hat{u}_d$ the estimate of the fixed effects and the prediction of the area-specific random effects, respectively, the resulting Empirical Best Linear Unbiased Predictor for θ_d can be written as a linear combination of a direct and a synthetic estimator:

$$\hat{\theta}_d^{sae} = \hat{\gamma}_d\,\hat{\theta}_d + (1 - \hat{\gamma}_d)\,\boldsymbol{X}_{\boldsymbol{d}}^{\boldsymbol{T}}\,\hat{\beta} \tag{2}$$

where $\hat{\gamma}_d$ is a shrinkage factor that represents the weight assigned to the direct estimator. In particular, it is given by:

$$\hat{\gamma}_d = \hat{\sigma}_u^2/\hat{\sigma}_e^2 + \hat{\sigma}_u^2, \tag{3}$$

where $\hat{\sigma}_e^2$ is the estimated sampling variance of the direct survey estimate, and $\hat{\sigma}_u^2$ is the estimated variance of the random effects.

4. The application results

The case study's objective is to estimate the SFL indicator at a specific level of disaggregation, defined by the cross-classification of provinces, sex, and two age classes (15-64; 65+), resulting in a total of 428 unplanned domains of interest. The planned domains with a targeted level of accuracy are the regions, having a coefficient of variation (CV) ranging from 5% to 12%. In order to compute estimates for the required level of disaggregation, the mixed area level model is specified using two auxiliary variables available at the area level:

- disability certification, available from INPS, the Italian National Social Security Institute, and
- the hospital attractiveness index, available from the Ministry of Health, that is used to monitor the healthcare service.

Additionally, valuable area level information is gathered from other small area estimation case studies, particularly for poverty estimation, where integration at the unit level between survey data and administrative information was feasible. This supplementary information is accessible through ISTAT's Integrated System of Registers, particularly from the Population Register and the Labour Register integrated with administrative data for income (Baldi et al., 2018). This additional auxiliary information comprises:

- population distribution for 7 age classes;
- population distribution for three education level classes (primary education, secondary education, university degree);
- at risk of poverty rate;
- quintiles of equivalent income at the national, regional, and provincial level;
- population distribution for work income, pension income and capital income grouped in five classes;
- population distribution for four classes according to the average number of working weeks, obtained by dividing the year into quarters.

All these variables' frequencies within the domains of interest have been considered to specify the small area level model.

Figure 1 shows the distribution of direct estimates in the 428 domains of interest. It is evident that the parameter of interest exhibits different intensities in the two age classes, with some overlaps between the upper tail of the estimates concerning the first age class and the lower tail of the estimates concerning the second class. Thus, the distribution of the estimates is characterized by a mixture of two distributions, as depicted in Figure 2. To address this situation, the fixed part of the mixed model is formulated incorporating the interaction of covariates with the two distinct age classes (15-64 and 65+). The final model is chosen through a

stepwise selection of relevant auxiliary information and its interaction with the two age classes. The SAEs of the target variable are calculated using the R package emdi (Kreutzmann et al. 2019), which stands for "Estimating and Mapping Disaggregated Indicators." This package is freely available on the R CRAN platform (https://cran.r-project.org/web/packages/emdi/index.html).

Figure 1 – *Distribution of direct estimates in the domains of interest*.

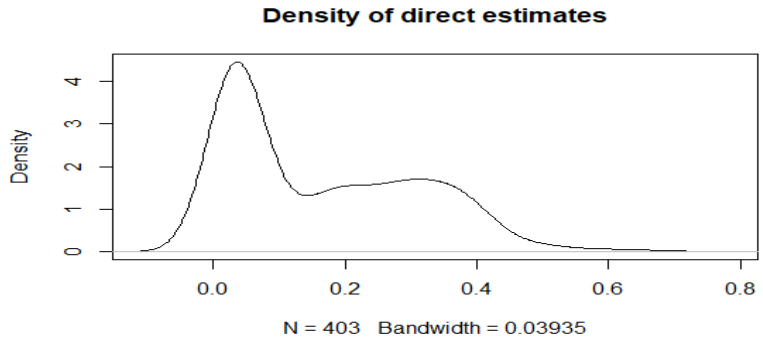
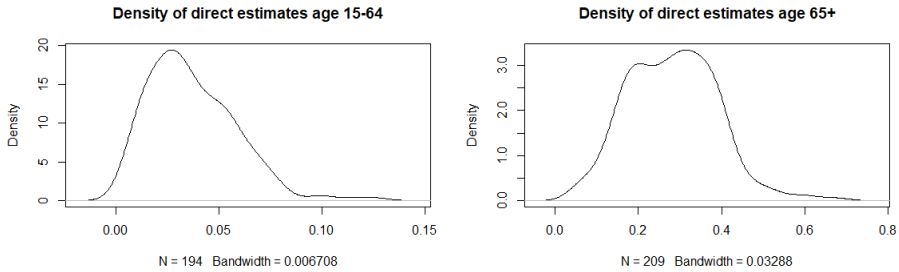


Figure 2 – *Distribution of direct estimates in the domains of interest in the two classes of age.*



The standard Fay-Herriot area-level estimator assumes normality and independence of the error terms. However, in this specific application, this assumption appears to be violated, as shown in Figure 3 which compares the distribution of the realized standardized residuals and of the standardized random effects of the FH model with their normal counterparts. To take this issue into account, the area level model has then been specified on the log-transformation of the direct estimates, and SAEs based on this model have been computed using emdi package. For more details on the log-transformed area-level model, see Kreutzmann et al. (2019). The log-transformation allows for a better suitability of

the fitted area level model to the assumptions of normality of the random error terms, as illustrated in Figure 4.

Figure 3 – *Distribution of the realized standardized residuals and of the standardized random effects of the standard FH model.*

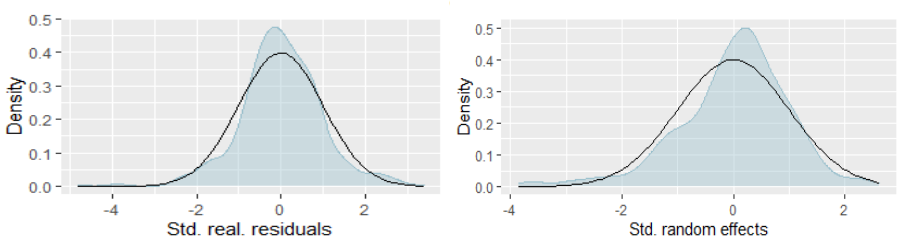
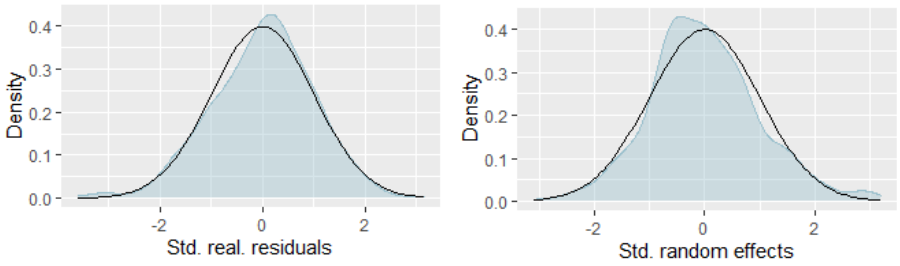


Figure 4 – *Distribution of the realized standardized residuals and of the standardized random effects of the log-transformed FH model.*



As with direct estimates, also their variance estimates can be very unstable, so that smoothing these variance has been considered and used for computing the log-transformed FH SAEs. Assuming that the CV of estimates depends on the area sample size and on the target variable, the smoothing model considered is given by

$$\ln(CV^2(\hat{\theta}_d)) = \beta_0 + \beta_1 \ln(n_d) + \beta_2 \ln(\hat{\theta}_d) \quad (5)$$

Figure 5 clearly illustrates the impact of employing a smoothed set of sampling variance estimates in contrast to the more unstable original estimates: the distribution of the shrinkage factor $\hat{\gamma}_d$ in estimator (2), as defined by expression (3), shows notable improvement with the use of smoothed variances. In fact, by incorporating the smoothed variances, the difference between the shrinkage factors for the two age classes becomes more evident, along with their respective trends with respect to the sample size. This enhancement significantly improves our

understanding of how the shrinkage factor influences the expression of estimator (2) in the determination of the SAEs of interest.

Figure 5 – *Distribution of the estimated shrinkage factor γ_d as a function of the sample size for the log-transformed FH model when the original (left) and the smoothed (right) variance is used to computed SAEs.*

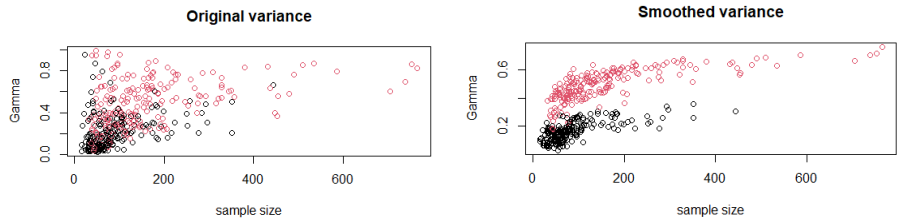


Figure 6 compares the values of the SAEs obtained using the standard FH method and the log-transformed FH method with respect to the direct estimates. The log transformation of data yields estimates that have higher consistency with direct estimates, in contrast to estimates derived from the standard area level model. This highlights the effectiveness of log-transformation in producing more reliable and precise SAEs in this case. This finding is further supported by Figure 7, which clearly shows a significant efficiency gain when applying the FH method to data transformed using a logarithmic function, as opposed to using the original data.

Table 2 displays the distribution of estimates across the three classes of %CV for direct estimates and for the SAEs computed under the standard FH and log-transformed FH models. It can be observed that all FH-LOG estimates have a coefficient of variation below 33%.

Figure 6 – *Direct estimates versus standard FH and LOG transformed FH SAEs.*

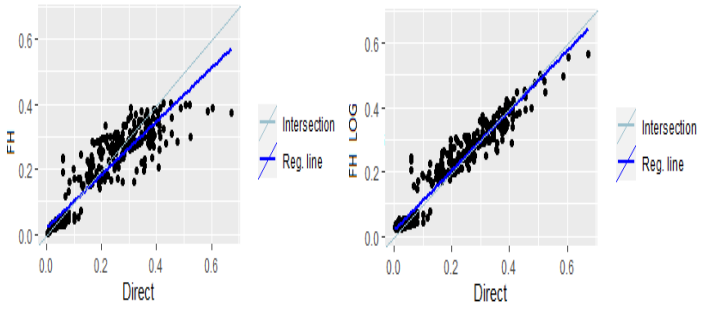


Figure 7 – *Distribution of Coefficient of Variation (CV) of Direct estimates versus standard FH and LOG transformed FH SAEs.*

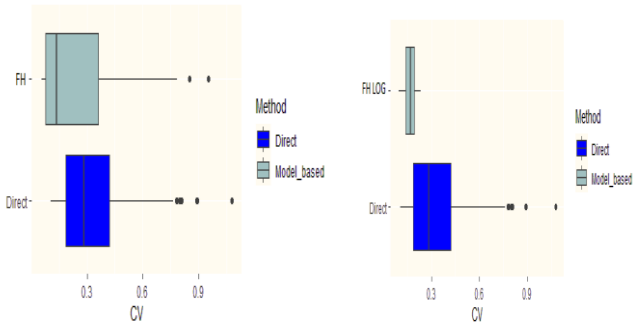
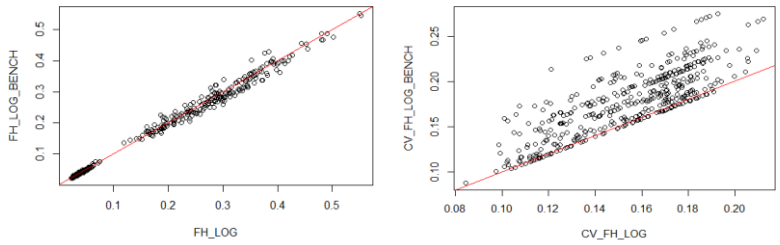


Table 2 – *CV distribution of direct and model based estimates.*

Estimator	CV%			
	< 16.6	*16.6-33.3*	*>33.3*	*Not available*
Direct	87	152	168	21
FH	213	65	150	0
FH LOG	196	232	0	0

Another crucial step is to ensure that all the estimates of the target indicator computed at different levels of disaggregation are consistent. To achieve this, the SAEs were benchmarked to the direct estimates produced for the planned regional domain. This process ensures that SAEs are aligned with the unbiased direct estimates at a regional level, enhancing also in this way the overall accuracy and reliability of the model based SAEs produced for the unplanned domains. The good performance of the FH-LOG estimator in terms of accuracy is further validated in Figure 8. The pictures show a comparison of the distribution of SAEs and their CV, before and after benchmarking. Following the benchmarking procedure, SAEs show small changes compared to the pre-benchmarked estimates, indicating a good fit of the specified model. As expected, the CVs of the estimates after benchmarking slightly increase in comparison to their respective pre-benchmarked counterparts, as the benchmark procedure introduces additional variability due to the adjustments needed to achieve the coherence among estimates.

Figure 8 – *Distribution LOG-transformed FH SAEs (left) and corresponding CV (right) before and after benchmarking.*



5. Conclusions

The application of a small area estimation method based on a mixed area level model with log-transformed data has yielded promising results for the target indicator of Severe Functional Limitation at the required unplanned domains from the European Health Interview Survey. It allows good gains of efficiency of the produced estimates with respect to direct estimates. Nonetheless, several further actions should be implemented to further enhance the small area estimation process. Firstly, as soon as it becomes available, auxiliary information from the ISTAT disability register could be considered. Moreover, it is important to further assess the quality of the SAEs produced, also by means of a process of validation of the estimates carried out by users and thematic experts.

Acknowledgements

We would express our gratitude for the fruitful collaboration of all the members of the sub-working group WP3 for Small Area Estimation (SAE), and in particular to Isabella Corazziari, ISTAT DIRM/DCME/MEB, Lidia Gargiulo and Laura Iannucci from ISTAT SWC, Gaia Bertarelli from the University of Venice, and Francesco Schirripa Spagnolo from the University of Pisa.

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.

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.

DEVAUD, D., TILLÉ, Y. 2019. Deville and Särndal's Calibration: Revisiting a 25-Years-Old Successful Optimization Problem, *Test*, Vol. 28, No. 4, pp. 1033-1065.

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.

KREUTZMANN, A., PANNIER, S., ROJAS-PERILLA, N., SCHMID, T., TEMPL, M. AND TZAVIDIS, N. 2019. The R Package Emdi for Estimating and Mapping Regionally Disaggregated Indicators, *Journal of Statistical Software*, Vol. 91, No. 7, pp. 1-33.

ISTAT. 2015. *Social inclusion of people with functional limitations, impairments or severe chronic diseases, Methodological Note*. Roma: Istituto Nazionale Di Statistica.

ISTAT. 2021. *Condizioni di salute e ricorso ai servizi saniotari in Italia e nell'Unione Europea, Nota metodologica*. Roma: Istituto Nazionale Di Statistica.

RAO J.N.K., MOLINA I. 2015. *Small area estimation*. New Jersey: John Wiley & Sons.

SÄRNDAL, C. E. 2007. Calibration Estimators in Survey Sampling, *Survey Methodology*, Vol. 33, No. 2, pp. 99-119.

Michele D'ALÒ, Istat Dirm/Dcme/Meb, dalo@istat.it
Andrea FASULO, Istat Dirm/Dcme/Meb, fasulo@istat.it
Francesco ISIDORI, Istat Dirm/Dcme/Meb, isidori@istat.it
Maria Giovanna RANALLI, University of Perugia, maria.ranalli@unipg.it