

Estimation of Health Indicators at Sub-regional Level

Stima di Indicatori sulle Condizioni di Salute a Livello Sub-regionale

Claudia De Vitiis, Stefano Falorsi, Aldo Lamberti, Alessandro Pallara

Italian National Institute of Statistics; Via Magenta, 2, Rome, Italy;

e-mail: devitiis@istat.it.

Riassunto: In questo lavoro si propone una strategia per la stima di indicatori sulle condizioni di salute, l'incidenza delle malattie e il ricorso ai servizi sanitari per domini territoriali (Azienda Sanitaria Locale) usualmente non contemplati nella progettazione del disegno d'indagine e, pertanto, non ottenibili utilizzando metodi diretti di stima.

Keywords: Small area estimation, Random-effect model, Fay-Herriot model.

1. Introduction

A very comprehensive examination of health status, health practices and health care utilization is conducted every five years by the Italian National Statistical Institute in the Health Status and Health Service Utilization Survey (HSHSUS). The HSHSUS provides direct national and regional estimates of these items, while little can be yielded directly from the survey data at sub-regional level, except for a few selected metropolitan areas. The need for more detailed health statistics exists, however, because health and health care characteristics are known to vary geographically and, also, health care planning often comprises local sub-regional areas, including sometimes very small fraction of territory, usually defined for administrative purposes, like, e.g., the *Azienda Sanitaria Locale* (hereafter denoted as Health Service Area, HSA).

The focus of this paper will be the production of small area (HSA) estimators from the HSHSUS. The HSHSUS is a two-stage sample survey. The sampling design fielded in 1999-2000 stratified municipalities (primary sampling units, PSUs) within each region into strata of approximately equal size, in terms of total population. Within each sampled PSU, households were randomly selected from population registers and each member of the household was interviewed. The selected sample in 1999-2000 HSHSUS consisted of approximately 60,000 households in 1463 municipalities.

In our study, the small areas are the HSAs, that are unplanned domains which cut across design strata. Therefore, sample size in most HSAs is not large enough to produce precise direct estimates. This will make it necessary to produce indirect estimates for the variable of interest using, e.g., data from other areas, through (implicit or explicit) models that assume similarities across areas. In this case, usually, the estimation process depends on some auxiliary variables that are known for the area of interest. Moreover, since the sample design is clustered, it occurs that the number of PSUs sampled in each HSA is very small and there will be the problem of how to compute estimates of the variance for a small area, that will depend on the PSUs within each small area and will, then, likely to be imprecise, with a very small number of degrees of freedom.

The paper is organized as follows. In section 2, having simply mentioned, for brevity, how small area direct estimates were yielded, attention is focused on obtaining improved (indirect) variance estimates for the small area, borrowing data from related areas. Indeed,

modeling variance for a small area from a clustered design is somewhat independent of whether the small area estimates of survey variables are for planned or unplanned domains. Then, section 3 will discuss briefly the small area estimators that have been used in our study. Depending on the amount and detail of auxiliary information available these will include more traditional indirect estimation methods as well as area-level or unit-level fixed or random effects linear models, which may further account for spatially correlated data structure. The proposed strategy for small area estimation will be illustrated throughout the paper with an application to estimation of some health indicators for HSAs in one region (Campania) from 1999-2000 HSHSUS.

2. Estimation of sampling error of the direct estimates

The first step to produce estimates at HSA level is evaluating the accuracy of the direct estimates, which are obtained by means of a calibration estimator (Deville e Särndal, 1992), using age and sex groups as auxiliary variables at regional level. Since the sample design is clustered, it may occur that very few PSUs fall in each HSA; therefore, the direct estimates of the variance for the small area, depending on the variability within the PSUs of each HSA, can be very imprecise, as it is likely to be calculated on a very small number of observations. In this situation, the variance estimation of the direct estimates at HSA level is an issue itself which is possible to deal with using indirect variance estimator, as suggested in Singh *et al.* (1994) and Marker (2001). The general idea is borrowing strength from other areas, by using a generalised variance function approach, in which it is assumed that the design effect of a larger area, comprising the small area of interest, applies to the small area. As a “macroarea” for each HSA it has been considered using either the whole region or the minimal union of design strata encompassing the small area. We finally preferred using the latter, as it allows to yield less biased estimates. The expression for the variance estimation of the direct estimate of the total Y_d for the small domain d is therefore defined as

$$\hat{Var}(\hat{Y}_d) = \hat{Var}_{srs}(\hat{Y}_d) \times d\hat{eff}_M(\hat{Y}_M) \quad (1)$$

with $\hat{Var}_{srs}(\hat{Y}_d)$ representing the estimated variance under a simple random sample (SRS) with the same sample size, in term of final units, as the effective one of the small area d ; $d\hat{eff}_M(\hat{Y}_M)$ is the design effect, estimated at the macroarea M level as the ratio between the estimated variance under the effective complex design and calibration estimator (based on the PSUs of the whole macroarea M) and the estimated variance under SRS. Expression (1) holds under the hypothesis of invariance of the design effect components (the intracorrelation coefficient and the number of final units per PSU) within the macroarea. The obtained values of the *deft* (square root of design effect) relative to four indicators from HSHSUS at regional, macroarea, as well as HSA level for the thirteen HSAs of the Campania region are presented in table 1.

As it can be observed from the table, using information from other areas resulted in a more stable variance estimate for the direct estimates at HSA level.

Note that out of the four variables presented in Table 1 the first two (myocardial infarction and cancer) show lower estimated prevalence combined with higher errors of direct estimates. It may then be useful trying to improve precision of estimates for these variables through small area estimators, which may benefit from the availability of auxiliary information from either survey data or administrative sources.

Table 1. Direct (D=HSA) and indirect (M=macroarea) estimates of square root design effects for four indicators from 1999-2000 HSHSUS

	Indicator							
	MYOCARDIAL INFARCTION		CANCER		AMNIOCENTESIS DURING PREGNANCY		MORPHOLOGIC FETAL ECHOGRAPHY	
Regional deft	0,96		1,11		1,39		0,88	
HSA	M-deft	D-deft	M-deft	D-deft	M-deft	D-deft	M-deft	D-deft
AV/1	1,04	1,14	1,40	0,61	1,33	-	0,79	1,97
AV/2	0,79	1,63	1,02	1,53	1,71	1,56	1,18	1,95
BN/1	1,13	0,49	1,41	0,00	1,10	1,97	0,52	2,55
CE/1	0,98	1,19	0,83	0,99	1,61	2,81	1,40	4,62
CE/2	0,79	0,68	1,66	-	1,71	3,51	0,52	4,08
NA/1	0,95	0,96	0,95	0,96	1,01	1,06	0,70	0,98
NA/2	0,85	0,91	1,13	1,19	1,29	1,73	0,54	2,79
NA/3	0,85	1,21	1,07	1,06	1,34	1,43	0,68	2,37
NA/4	1,18	0,00	1,24	1,09	1,95	0,53	0,98	2,86
NA/5	0,78	1,07	0,96	1,01	1,22	1,34	0,79	3,17
SA/1	0,94	1,15	0,79	0,76	1,51	2,70	1,34	4,43
SA/2	0,99	1,16	1,33	0,88	1,37	1,26	0,32	3,65
SA/3	1,09	1,06	1,28	1,30	1,06	1,30	0,79	1,78

3. Small area estimators

In table 2 the estimated coefficient of variation (CV) of direct estimates from HSHSUS for myocardial infarction are presented, as well as the total population and sampled units in each HSA. The table also contains the total number of hospitalized persons with primary diagnosis of suspected heart attack available from administrative source (*Schede di Dimissione Ospedaliera*, SDO).

Table 2. CVs of direct estimates in each HSA and auxiliary information from administrative source for myocardial infarction

HSA	CV	SDO	HSHSUS sample	Total population
AV/1	60,2	112	308	169658
AV/2	41,2	168	766	270542
BN/1	232,6	138	428	292829
CE/1	82,6	213	695	397884
CE/2	53,5	222	910	458979
NA/1	35,9	667	1461	1000470
NA/2	65,1	153	877	518780
NA/3	54,7	137	406	391297
NA/4	-	204	450	546399
NA/5	32,6	373	1229	642942
SA/1	81,6	198	797	321396
SA/2	46,3	311	675	499790
SA/3	58,9	216	669	271278

The figures in the table show that CVs of direct estimates are always greater than 30 percent and generally too large to be published or used for health care planning purposes. Although it is possible to make design choices based on optimal stratification and oversampling that would improve accuracy for many small areas (cfr. Marker, 2001), it will never be possible

to allocate sufficient sample sizes to all domains of interest, so indirect estimators will always be needed. When turning to indirect estimation techniques it is possible to consider two different classes of small area estimators (cfr. Rao, 2003): (i) the standard *design-based* estimators, which refer to implicit models, using, e.g., the known age/sex group totals as the only supplementary information (this class comprises the well known synthetic, composite and sample size dependent estimators), (ii) estimators that make explicit use of a suitable models which allow for between-area variation. Most of the models used may be regarded as special cases of the general mixed linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{y}$ is the vector of sample observations, $\mathbf{X}$ and $\mathbf{Z}$ are known matrices, $\boldsymbol{\beta}$ is a vector of unknown parameters. $\mathbf{u}$ represents the area-specific random effects and $\boldsymbol{\varepsilon}$ the errors of the model. The components $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are assumed to be distributed independently with means $\mathbf{0}$ and covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\varepsilon$ respectively. Depending on available auxiliary information it is then possible to fit different models. If area-specific auxiliary data $\mathbf{X}_i$ are available, area level models are obtained as (Fay and Herriot, 1979)

$$\hat{\theta}_i = \mathbf{X}_i^T \boldsymbol{\beta} + u_i + \varepsilon_i \quad i=1 \dots m \quad (3)$$

where $\hat{\theta}_i$ is the direct estimator of the parameter θ_i and ε_i is the sampling error.

If the available auxiliary variables $\mathbf{X}_{ij}$ are unit-specific, it is possible to fit the nested error regression model (Battese *et al.* 1988)

$$y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + u_i + \varepsilon_{ij} \quad i=1 \dots m, j=1 \dots n_i \quad \dots (4)$$

where the $\mathbf{X}_{ij}$ are available for each sampled element j in each small area i . In this case the specific area effects u_i are assumed to be *iid* random variables with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_u$, independent of ε_{ij} . Note that both with area-level and unit-level models it is possible to handle the spatial dependence of random small area effects and fit the model with or without spatial correlation among small areas.

References

- Battese, G.E., Harter, R.M. and 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, 28-36.
- Deville, J.C. and Särndal, C.E. (1992), "Calibration Estimators in Survey Sampling", *Journal of The American Statistical Association*, 87, 376-382.
- Fay, R.E.III 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*, 74, 269-277.
- Marker, D.A. (2001), "Producing Small Area Estimates from National Surveys: Methods for Minimizing Use of Indirect Estimators", *Survey Methodology*, 27, 183-188.
- Rao, J.N.K. (2003), *Small Area Estimation*, Wiley, New York.
- Singh M.P., Gambino J. and Mantel H.J. (1994), "Issues and Strategies for Small Area Data", *Survey Methodology*, 20, 3-22.