

A PROPOSAL FOR LONGITUDINAL SMALL AREA ESTIMATION¹

Stimatori Longitudinali per Piccole Aree: una Proposta

Maria Rosaria Ferrante, Silvia Pacei

Department of Statistics,

University of Bologna, Via Belle Arti n. 41, 40126 Bologna

ferrante@stat.unibo.it, pacei@stat.unibo.it

Riassunto: In questo lavoro si affronta il problema della stima per piccole aree in campioni ruotati, tipici delle indagini sulle forze di lavoro. I risultati ottenuti costituiranno la base di partenza per la costruzione di indicatori affidabili relativi a mercati del lavoro locali. A tale riguardo vengono proposte due procedure di stima che rappresentano un'estensione di uno stimatore suggerito da Fuller (1990). Uno studio basato su simulazioni evidenzia che, nei casi considerati, i due stimatori proposti sono più efficienti rispetto ad uno stimatore composto usualmente impiegato nella stima per piccole aree.

Keywords: Small Area Estimation, Linking Model, Rotating Samples.

1. Introduction

Statistical Agencies are often required to provide reliable estimates of parameters referring to the labour market in local areas. Nevertheless, the sample size of the official surveys within those areas is usually too small to obtain efficient results using direct survey estimators. Many studies have dealt with the small area estimation problem (Ghosh and Rao, 1994; Rao, 1999) and, more precisely, in Italy, many research project have been carried out to identify a suitable small area estimation strategy for evaluating small area labour force indicators, based on the Italian Labour Force Survey. However, although many surveys are conducted on partial overlap samples, such as labour force surveys, very few studies are reported in the literature about the possibility of exploiting repeated measurements on the same unit, in order to obtain further gains in efficiency. Taking into account survey repetition, Falorsi *et al.* (1998) have evaluated the performance of time series estimators (Singh *et al.*, 1994), finding that these estimators do not perform better than traditional estimators.

In view of the interest shown in Italy in the estimation of labour force indicators for local areas, we have considered the small area estimation problem for repeated surveys where a partial replacement of units from one occasion to the next is carried out in each small area, as happens for the Italian Labour Force Survey. More precisely, in dealing with the small area problem, we propose to take into consideration longitudinal observations of the same unit, through individual correlation. We suggest two strategies in order to derive small area estimators as extensions of the longitudinal estimators already proposed for partial

¹ The present paper is financially supported by Murst-Cofin (2001) “L'utilizzo di informazioni di tipo amministrativo nella stima per piccole aree e per sottoinsiemi della popolazione”.

overlap samples (Fuller, 1990). Unlike the time series estimators, that are heavily model-based, we would like to obtain a further gain in efficiency, both from individual temporal correlation and by specifying a model linking different small areas. We suggest two small area estimators for rotating samples, which combine gains in efficiency deriving from the longitudinal and spatial dimensions, called, respectively, two-stage longitudinal composite and one-stage longitudinal composite estimators. A simulation study is carried out to evaluate the gain in efficiency allowed by the proposed solutions, for different levels of autocorrelation between repeated measurements. The outline of the paper is as follows. The two strategies that we suggest are described in section 2, while their performances, based on the implemented simulation study, are compared in section 3.

2. Small area estimation in rotating samples

For the sake of simplicity, we consider the case of a rotating sample on two occasions ($t = 1, 2$), where n units are randomly selected from the population in $t = 1$ and some of them are randomly retained for occasion $t = 2$, while the remaining units are replaced. The parameters of interest are the small area population means. We will use ${}_d\bar{Y}_{t,i}'$ to indicate the direct estimator for the small area i ($i = 1, \dots, m$), associated with the units that are not common to both occasions (non panel), and ${}_d\bar{Y}_{t,i}''$ as the direct estimator associated with the units observed on both occasions (panel). The same notation will be used for the synthetic estimators (${}_s\bar{Y}_t'$ and ${}_s\bar{Y}_t''$) which may be generally defined as unbiased estimators for a larger area having acceptable standard errors. In order to simplify the notation, in this section every vector and matrix is referred to the i .th small area, even if it is not indicated.

2.1 - Two-stage longitudinal composite estimator

In the first stage of this estimation strategy a regression model $\bar{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is specified for each small area, where $\bar{Y}^T = [{}_d\bar{Y}_{1,i}' \quad {}_d\bar{Y}_{2,i}'' \quad {}_d\bar{Y}_{1,i}' \quad {}_d\bar{Y}_{2,i}'' \quad {}_s\bar{Y}_1'' \quad {}_s\bar{Y}_2'' \quad {}_s\bar{Y}_1' \quad {}_s\bar{Y}_2']$ is the vector of the elementary direct and synthetic estimates, and $\mathbf{X}$ is a dummy variables matrix indicating (assuming 1) the elementary estimates to use in the final estimates. To obtain the longitudinal direct and synthetic estimates, $\mathbf{X}$ has to be specified in the following form:

$$\mathbf{X}^T = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Moreover $\boldsymbol{\beta}^T = [\mu_{1,i} \quad \mu_{2,i} \quad \mu_1 \quad \mu_2]$ denotes the population means, respectively, for the small and large areas, at the two considered time instants, and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{\Sigma}$ contains variances of $\bar{Y}^T$ in the principal diagonal and covariance between estimators referring to different times and the same units. Remaining covariances are assumed to be 0. Deriving GLS estimates for $\boldsymbol{\beta}$ leads to the longitudinal synthetic and direct estimators

$\hat{\beta}^T = [\bar{Y}_{1,i} \quad \bar{Y}_{2,i} \quad \bar{Y}_1 \quad \bar{Y}_2]$. In the second stage, the following result will be used. Using a similar model and defining suitable $\mathbf{X}$ and covariance matrices, a usual (cross-sectional) small area composite estimator can be obtained from the direct and synthetic estimators. We consider the usual direct and synthetic estimates as the vector of the basic statistics, $\tilde{\mathbf{Y}}^T = [\bar{Y}_{1,i} \quad \bar{Y}_{2,i} \quad \bar{Y}_1 \quad \bar{Y}_2]$, for $t=1,2$, and

$$\tilde{\mathbf{X}}^T = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Then the synthetic and direct estimates are combined on the basis of the model $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\tilde{\beta} + \tilde{\epsilon}$, where $\tilde{\beta}^T = [\mu_{1,i} \quad \mu_{2,i}]$. Since the synthetic estimators can be heavily biased for some of the small areas, the MSE matrix of $\tilde{\mathbf{Y}}^T$ has to be considered in the GLS estimation procedure. The resulting GLS estimator of $\tilde{\beta}$ coincides with the usual small area composite estimator. In fact, this small area estimator is expressed as a weighted average of the direct and synthetic estimators, with optimal weights obtained by minimising the MSE of the composite estimator.

The above result is used in the second stage as follows. The longitudinal estimates in $\hat{\beta}$, obtained in the first stage, are considered as the elementary estimate, that is, $\tilde{\mathbf{Y}} = \hat{\beta}$. However, remembering that $\tilde{\Sigma} = \text{Var}(\tilde{\mathbf{Y}}) = \text{Var}(\hat{\beta})$, to obtain a small area composite estimator, the bias of the synthetic estimators has to be considered, as well as the covariance matrix $\tilde{\Sigma}$. Thus, in the diagonal elements of $\tilde{\Sigma}$ referring to the synthetic estimates, the MSE has to be considered instead of variance and, on the basis of the above result, $\tilde{\Sigma}$ has to be diagonalized to obtain a composite estimator.

Denoting with $\tilde{\Sigma}^*$ the transformed covariance matrix, the GLS two-stage longitudinal composite estimator is given by $\hat{\beta} = (\tilde{\mathbf{X}}^T \tilde{\Sigma}^{*-1} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \tilde{\Sigma}^{*-1} \mathbf{Y} = (\tilde{\mathbf{X}}^T \tilde{\Sigma}^{*-1} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \tilde{\Sigma}^{*-1} \hat{\beta}$.

2.2 - One-stage longitudinal composite estimator

The vector $\bar{\mathbf{Y}}^T$ of elementary direct and synthetic estimates, already introduced in section 2.1, is used to estimate directly the vector of parameters $\beta^T = [\mu_{1,i} \quad \mu_{2,i}]$, that is, the population means for the small area at the two considered time instants. In order to obtain longitudinal composite estimates using a one-stage strategy, the dummy matrix in the regression model $\bar{\mathbf{Y}} = \ddot{\mathbf{X}}\beta + \ddot{\epsilon}$ has to be specified in the following form:

$$\ddot{\mathbf{X}}^T = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Because of the bias of synthetic estimates in $\bar{\mathbf{Y}}^T$ the MSE matrix of $\bar{\mathbf{Y}}^T$ has to be considered, also in this case, in the GLS estimation procedure. Deriving GLS estimates for β allows us to obtain the longitudinal composite one-stage estimator $\hat{\beta}^T = [\bar{Y}_{1,i} \quad \bar{Y}_{2,i}]$.

3. The performance of the two estimators

In order to evaluate the properties of the proposed procedure in a simple setting, a finite population, subdivided into 5 equally weighted small areas, is simulated on two occasions, $t=1,2$, from a normal distribution, assuming a first order auto-regressive process with parameter ρ , and then considering different values for it (0.5, 0.7 and 0.9). Small area observations are generated for 5 small areas and for two different situations: the case in which the ordered small area means increase linearly, that is, the differences between the means are small, and a case in which the ordered small area means increase exponentially, that is, the differences between the means are high. A simple random sample is drawn from each small area population at $t=1$. Half of each sample is also observed on the second occasion and the other half is replaced with a simple random sample selected from the corresponding small area domain population. Therefore, a stratified sample design is assumed where strata correspond to small areas (for details of the simulation see Ferrante and Pacei (2002)). The procedure described has been repeated 1,000 times, since the MSE values of the estimators become quite stable after about 900 simulation runs.

To evaluate the performance of the estimators proposed, we compared their MSE to that of a usual small area estimator, which is the composite estimator that uses the mean of the whole sample as synthetic estimator. On the whole, the suggested estimators perform better than the composite one, and lead to a considerable gain in efficiency in almost all cases. Obviously, their efficiency increases when the autocorrelation coefficient increases, and when the difference between the small area and whole population means is low. The comparison between the two estimators proposed shows that neither of them always outperforms the other. Moreover, their performances may lead to different and probably less exciting results when parameters population have to be estimated from a sample. Hence further analysis and application to real data have to be carried out to establish a criterion that allow us to choose between them and to evaluate their efficiency in practice.

References

- Falorsi P.D., Falorsi S., Russo A. (1998), Small area estimation at provincial level in the Italian Labour Force Survey, *Journal of the Italian Statistical Society*, 1, 93-109.
- Ferrante M.R., S. Pacei (2002), Small area estimation in partial overlap samples, in *Qualità e informazione statistico-economica territoriale: aspetti del processo di formazione dei dati*, a cura di S. Biffignandi e G. Gozzi, F. Angeli, Milano, in corso di stampa.
- Fuller W. (1990), Analysis of Repeated Surveys, *Survey Methodology*, 16, 2, 167-180.
- Ghosh M., Rao J. N. K. (1994), Small Area Estimation: An Appraisal, *Statistical Science*, 9, 55-93.
- Rao J. N. K. (1999), "Some Recent Advances in Model-Based Small Area Estimation", *Survey Methodology*, 25, 2, 175-186.
- Singh A.C., Mantel H.J., Thomas B.W. (1994), Time series EBLUPs for Small Areas Using Survey Data, *Survey Methodology*, 20, 33-43.