Small Area Estimation of Inequality Measures
La stima per piccole aree di indicatori di disuguaglianza

Maria Rosaria Ferrante and Silvia Pacei

Abstract In order to estimate inequality measures at local level, small area estimation methods may be used to improve the reliability of estimates when the sample size is low. Small area models specified at area level, incorporate the design based estimates (direct estimates) as inputs, that are typically unbiased even though unreliable for small samples. Nevertheless, in the case of inequality measures, design based estimates are instead known to be biased for small sample sizes. In this work we focus on the search for a correction that can produce approximately unbiased direct estimators, taking into account the complexity of the survey design. We use data taken from the EU-SILC sample survey for Italy in 2013. Those modified estimators can then be used in small areas models.

Abstract *Allo scopo di stimare indicatori di disuguaglianza a livello locale, si possono impiegare metodi di stima per piccole aree per migliorare l'affidabilità delle stime quando la dimensione campionaria è piccola. I modelli per piccole aree specificati a livello di area, si basano su stimatori basati sul disegno (diretti), tipicamente corretti ma non affidabili per piccoli campioni. Gli stimatori degli indicatori di disuguaglianza sono invece distorti per piccoli campioni. L'obiettivo di questo lavoro è proporre una correzione che possa portare a stimatori diretti approssimativamente corretti, tenendo conto della complessità del disegno campionario. A questo scopo si usano i dati ottenuti per l'Italia dall'indagine* EU-SILC *del* 2013.

Key words: mean log deviation, complex sample survey, Fay-Herriot model.

[1] Maria Rosaria Ferrante; University of Bologna; email: maria.ferrante@unibo.it.
Silvia Pacei; University of Bologna; email: silvia.pacei@unibo.it.

1 Introduction

The increased interest for reliable information for restricted domain with reference to inequality measures is due to different reasons. One of the most relevant is to better plan policies to reduce inequality at local level. In this regards, an increasing gap in inequality and social exclusion among regions within the different EU member States has been observed in recent years. This issue is particularly relevant for Italy whose economic system is characterized by a strong territorial disparity.

Using data taken from the EU-SILC sample survey for Italy in 2013, we consider the estimation of inequality measures for the Italian provinces. Nevertheless the number of units sampled from many provinces is too low to provide reliable estimates using a "direct" estimator, that is an estimator calculated simply using the sample weights. This problem happens because EU-SILC survey is planned to provide reliable estimates for areas that are larger than those we are interested in.

To solve that problem we may resort to a small area estimation strategy. We consider area level models that incorporate the direct estimates as inputs. These estimates are typically obtained through unbiased estimators even though unreliable for small samples. Nevertheless, in the case of inequality measures, design based estimators are instead known to be biased for small sample sizes. The reason is that inequality measures can be written as ratios of random variables, both of which are estimated from the sample. They are thus biased in small sample, because the expected value of a ratio of random variables is not generally equal to the ratio of the expected values. The bias of the sample measure is $O\left(\frac{1}{n}\right)$, where n is the sample size.

In this work we focus on the search for a correction that can produce approximately unbiased direct estimators, taking into account the complexity of the survey design. Those modified estimators can then be used in small areas models.

We consider the class of generalized entropy (GE) measures, having the merit of satisfying the decomposability axiom, that allows to decompose the total inequality into the part due to inequality within areas and the part due to differences between areas. GE measures can be expressed as:

$$GE(\alpha) = \frac{1}{\alpha(\alpha-1)}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_j}{\bar{y}}\right)^{\alpha} - 1\right]; \quad j = 1, \dots, n; \quad \alpha \in [0,\infty) \qquad [1]$$

where $\bar{y}$ denotes the sample mean.

Specific special members of this family include Theil's mean log deviation ($\alpha = 0$), Theil's Index ($\alpha = 1$) and half the squared coefficient of variation ($\alpha = 2$). We start considering the mean log deviation for different reasons: *i*) it is used to study the "Inequality of opportunity" (Checchi and Peragine, 2010) with the purpose of assessing to what extent circumstances and efforts determine advantages; *ii*) it is particularly sensitive to changes in the tails of distribution, that are particular interesting in the case of income data; *iii*) it is found to be the less biased among the three indices mentioned (Breunig and Hutchinson, 2008).

2 Estimating the Theil's mean log deviation

We are interested in estimating the mean log deviation of the individual equivalized income, *Y*, for small areas indexed by $i = 1, \dots, m$. In the case of complex sample surveys, the direct estimator may be calculated using sample weights as follows:

$$ge(0)_i = \frac{1}{\hat{N}_i}\left[\sum_{j=1}^{n_i} w_j log\left(\frac{\bar{y}_i}{y_j}\right)\right]; \quad i = 1, \dots, m \qquad [2]$$

where $\bar{y}_i$ denotes the small domain sample mean calculated using sample weight, $\bar{y}_i = \frac{\sum_{j=1}^{n_i} w_j y_j}{\sum_{j=1}^{n_i} w_j}$, and $\hat{N}_i = \sum_{j=1}^{n_i} w_j$.

In the literature a few papers consider the small sample bias issue for inequality measures (see Breunig, 2001, 2002; Giles, 2005; Breunig and Hutchinson, 2008) and propose a correction, but only in the simple random sample context. Breunig and Hutchinson (2008), for example, write the GE measures as functions of the population mean, μ, and some other population functions and then derive corrections for the GE measures, based on a second-order Taylor's series expansion of the sample estimates around the population values.

Regarding the mean log deviation, they obtain the following result for the approximate bias:

$$ABias\big(ge(0)\big) = -\frac{1}{2}\mu^{-2}Var(\hat{\mu}) \qquad [3]$$

They suggest to estimate [3] from sample data and then subtract it from $ge(0)$ to obtain a bias approximately corrected inequality value.

They also warn about the fact that the correction tends to increase the variability of the estimator, and that the overall reliability of estimates have to be considered.

Extension of this bias correction to the weighted estimator in equation [2] is not trivial. We consider an heuristic solution by substituting μ with the weighted sample mean and $Var(\hat{\mu})$ with the estimate obtained using the standard procedure used by Eurostat for a two-stage stratified sample (Eurostat, 2013). In particular, in EU-SILC survey carried out in Italy a stratified sample of municipalities is selected in the first stage and, in the second stage, a sample of households is randomly selected from the municipalities included in the first stage. The largest municipalities are always included in the sample (therefore they are called auto-representative or AR), while the other ones are selected according on a stratified sample where strata are defined by the administrative regions and the number of inhabitants (non auto-representative municipalities or NAR). The procedure used for estimating $Var(\hat{\mu})$ involves two different methods for AR and NAR municipalities. In our case, both estimates of μ and $Var(\hat{\mu})$ are calculated at small area level.

3 Simulation study

We carry out a simulation study to assess both the magnitude of the bias of the non-corrected estimator and the effectiveness of the correction adopted to reduce that small sample bias. To this purpose we consider the EU-SILC sample as target population and then repeatedly select random samples from it. We prefer to base our study on the EU-SILC dataset, rather than use data generated under some distribution model, to have a more realistic view of the small area estimation problem considered.

We consider as small areas the administrative regions and repeatedly select 1,000 two-stage stratified samples, mimicking the sample strategy adopted in the EU-SILC itself: in the first stage, AR municipalities are always included in the sample, while a stratified sample of NAR municipalities are selected; in the second stage, a simple random sample of households is selected from each municipality included in the first stage. We consider two overall sampling rates, 1.5 and 3%, to better understand the extent of the problem and the effectiveness of the solution with reference to different sample sizes. In our simulation setting the small area sample size ranges from a minimum of 6 to a maximum of 28 for the 1.5% sample, and almost twice for the 3% sample. $ge(0)$ and its bias corrected version, from now on $geCorr(0)$, are calculated considering the individuals, as usual. Individual equivalized income is, by definition, the same for all members of the same household.

$ge(0)$ and $geCorr(0)$ are compared in terms of bias and accuracy using the average absolute relative bias (AARB) and the average absolute relative error (AARE):

$$AARB = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{1000}\sum_{r=1}^{1000}(est_{ri}/GE(0)_i - 1)\right| \quad [4.a]$$

$$AARE = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{1000}\sum_{r=1}^{1000}|est_{ri}/GE(0)_i - 1| \quad [4.b]$$

where est_{ri} denotes the value of an estimator (alternatively $ge(0)$ or $geCorr(0)$) obtained for the r.th simulated sample and i.th small area, and $GE(0)_i$ is the true small area mean log deviation.

Percentage values of indicators in [4.a] and [4.b] are reported in Table 1. Results show that the correction considered greatly reduces the bias of the non-corrected estimator, although the corrected estimates remain a little biased on average. On the other hand, with respect to the concern about the reduction of the overall reliability of the estimates due to the correction, we find instead a negligible increase in the accuracy indicator.

Table 1: Percentage performance measures based on the simulation study

	1.5% sample		***3% sample***	
	$ge(0)$	$geCorr(0)$	$ge(0)$	$geCorr(0)$
AARB%	15.9	4.0	7.9	2.6
AARE%	51.8	52.3	37.8	38.2

References

1. Breunig R. (2001), An almost unbiased estimator of the coefficient of variation, Economics Letters, 70, 15-19
2. Breunig R. (2002), Bias correction for inequality measures: an application to China and Kenia, Applied Economics Letters, 9, 783-786
3. Breunig R., Hutchinson D.L.A. (2008), Small sample bias corrections for inequality indices, in "New Econometric Modeling Research", William N. Toggins ed., Nova Science Publishers: New York
4. Checchi D., Peragine V. (2010), Inequality of Opportunity in Italy, Journal of Economic Inequality, 8(4), 429-450.
5. Giles D.E. (2005), The bias of inequality measures in very small samples: some analytic results, Econometric Working Paper EWP0514, ISSN 1485-6441, University of Victoria, Department of Economics
6. EUROSTAT (2013), Standard error estimation for the EU-SILC indicators of poverty and social exclusion, EUROSTAT methodologies and working papers