Small Area Estimation of Entropy Inequality Measures: a Comparison Between Alternative Distribution Models

Maria Rosaria Ferrante[1], Silvia Pacei[2]
[1]Department of Economics, Piazza Scaravilli 2, 40126 Bologna, Italy
[2]Department of Statistical Sciences, Via delle Belle Arti 41, 40126 Bologna, Italy

Abstract
Small area statistics on economic inequality are becoming important for better planning public regional policies. We focus on the estimation of entropy inequality measures in Italian provinces by using data taken from the EU-SILC sample survey for Italy. In EU-SILC survey the number of units sampled at provincial level is generally too small to obtain reliable estimates, and the use of small area estimation models is advisable. We consider small area models specified at area level that include the "direct" survey weighted estimators. In these models "direct" estimators are usually assumed to be unbiased and normally distributed. Nevertheless, in the case of inequality measures, design based estimators are known to be biased for small sample sizes. To solve this problem, we search for a correction that can produce approximately unbiased direct estimators. Moreover, due to the range of values that these estimators can assume and to the possible asymmetry of their distribution, the normality assumption could be inadequate in small area estimation models. In this regard we propose a small area model based on more flexible distributions as alternative to the normal one.

Key Words: Fay-Herriot model, Skewed distributions, Hierarchical Bayes

1. Introduction

The aim of this work is to propose a small area estimation strategy for the estimation of entropy inequality measures. The interest for this objective is due to the increment in the gap in inequality and social exclusion observed among regions, for example, within the EU member States. The availability of reliable information at local level may help to plan policies to reduce such inequality. This issue is particularly relevant for Italy, whose economic system is characterized by a strong territorial concentration of productive activities. Moreover, we have heard about the small area estimation problem above all with reference to poverty, but inequality mapping represents a different concept, which also deserves to be investigated.

Using data taken from the EU-SILC sample survey for Italy in 2013, we consider the estimation of Generalized Entropy (GE) measures for the Italian provinces. It is well known that there are many other possible indices candidate to measure inequality, for example the popular Gini index. However, the Gini index is positional transfer sensitive, then the change in Gini depends on the ranks of the donor and recipients, whereas the GE measures are transfer sensitive, that means that they reacts to transfers depending on donor and recipient income levels. Moreover, this class of measures has the merit of satisfying the additive 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. In the analysis of regional disparity this between-regions component is interesting for the analysis of regional inequality, as it suggests the relative importance of spatial dimension of inequality.

GE measures can be expressed as:

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

where $\bar{y}$ denotes the sample mean. In particular in this work we consider a specific member of this class of measures, the Theil's mean log deviation, from now on GE(0), of the individual equivalized income, obtained by setting $\alpha \to 0$. This index is found to be the least biased of its class in small samples (Breunig and Hutchinson, 2008).

The small area of interest are the Italian provinces. Nevertheless the number of units sampled from many provinces is too low to provide reliable estimates of GE(0) 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. Hence we have to resort to a small area estimation strategy.

To this purpose we consider area level models, which consist of two models: a sampling model, which connects the small area direct estimates $\hat{\theta}_i$ (where *i* denotes the small area) to the small area parameters θ_i, and a model linking the small area parameters to some small area specific auxiliary data $\boldsymbol{x}_i$ (Rao and Molina, 2015). To estimate the models we adopt a Hierarchical Bayesian approach. Small area models rely on the assumptions of unbiasedness for the direct estimates, and normality for both the direct estimates and the underling parameters. Such assumptions are both inadequate in our case. First of all, design based estimators of inequality measures are known to be biased for small sample sizes (Breunig and Hutchinson, 2008). 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. Secondly, the normality assumption is inadequate for asymmetric outcomes, when sample sizes are particularly low, and for limited-range outcomes.

We try to solve these problems looking for a correction for the bias of the direct estimator (Section 2) and, after having found a correction for the bias, looking for a more suitable distribution than the normal one for the corrected estimator (Section 3). The results obtained from the chosen distributions, skew-normal and skew-*t*, are reported in Section 4. Section 5 offers some conclusions and future research directions.

2. Correction for the Direct Estimator

2.1 Bias of the Direct Estimator

In the case of complex sample surveys, the direct estimator of the mean log deviation of the individual equivalized income, *Y*, for small areas, may be calculated by using the sample weights as follows:

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

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

In the literature a few papers consider the small sample bias issue for inequality measures and propose a correction, but only in the simple random sample context (for a review see Ferrante and Pacei, 2019). 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 to 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.

2.1 Simulation Study to Evaluate the Correction for the Bias

To evaluate the effectiveness of the correction adopted we carry out a design-based simulation study. In this simulation we consider the EU-SILC sample as the target population and the administrative Regions as small areas. We prefer to base our study on the EU-SILC dataset, rather than use data generated under some distribution models, to have a more realistic view of the small area estimation problem considered.

We repeatedly select 1,000 two-stage stratified samples, mimicking the sample strategy adopted in the survey 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 for the corrected and non-corrected estimators (simulation study based on EU-SILC survey data)

	1% 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

Looking at the Relative Bias obtained for the corrected and non-corrected estimators for each small area (Figure 1), and observing the relationship between the bias and the small area sample size, we can notice that the bias of the non-corrected estimator is negative, decreases as the sample size increases, and is small but not zero for large sample sizes. Moreover, the reduction of the bias provided by the correction is noticeable.

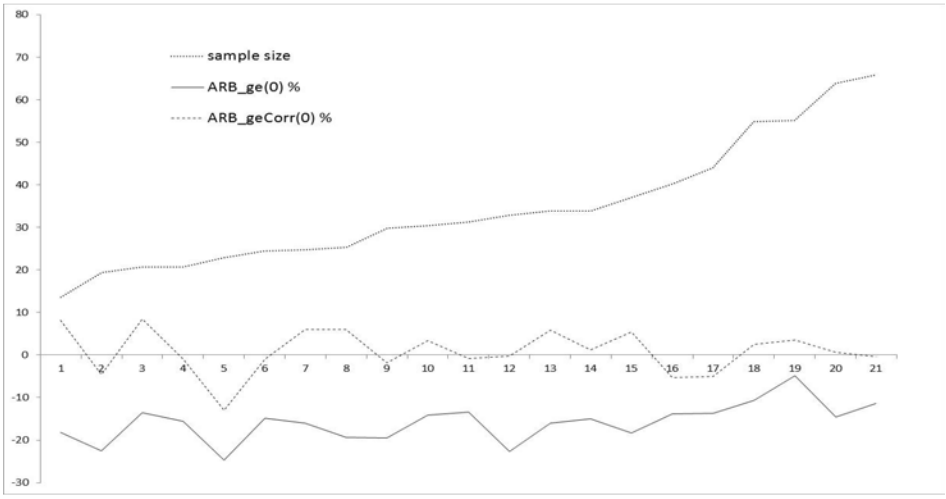

Figure 1: Relative Bias for *ge(0)* and *geCorr(0)* estimators - 1.5% sample

3. Small Area Model: Alternative Distributional Assumptions

3.1 Analysis of the Empirical Distribution of the Corrected Estimator

It is well known that the distribution of the equivalized income is characterized by a positive skewness. However, little is known about the distribution of entropy inequality estimators in small samples. To analyse the empirical distribution of the corrected estimator proposed, in small areas, we use the simulation study described in the previous Section. We consider the 3% sample. Looking at the histogram by region (Figure 2), the distribution of the corrected estimator appears skewed in all regions, with different degrees of skewness. The skewness, as expected, is always positive.

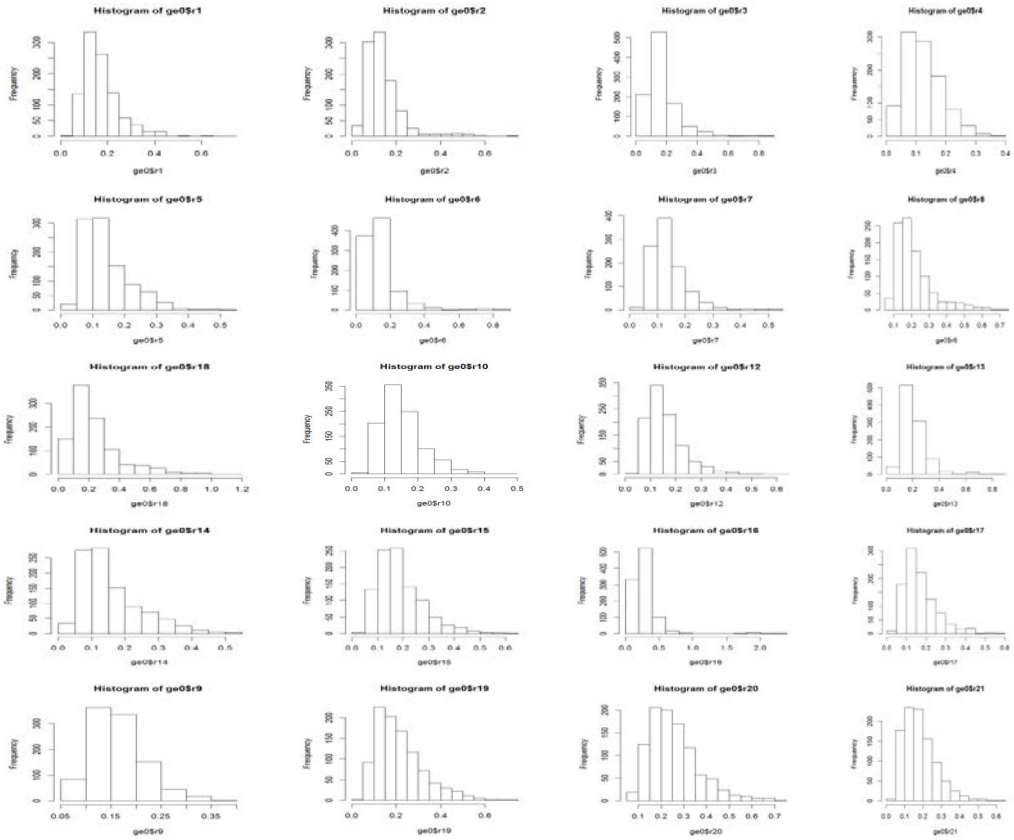

Figure 2: Histograms of *geCorr(0)* empirical distribution in Italian regions (3% sample).

3.2 Skew-normal and skew-*t* models

In the presence of skewed data, the assumption of normality at the sampling level of a small area level model cannot be justified invoking the central limit theorem when dealing with small sample sizes. To take into account the asymmetry of data, we relax the normality assumption of the most popular so-called normal–normal model (Fay and Herriot, 1979) by adopting a skew normal or a skew t distribution in the sampling models. We consider an asymmetric distribution only in the sampling model, because we noticed in a previous work (Ferrante and Pacei, 2017) that the specification of a skewed distribution also in the linking model tends to have a negligible effect on the results.

3.2.1 Skew-normal and skew-t distributions

The class of skew-normal distributions proves to be quite useful in modelling real data sets and enjoys remarkable properties in terms of mathematical tractability (Azzalini, 1985; Azzalini and Capitanio, 1999). Moreover, the skew-normal specification offers some other advantages with respect to other non-symmetric distributions as, for example, the inclusion of the normal distribution as a special case.

According to the definition of Azzalini (1985), *Y* is a skew-normal variable with location parameter ξ, scale parameter ϕ and shape parameter λ, Y~*SN*(ξ, ϕ^2, λ), if it has the following pdf

$$\frac{2}{\phi}\varphi(z)\Phi(\lambda z), \qquad z = \frac{y-\xi}{\phi} \qquad [5]$$

where $\varphi(.)$ and $\Phi(.)$ are the density and the distribution function, respectively, of the *N*(0,1) distribution. The mean and the variance of the skew-normal distribution are:

$$\mu = \xi + \sqrt{\frac{2}{\pi}}\phi\delta \qquad \sigma^2 = \phi^2\left(1 - \frac{2}{\pi}\delta^2\right) \qquad [6]$$

where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

However, the presence of a single parameter to regulate the density shape in the skew normal distribution could not be sufficient to handle adequately the very diverse types of situations met in practical work (Arellano-Valle and Azzalini, 2013). In this regard, a more flexible distribution than the skew-normal one is given by the skew-*t* distribution, which allows the capture of heavy tailed data. The density function of the skew-*t* distribution is:

$$\frac{2}{\phi}t(z;\tau)\mathrm{T}\left(\lambda z\sqrt{\frac{\tau+1}{\tau+z^2}};\tau+1\right), \qquad z = \frac{y-\xi}{\phi} \qquad [7]$$

where $t(.;\tau)$ and $\mathrm{T}(.;\tau)$ are the density and the distribution function, respectively, of the Student's *t* distribution with τ degrees of freedom. The mean and the variance of the skew-*t* distribution are given by the same functions showed for the skew-normal distribution (equations [6]).

3.2.2 Skew-normal and skew-t area level models

In the context of area level model-based estimation and of the Bayesian framework for inference, Ferraz and Moura (2011) tackled the problem of skewness by assuming a skew normal distribution at the sampling model level. Ferrante and Pacei (2017), dealing with correlated outcomes, proposed a multivariate skew normal area level model, where a multivariate skew normal model is specified both in the sampling and linking models.

In the skew-normal area level model considered in this work, the corrected estimator of the mean log deviation for area *i*, $\hat{\theta}_i$ = *ge(0)corr*$_i$, is supposed to be skew normally distributed with location parameter θ_i, scale parameter ϕ_i and shape parameter λ_i:

$$\hat{\theta}_i|\theta_i,\phi_i,\lambda,n_i \sim SN(\theta_i,\phi_i,\lambda_i) \qquad i=1,\dots,m. \qquad [8]$$

Each shape parameter λ_i is set equal to a parameter common for every area λ, divided by the square root of the sample size in the small area (Gupta and Kollo, 2003), so that, when the sample size increases, the shape parameter tends to 0 and the skew-normal distribution tends to the normal one:

$$\lambda_i = \lambda/\sqrt{n_i} \qquad [9]$$

The scale parameter, ϕ_i, can be obtained as a function of the variance of the direct estimates and the shape parameter, according to the relationship between the variance and the parameters in the skew-normal distribution:

$$\phi_i = V(\hat{\theta}_i)/\left(1-\frac{2}{\pi}\cdot\delta_i^2\right) \qquad \delta_i^2 = \lambda_i/\sqrt{1+\lambda_i^2} \qquad [10]$$

In the linking model, parameter θ_i is supposed to be normally distributed with mean μ_i, which is a linear function of the area level auxiliary variables $\boldsymbol{x}_i$, and common variance σ_v^2:

$$\theta_i|\mu_i,\sigma_v \sim N(\mu_i,\sigma_v^2) \qquad [11a]$$
$$\mu i = \boldsymbol{x}_i^T\boldsymbol{\beta} \qquad [11b]$$

Our parameter of interest is the expectation of the distribution of $\hat{\theta}_i$, under the model described, which, because of the properties of the skew-normal distribution, is given by a function of the parameters of the sampling distribution:

$$\theta_i^* = \theta_i + \sqrt{\frac{2}{\pi}}\phi_i\delta_i \qquad [12]$$

In the skew-*t* area level model, the corrected estimator is supposed to follow a skew-*t* distribution with location parameter θ_i, scale parameter ϕ_i and shape parameter λ_i:

$$\hat{\theta}_i|\theta_i,\phi_i,\lambda,n_i \sim St(\theta_i,\phi_i,\lambda_i,n_i-1) \qquad [13]$$

The same assumptions of the skew-normal model [9]-[10] apply to the parameters of the skew-*t* model. The degrees of freedom allow for the convergence to the skew-normal distribution. In the linking model the normal distribution is again assumed. The parameter of interest is given by the same function of the sampling model parameters reported in [12].

As it is customary, we assume that the variances of the direct estimates, $V(\hat{\theta}_i)$, are known, and we substitute them with their respective estimates. These estimates are obtained using a Bootstrap strategy, that is by repeatedly selecting random samples with replacement from the survey sample by province, calculating the corrected estimates for each replication by province, and then calculating the variance of these estimates by province.

In the linking model we include three auxiliary variables: the proportion of income tax returns on the total population, the mean income of the population and an indicator of the

aging population rate given by the ratio between the number of over 64-year-olds and the population between 15 and 64 years old. We obtain this information using data available from the registry offices and the tax archives.

To estimate the models we adopt a Hierarchical Bayesian approach implemented by using MCMC computational methods. To this purpose we use OpenBugs program. This program does not take the described distributions into consideration, therefore we explicitly write the density formulas into the BUGS code, using what is known as "the trick for specifying new distributions" (Spiegelhalter et al., 2003).

To complete the specification of our Bayesian model we use non informative priors for the variance and the regression coefficients in the linking model. Only for the shape parameter we presume a positive value, and specify a normal distribution truncated at zero:

$$\beta_k \sim N(0, B), \quad \sigma_v^{-1} \sim Gamma(a_v, b_v), \quad \lambda \sim TN_{[0,\infty]}(0, D) \qquad [14a]$$

$$(k = 1,2,3; B = 0.0001; a_v = 0.001, b_v = 0.001, D = 0.0001) \qquad [14b]$$

4. Results

To evaluate the performance of the models considered we compare them with the Fay-Herriot model. The comparison is carried out in terms of fit of the data and gain in efficiency provided by the small area estimators compared with the direct estimator. For the fit of the data we use the Logarithm of the Pseudo Marginal Likelihood (LPML), while we use the percentage Coefficient of Variation Reduction (CVR) of small domain model estimator (HB) versus the direct one (*dir*) to measure the gain in efficiency:

$$CVR_i = 100 \cdot (1 - CV_i^{HB} / CV_i^{dir}) \qquad [15]$$

Results are reported in Table 2.

Table 2: Performance measures: LPML and summaries for the CVR of the HB estimators versus the direct estimator (EU-SILC sample survey)

		N-N (Fay-Herriot)	*SN-N*	*St-N*
LPML		189.9	237.0	273.6
CVR_i	Mean	37.5	38.2	41.3
	Median	36.4	35.9	40.7
	25° percentile	26.5	20.5	26.7
	75° percentile	47.0	54.0	58.4

As models with the highest LPML are better supported by the data, the skew-*t* model shows the best fit, followed by the skew-normal model, both preferable to the Fay-Herriot model. Regarding the gain in efficiency, it is relevant for all the small area models and in particular for the skew-*t* model. The reduction of the coefficient of variation for the skew-*t* model is more than 40% on average and on median.

To better understand how much the small area estimates differ from the direct estimates we carry out a graphical analysis. Figure 3 shows the model-based estimates versus the

direct estimates. For the skew-normal model and even more for the skew-*t* model the points are closer to the bisector (the continuous line) than for the Fay-Herriot model. This suggests that the skew-normal and skew-*t* small area estimates are approximately design-unbiased, even though we can observe a slight shrinkage for the highest estimates.

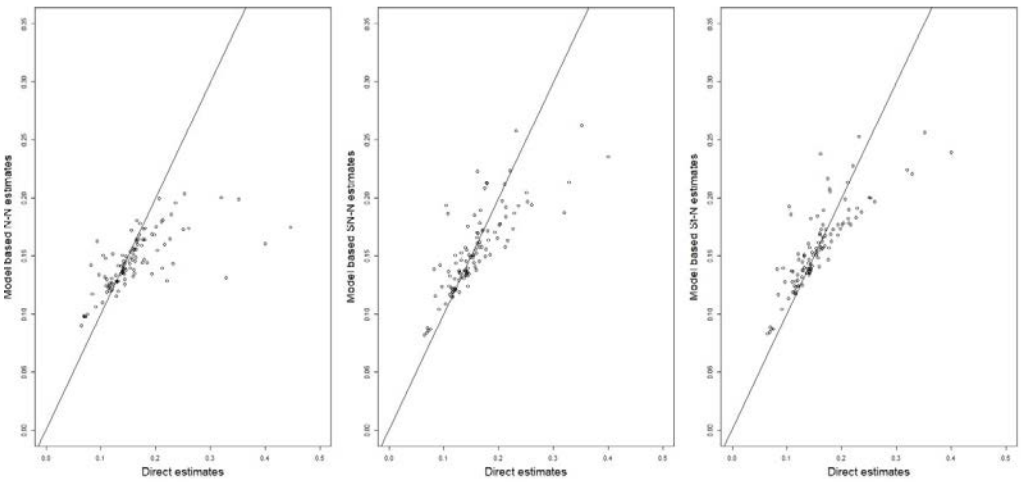

Figure 3: Direct estimates *versus* model-based estimates

Finally, in Figure 4 small areas are plotted according to the coefficient of variation obtained for the skew-*t* estimates and that obtained for the direct estimates. The coefficient of variation of the skew-*t* estimates appears markedly smaller than that of the direct estimates for most of the small areas.

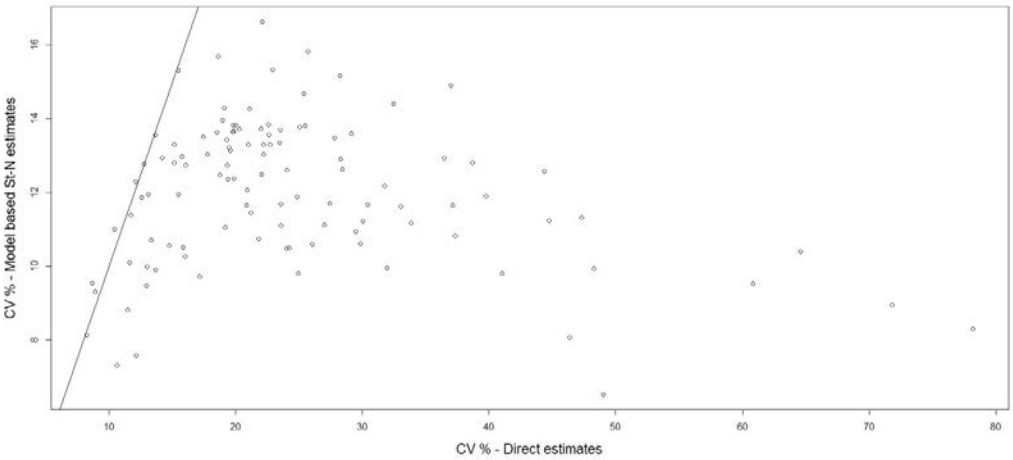

Figure 4: CV% comparison - model-based skew-*t* estimates *versus* direct estimates (continuous line = bisector)

4. Conclusions

This work proposes a small area estimation strategy for the estimation of entropy inequality measures. The strategy includes the adoption of a correction for the bias of the

direct estimator, and the choice of a skewed distribution to be used in the small area level model. The results obtained for the correction for the bias are promising, as well as those obtained from the specification of an asymmetric distribution in the small area model. However, it is necessary to carry out a simulation study to better understand the properties of the strategy proposed. Furthermore, since the analysis of a set of inequality measures is usually required, the joint estimation of several measures through multivariate models should be addressed.

References

Arellano-Valle, R.B., and A. Azzalini. 2013. The centred parameterization and related quantities of the skew-*t* distribution. *Journal of Multivariate Analysis*. 113. 73-90.

Azzalini, A.. 1985. A class of distributions which includes the normal ones. *Scand. J. Statist.*. 12. 171–178.

Azzalini, A., and A. Capitanio. 1999. Statistical application of the multivariate skew normal distribution. *J. R.Statist. Soc.* B. 61. 579–602.

Breunig, R., and D.L.A. Hutchinson. 2008. Small sample bias corrections for inequality indices, in *New Econometric Modeling Research.* William N. Toggins ed., Nova Science Publishers: New York.

EUROSTAT. 2013. Standard error estimation for the EU-SILC indicators of poverty and social exclusion. EUROSTAT methodologies and working papers.

Fay, R., and R. Herriot. 1979. Estimates of income for small places: an application of James–Stein procedures to census data. *J. Am. Statist. Ass.*. 74. 269–277.

Ferraz, V. R. S., and F. A. S. Moura. 2011. Small area estimation using skew normal models. *Computnl Statist.Data Anal.*. 56. 2864–2874.

Ferrante, M.R., and S. Pacei. 2017. Small domain estimation of business statistics by using multivariate skew normal models. *J. R.Statist. Soc.* A. 180. 1057–1088.

Ferrante, M.R., and S. Pacei. 2019. Small Sample Bias Corrections for Entropy Inequality Measures. Biostat Biometrics Open Acc J. 2019; 9(3): 555765. DOI: 10.19080/BBOAJ.2019.09.555765.

Gupta, A. K. and T. Kollo. 2003. Density expansion based on themultivariate skewnormal distribution. *Sankhya*. 66. 821–835.

Rao, J. N. K., and I. Molina. 2015, *Small Area Estimation*. 2nd edn. Hoboken: Wiley

Spiegelhalter, D. J., N.G. Best, B. P. Carlin, and A.van der Linde. 2002. Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B. 64. 583–639.