

Estimation of poverty indicators at the small area level in Italy¹

Caterina Giusti

Dept. of Statistics and Mathematics Applied to Economics, University of Pisa, e-mail:
caterina.giusti@ec.unipi.it

Stefano Marchetti

Scuola Superiore Sant'Anna of Pisa, e-mail: stefano.marchetti@sssup.it

Monica Pratesi

Dept. of Statistics and Mathematics Applied to Economics, University of Pisa, e-mail:
m.pratesi@ec.unipi.it

Abstract: Available data to measure poverty and living conditions in Italy come mainly from sample surveys, such as the Survey on Income and Living Conditions (EU-SILC). However, these data can be used to produce accurate estimates only at the national or regional level. To obtain estimates referring to smaller unplanned domains small area methodologies can be used. The aim of this paper is to provide a general framework in which the joint use of large sources of data, namely the EU-SILC and the Population Census data, can fulfill poverty and living conditions estimates for Italian Provinces and Municipalities such as the Head Count Ratio and the quantiles of the household equivalised income.

Keywords: poverty mapping, M-quantile models, bootstrap methods.

1. Introduction

In many different application fields, National Statistical Institutes provide data coming from large national and transnational surveys. Often, these data are integrated with administrative data sets for purposes of validation. This is the case with the EU-SILC survey in Europe. The principal aim of this survey is to fulfill timely and comparable estimates on income and living conditions in the EU countries, both in a cross-sectional and longitudinal perspective. In Italy, ISTAT integrates data coming from the EU-SILC survey with those from the Italian Social Security Service and others administrative data sets mainly to correct for nonresponse and under-report of income values. Thus, the final EU-SILC data sets represent a complete and valuable source of information to produce poverty and living condition estimates in Italy.

However, EU-SILC data can be used to produce accurate estimates only at the NUTS 2 level (that is, regional level). Thus, to satisfy the increasing demand from official and private institutions of statistical estimates on poverty and living conditions referring to smaller domains (LAU 1 and LAU 2 levels, that is Provinces and Municipalities), there is the need to resort to small area methodologies.

¹The paper is financially supported by the European Project SAMPLE “Small Area Methods for Poverty and Living Condition Estimates”, European Commission 7th FP - www.sample-project.eu, and by the Research Project PRIN2007 “Inference in the presence of imperfect auxiliary information: sampling from elusive populations and estimation for unplanned domains”.

Small area estimation techniques are employed when sample data are insufficient to produce accurate direct estimates in the domains of interest. The idea of these methods is to use statistical models to link the survey variable of interest with covariate information that is also known for units not in the sample. Among these models linear mixed models (Rao, 2003) are the most popular. A novel approach is represented by M-quantile models (Chambers and Tzavidis, 2006).

In this paper we focus on the estimation of poverty measures, such as Laeken indicators, at the small area level in Tuscany. For this purpose we combine data coming from the EU-SILC survey 2004 with those from the Population Census 2001. Besides poverty indicators, our objective include the estimation of the cumulative distribution function of the household equivalised income, since the estimation of its quantiles can fulfill more in deep knowledge on the living conditions in the small areas. For this purpose we propose the use of a bias adjusted estimator of the cumulative distribution function and a bootstrap technique for the estimation of the mean squared error for the bias adjusted estimator.

2. Small area methods for poverty estimates

The knowledge of the cumulative distribution function of an income variable in a given area of interest represents an important source of information for the living conditions in that area. From the cumulative distribution function of the household disposable income many quantities (e.g. the median income, the income quantiles) and indicators (e.g. the Head Count Ratio or *at-risk-of-poverty-rate*) can be computed.

Let $\mathbf{x}_i$ be a known vector of auxiliary variables for each population unit i in small area j and assume that information for the variable of interest y_{ij} , the household disposable income on an equivalised scale for unit i in small area j , is available only for the units in the sample. The target is to use these data to estimate the cumulative distribution function of the household income, which can be defined as:

$$F_j(t) = N_j^{-1} \left(\sum_{i \in s_j} I(y_{ij} \leq t) + \sum_{i \in r_j} I(y_{ij} \leq t) \right) \quad (1)$$

where N_j is the number of population units in small area j , s_j denotes the n_j sampled units in area j and r_j the remaining $N_j - n_j$ not sampled units. The y_{ij} for the r_j units in each area j need to be predicted under a given small area model. For this purpose we consider linear mixed models, and in particular the unit level nested error regression model (Battese *et al.*, 1988), and M-quantiles models, an approach to small area estimation based on the quantiles of the conditional distribution of the variable of study y given the covariates ((Chambers and Tzavidis, 2006). In more detail, the q th M-quantile $Q_q(x; \psi)$ of the conditional distribution of y given x satisfies:

$$Q_q(\mathbf{x}_{ij}; \psi) = \mathbf{x}_{ij}^T \beta_\psi(q) \quad (2)$$

where ψ denote the influence function associated with the M-quantile. For specified q and continuous ψ , an estimate of $\hat{\beta}_\psi(q)$ of $\beta_\psi(q)$ is obtained via an iterative weighted least squares algorithm. Unlike linear mixed models, M-quantile models do not depend on strong distributional assumptions and they are robust against outlying area values.

However, by simply plugging-in the values predicted under the M-quantile or linear mixed model in (1), using the so called *naïve* estimator, we obtain bias estimates of the mean of y in the small areas. A possible solution is represented by the use at the small area level of

the bias adjusted estimator of the cumulative distribution function proposed by Chambers and Dunstan (1986) (hereafter CD). When (2) holds, this estimator for small area j is:

$$\hat{F}_j^{CD}(t) = N_j^{-1} \left\{ \sum_{i \in s_j} I(y_{ij} \leq t) + n_j^{-1} \sum_{i \in r_j} \sum_{k \in s_j} I\left\{ [\mathbf{x}_{ij}^T \hat{\beta}(\hat{\theta}_j) + (y_{kj} - \hat{y}_{kj})] \leq t \right\} \right\} \quad (3)$$

where $\hat{y}_{kj} = \mathbf{x}_{kj}^T \hat{\beta}(\hat{\theta}_j)$ is a linear combination of the auxiliary variables and $\hat{\theta}_j$ is an estimate of the average value of the M-quantile coefficients of the units in area.

Analytic estimation of the standard error of estimator (3) is complex. A linearization-based prediction variance estimator is defined only when the estimator of interest can be written as a weighted sum of sample values, which is not the case with (3). As a consequence, we adapt to the small area problem the bootstrap procedure suggested by Lombardina *et al.* (2003). In particular, we define a set of bootstrap schemes that resample residuals from an M-quantile model fit. Using (one of) these schemes, we generate a bootstrap population, from which we draw bootstrap small area samples.

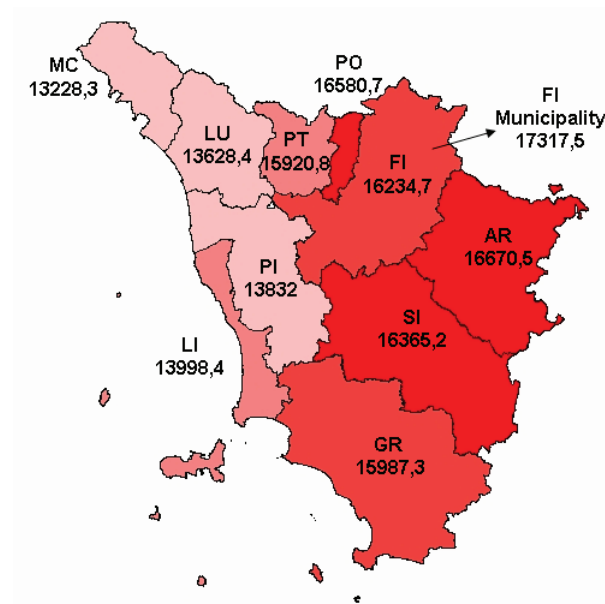
3. Estimation of poverty measures in Tuscany

Tuscany Region is a planned domain for which EU-SILC estimates are published, while the Provinces and Municipalities are unplanned domains. Direct estimates at the Province level may therefore have large errors and they may not even be computable at Municipality level, thereby requiring resort to small area estimation techniques. Data sources for the present application are the 2004 EU-SILC survey (for the n_j sampled units in area j) and the 2001 Population Census of Italy, with a total of 1388260 individuals in Tuscany (for the not sampled r_j units in area j). In 2004 the EU-SILC regional sample size in Tuscany was of 1751 households; 54 municipalities were included in the sample. The small areas of interest are the 10 Tuscany provinces, with sample sizes n_j ranging from 70 (Province of Grosseto) to 545 (Province of Firenze). Due to the large sample size in the Province of Firenze and to the differences characterizing that territory, we consider the Municipality of Florence, with 178 units out of 545, as a stand-alone small area. The characteristic of interest y is the household disposable income, equivalised according to the Eurostat guidelines. As a preliminary result, we are interested in estimating the small area mean of y in the small areas. The following auxiliary variables are known for each unit in the population and have resulted significant in the models for income:

- size of the household in terms of the number of components of the household i in the small area j (integer value);
- age of the head of the household i in the small area j (integer value);
- years in education of the head of the household i in the small area j (integer value);
- working position of the head of the household i in the small area j (employed/unemployed in the previous week);
- tenure status of household i in the small area j (owner/tenant).

Figure 1 represents the map of Tuscany with the values of the median household equivalised income in each Province estimated using the CD M-quantile predictor. As we can see, there are differences between the estimates: the richest areas are the Municipality of Florence and the Provinces of Prato, Arezzo e Siena (darkest color); the poorest the Provinces of Pisa, Lucca and Massa Carrara (lighter color).

Figure 1: *Median household equivalised income.*



To better describe and compare the living conditions in these areas, different estimates such as the Head Count Ratio and the quantiles of the income will be also considered, together with an estimate of the their standard error using the proposed bootstrap approach.

References

- Battese G., Harter R. and Fuller W. (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.
- Chambers R. and Dunstan R. (1986) Estimating distribution functions from survey data, *Biometrika*, 73, 597–604.
- Chambers R. and Tzavidis N. (2006) M-quantile models for small area estimation, *Biometrika*, 93, 255–268.
- Lombardia M., Gonzalez-Manteiga W. and Prada-Sanchez J. (2003) Bootstrapping the Chambers-Dunstan estimate of finite population distribution function, *Journal of Statistical Planning and Inference*, 116, 367–388.
- Rao J. (2003) *Small Area Estimation*, Wiley, New York.