



Policy, Welfare and Financial Resources

The Impact of the Crisis on Territories

edited by

Alessandra Coli
Barbara Pacini
Elettra Stradella

P  S A
UNIVERSITY
PRESS

Small Area Methods for Estimating Local Poverty Indicators

C. GIUSTI, S. MARCHETTI

The estimation and dissemination of local (sub-regional) welfare indicators all over the European Union has become one topic of primary interest. These local indicators are essential for policy makers in planning and verifying welfare interventions. To estimate these indicators for sub-regions of the survey population, data coming from national official surveys are usually employed. However, if the sample from a sub-region is small, then a traditional design-based survey estimator may have unacceptably large variance. These sub-regions are commonly known as small areas. Different approaches to small area estimation are possible: direct estimators, indirect estimators, model-based estimators. Direct estimators use only area-specific units, so when the sample size is small, they usually produce unreliable estimates due to the large variance. An estimator is called indirect if a reliable direct estimator for a large area is used to derive indirect estimators for the small areas under the assumption that the small areas have the same characteristics as the large area. This estimator has usually a smaller variance compared to that of a direct estimator. Model-based estimators are indirect estimators based on small area models, which are explicit linking models based on random area-specific effects that account for between area variation beyond that is explained by auxiliary variables included in the model. Model-based estimators can be classified into two types: area-level models, relating small area direct estimates to area-specific auxiliary variables and unit-level models, relating the unit values of a study variable to unit-specific auxiliary variables available for all the population. In this chapter, we estimate the incidence and the intensity of relative poverty in the Italian provinces using data from the EU-SILC 2013 survey, the Income Tax Office data base and Population Census 2011 by means of area level models.

1 Introduction

In recent years, there has been an increase worldwide in the demand for poverty and living conditions estimates, as such data can help in planning the local policies which are aimed at decreasing poverty and social exclusion. In September 2015, the UN General Assembly approved the Sustainable Development Goals (SDGs), which provided a new policy framework worldwide with the purpose of ending all forms of poverty, fighting inequalities and tackling climate change, while ensuring that no one be left behind. Eurostat (2016) provided a first limited number of indicators which are relevant for the EU which aim to capture the broader ambition of each SDG, highlighting the importance of computing indicators, not only at the national level but at the regional level as well. The main reason is that in Italy, social protection benefits and allowances are often managed by local administrations, such as Provinces and Municipalities. Moreover, in times of increasing financial difficulties and reduction of public funding, welfare systems are subject to several transformations which may impact the spatial quality and homogeneity of the offered services. Thus, the local dimension is essential when the interest is in analyzing the living conditions as well as the impact on social protection policies, especially in times of financial crisis.

The aim of this chapter is to estimate two poverty indicators, namely the incidence and the intensity of poverty of Italian households at the provincial level. Our focus is therefore on the demand for social protection from households, with a special interest on poverty. To compute these indicators, we use data from the EU-SILC survey, an annual survey dedicated to the collection of comparable longitudinal and cross-sectional data at the European level concerning income, poverty, social exclusion and living conditions. This survey is currently one of the primary sources of information in Europe to obtain estimates on the demand for social protection services from households and individuals. However, data from the EU-SILC can be used to compute reliable indicators at the national level or at the macro-regional level corresponding to NUTS 1 definition in the nomenclature of territorial units for statistics of the EU. To compute the indicators of interest at a more detailed geographical level - for example at LAU 1 or 2 levels (e.g. Provinces and Municipalities in Italy) - it is necessary to resort to appropriate methodologies, since the sample size is usually too small (or even equal to zero) to obtain reliable direct estimates (estimates based only on the information coming from the EU-SILC survey).

In this chapter we use the appropriate methodology, namely area level

small area models, to estimate the Head Count Ratio (HCR, also known as the at risk of poverty rate, ARPR) and the Poverty Gap (PG) of households living in the 110 Italian provinces. Nowadays, poverty is recognized as a multidimensional concept (Whelan et al., 2014), which includes both monetary and non-monetary deprivation dimensions. Among available indicators of monetary poverty, the HCR and PG, which belong to the poverty measures first introduced by Foster et al. (1984), have primary relevance, as they have been indicated as dashboard indicators by the EC Social Protection Committee (EC, 2012).

With the poverty line denoted by t , the Foster et al. (1984) (FGT) poverty measures for a small area i are defined as:

$$F_{\alpha i} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{t - y_{ji}}{t} \right)^{\alpha} I(y_{ji} \leq t). \quad (1)$$

In formula (1), the poverty line t is the level of income which defines the state of poverty (households with incomes below t are considered poor), y is a measure of income for household j , N_i is the number of households in area i , I is the indicator function (equal to 1 when $y_{ji} \leq t$ and 0 otherwise) and α is a “sensitivity” parameter. When $\alpha = 0$, $F_{\alpha i}$ is the *Head Count Ratio* whereas when $\alpha = 1$, $F_{\alpha i}$ is the *Poverty Gap*. The HCR measures the incidence of poverty while PG measures its intensity.

Small area methodologies have been used in several studies to compute poverty indicators using data from the EU-SILC survey (see for example Giusti et al. (2012)). Since the availability of unit-level data having a fine territorial reference is not a trivial matter, as best explained in section 2, small area models defined at the area level can represent a flexible method to obtain reliable local estimates.

Finally, we remark that the small area method explained in this chapter could be used to estimate other indicators of interest at the local level, using EU-SILC data, e.g. the HCR, including or excluding social cash transfers and the HCR broken down with respect to socio-demographic characteristics.

2 Methods

Data obtained from surveys are often used to estimate characteristics for subsets of the survey population. If the sample from a subset is small, then a traditional design-based survey estimator may have an unacceptably large variance. These subsets are commonly known as small areas.

A wide range of methods have been used in literature to obtain reliable small-area estimates (Pfeffermann, 2013; Rao and Molina, 2015), mostly model-based estimators. The applicability of small area model-based methods depends on data availability. From this point of view, we can classify small area techniques into two main approaches: unit-level and area-level. The unit-level approach requires survey micro-data availability and some of the variables collected in the survey have to be known at the population-level - i.e. obtained from population micro-data - for each small area. The area-level approach can be applied when direct estimates, their MSEs and aggregate auxiliary information are available for those small areas of interest.

In the context of monetary poverty indicators, namely HCR and PG, taxable income represents the auxiliary variable mostly correlated with target variables. Unfortunately, taxable income micro-data are not accessible, thus making the unit-level approach impractical. Therefore we referred to the area-level approach.

The Fay and Herriot (1979) estimator is the “industry standard” for the area-level approach. However, when spatial information is available and when there is evidence of spatial autocorrelation in the target variable, the use of spatial model-based area-level estimators may be appropriate. In literature, there are three main area-level estimators which use spatial information: i. the spatial-EBLUP (Petrucci and Salvati, 2006; Singh et al., 2005); ii. the non-parametric EBLUP (Giusti et al., 2012); iii. the non-stationary EBLUP (Chandra et al., 2015). A brief review of the method used in section 3 is presented hereunder.

Let us assume that there are m small areas of interest and that θ_i represents the population characteristic of interest in area i , such as a mean, a proportion or a percentile. A survey provides a direct estimator $\hat{\theta}_i^{dir}$ of θ_i for some or all of the small areas. In our case $\hat{\theta}_i^{dir} = \hat{F}_{0i}^{dir}$ for the HCR and $\hat{\theta}_i^{dir} = \hat{F}_{1i}^{dir}$ for the PG. As usual, we assume that under the sampling design $E[\hat{\theta}_i^{dir}] = \theta_i$. A p -vector X_i contains the auxiliary data sources of population characteristics for area i . By using auxiliary data, it is possible to reduce the mean squared error of direct estimates. Let us assume that the auxiliary variables X_i are exactly known. The area-level model is as follows:

$$\hat{\theta}_i^{dir} = X_i^T \beta + u_i + e_i \quad i = 1, \dots, m, \quad (2)$$

where $u_i \stackrel{iid}{\sim} N(0, \sigma_e^2)$, $i = 1, \dots, m$ are the model errors and $e_i \stackrel{ind}{\sim} N(0, \psi_i^2)$, $i = 1 \dots, m$ are the design errors, with e_i independent from u_j for all i and j . It is assumed that the quantity of interest in area i is $\theta_i = X_i^T \beta + u_i$.

Under these normality assumptions the best linear unbiased predictor (BLUP) of θ_i is:

$$\tilde{\theta}_i^{BLUP} = \gamma_i \hat{\theta}_i^{dir} + (1 - \gamma_i) \mathbf{X}_i^T \beta, \quad \gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2}. \quad (3)$$

The predictor $\tilde{\theta}_i^{BLUP}$ is a convex combination of the direct estimator $\hat{\theta}_i^{dir}$ and the predicted value $\mathbf{X}_i^T \beta$ from the regression model. The extent to which it depends on the two terms is determined by the relative sizes of the model error variance σ_u^2 and the sampling error variance ψ_i^2 .

The parameters β and σ_u^2 are unknown and must be estimated while ψ_i^2 is assumed to be known (Rao and Molina, 2015). The estimators of the ψ_i^2 s are often smoothed, and the smoothed estimators are treated as if they were the true sampling variances (Datta et al., 2005). A discussion concerning unknown sampling error variance can be found in Rao and Molina (2015).

Estimators of β and σ_u^2 can be obtained using the restricted maximum likelihood from the marginal distribution $\hat{\theta}_i^{dir} \sim N(\mathbf{X}_i^T \beta, \sigma_u^2 + \psi_i^2)$. By plugging in the estimates of β and σ_u^2 into equation (3) the empirical best linear unbiased predictor (EBLUP) is obtained:

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i \hat{\theta}_i^{dir} + (1 - \hat{\gamma}_i) \mathbf{X}_i^T \hat{\beta}, \quad \hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}. \quad (4)$$

The terms $\hat{\gamma}_i$ are commonly known as shrinkage factors.

One advantage of the model-based estimation approach is that synthetic estimation is possible in areas in which direct estimation is not possible, generally due to the absence of sample observations. However, auxiliary variables must be available in those areas to allow for synthetic estimation.

When all the parameters are known, the mean squared error (MSE) of the estimator (3) is:

$$MSE(\tilde{\theta}_i^{BLUP}) = E[(\tilde{\theta}_i^{BLUP} - \theta_i)^2] = \gamma_i \psi_i^2 = g_{1i}(\sigma_u^2). \quad (5)$$

When the parameters in (3) are estimated, we obtain the estimator (4) that has the following MSE:

$$\begin{aligned} MSE(\hat{\theta}_i^{EBLUP}) &= \gamma_i \psi_i^2 + (1 - \gamma_i)^2 \mathbf{X}_i^T V(\hat{\beta}) \mathbf{X}_i + \psi_i^4 (\psi_i^2 + \sigma_u^2)^{-3} V(\hat{\sigma}_u^2) \\ &= g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2), \end{aligned} \quad (6)$$

where $g_{2i}(\sigma_u^2)$ is the contribution to the MSE from estimating β and $g_{3i}(\sigma_u^2)$ is the contribution to the MSE from estimating σ_u^2 . In equation (6) $V(\hat{\beta})$

and $V(\hat{\sigma}_u^2)$ are the asymptotic variances of an estimator $\hat{\beta}$ of β and an estimator $\hat{\sigma}_u^2$ of σ_u^2 , respectively. An estimator of (6) is as follows:

$$mse(\hat{\theta}_i^{EBLUP}) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2), \quad (7)$$

where $g_{1i}(\hat{\sigma}_u^2) = \hat{\gamma}_i \psi_i^2$, $g_{2i}(\hat{\sigma}_u^2) = (1 - \hat{\gamma}_i)^2 \mathbf{X}_i^T [\sum_{i=1}^m \mathbf{X}_i \mathbf{X}_i^T / (\psi_i^2 + \hat{\sigma}_u^2)]^{-1} \mathbf{X}_i$, $g_{3i}(\hat{\sigma}_u^2) = \psi_i^4 (\psi_i^2 + \hat{\sigma}_u^2)^{-3} 2 [\sum_{i=1}^m 1 / (\hat{\sigma}_i^2 + \psi_i^2)^2]^{-1}$. More details concerning analytic MSE estimation for area level models can be found in Rao and Molina (2015); Datta and Lahiri (2000); Prasad and Rao (1990)

The described method can be extended to account for spatial information to improve the efficiency of the EBLUP. The target variables, HCR and PG direct estimates, are spatially autocorrelated according to a Moran index ρ , which is 0.65 for the HCR and 0.59 for the PG. The spatial autocorrelation can be accounted non-parametrically using a p-spline based estimator or can be incorporated in the area random effects u_i . When there are no out-of-sample-areas and the auxiliary variables are available for all the areas, as in our application, a popular choice is to include the spatial information in the area random effects and/or in the regression parameters.

If the regression parameters associated with the model do not vary spatially, then there is spatial stationarity. There are situations, however, where this assumption is inappropriate, a phenomenon referred to as spatial nonstationarity. We tested if in our application there is evidence of spatial nonstationarity, using a bootstrap test proposed by Chandra et al. (2015). The hypothesis of spatial stationarity is not rejected, both for the HCR and PG targets, so our choice falls on the spatial-EBLUP described hereafter.

The model (2) can be extended to allow for stationary spatially correlated area effects, as follows. Let $\mathbf{v}$ be the result of a SAR process having an unknown autoregression parameter ρ and a known proximity matrix $\mathbf{W}$ (Cressie, 1993)

$$\mathbf{v} = (\mathbf{I}_m - \rho \mathbf{W})^{-1} \mathbf{u}, \quad (8)$$

where $(\mathbf{I}_m - \rho \mathbf{W})^{-1}$ is supposed to be non-singular, $\mathbf{u} = [u_1, \dots, u_m]^T \sim N(0, \sigma_u^2 \mathbf{I}_m)$, $\mathbf{I}_m$ is an $m \times m$ identity matrix and $\mathbf{W}$ is defined in a row-standardized way. Putting together equations (2) and (8), the spatial area-level model is:

$$\hat{\theta}^{dir} = \mathbf{X}\beta + (\mathbf{I}_m - \rho \mathbf{W})^{-1} \mathbf{u} + \mathbf{e}, \quad (9)$$

where $\hat{\theta}^{dir} = [\hat{\theta}_1^{dir}, \dots, \hat{\theta}_m^{dir}]^T$, $\mathbf{e} = [e_1, \dots, e_m]$ and $\mathbf{X} = [\mathbf{x}_1^T, \dots, \mathbf{x}_m^T]$. The BLUP which follow from model (9) is:

$$\tilde{\theta}_i^{SEBLUP} = \mathbf{X}_i^T \tilde{\beta} + \mathbf{d}_i^T \mathbf{G} \mathbf{V}^{-1} [\mathbf{y} - \mathbf{X} \tilde{\beta}], \quad (10)$$

where $\mathbf{V} = \mathbf{G} + \text{diag}(\psi_i^2)$ is the covariance matrix of (9), $\mathbf{G} = \sigma_u^2[(\mathbf{I}_m - \rho \mathbf{W})^T (\mathbf{I}_m - \rho \mathbf{W})]^{-1}$ and $\mathbf{d}_i^T = [0, \dots, 1, 0, \dots, 0]$ is a selection vector with 1 in the i th position. Estimating the parameter $\tilde{\beta}$, ρ and σ_u^2 using REML we obtain the spatial-EBLUP (SEBLUP):

$$\hat{\theta}_i^{SEBLUP} = \mathbf{X}_i^T \hat{\beta} + \mathbf{d}_i^T \hat{\mathbf{G}} \hat{\mathbf{V}}^{-1} [\mathbf{y} - \mathbf{X} \hat{\beta}]. \quad (11)$$

Under normality and independence of random area effects and errors, the MSE of the SEBLUP can be decomposed as:

$$\begin{aligned} \text{MSE}(\hat{\theta}_i^{SEBLUP}) &= g_{1i} + g_{2i} + g_{3i} \\ &= \mathbf{d}_i^T [\mathbf{G} - \mathbf{G} \mathbf{V}^{-1} \mathbf{G}] \mathbf{d}_i \\ &\quad + \mathbf{d}_i^T [\mathbf{I}_m - \mathbf{G} \mathbf{V}^{-1}] \mathbf{X} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T [\mathbf{I}_m - \mathbf{V}^{-1} \mathbf{G}] \mathbf{d}_i \\ &\quad + \text{tr}\{\mathbf{L}_i \mathbf{V} \mathbf{L}_i^T \mathcal{I}^{-1}\}, \end{aligned} \quad (12)$$

where $\mathcal{I}$ is the information matrix and

$$\mathbf{L}_i = \begin{bmatrix} \mathbf{d}_i^T (\mathbf{C}^{-1} \mathbf{V}^{-1} - \sigma_u^2 \mathbf{V}^{-1} \mathbf{V}^{-1} \mathbf{C}^{-1} \mathbf{V}^{-1}) \\ \mathbf{d}_i^T (\mathbf{A} \mathbf{V}^{-1} - \sigma_u^2 \mathbf{C}^{-1} \mathbf{V}^{-1} \mathbf{A} \mathbf{V}^{-1})^T \end{bmatrix}$$

with $\mathbf{C} = (\mathbf{I}_m - \rho \mathbf{W})^T (\mathbf{I}_m - \rho \mathbf{W})$ and $\mathbf{A} = \sigma_u^2 \mathbf{C}^{-1} (\mathbf{W} + \mathbf{W}^T - 2\rho \mathbf{W}^T \mathbf{W}) \mathbf{C}^{-1}$. To estimate the MSE, we assume that under SAR models the covariance depends on a proximity matrix which specifies the proximity between the areas, therefore an approximately unbiased estimator of the MSE (Zimmerman and Cressie, 1992) is $mse(\hat{\theta}_i^{SEBLUP}) = \hat{g}_{1i} + \hat{g}_{2i} + 2\hat{g}_{3i}$, where estimates of ρ and σ_u^2 , obtained by REML, have been plugged-in (12). Molina et al. (2009) proposed alternatives based on bootstrap to estimate the MSE in (12).

In section 3 we use data obtained from EU-SILC together with administrative data provided by the Italian Ministry of Economy and Finance and ISTAT, with the aim of estimating the HRC and PG at the provincial level (LAU 1 according to EU definition).

It is important to remark that the EU-SILC in Italy is designed to obtain reliable estimates at the regional level (NUTS 2). Therefore, direct estimates in smaller domains - such as provinces - may lead to unreliable estimates, as

we show in section 3. For this reason, we referred to small area techniques in order to increase the reliability of direct estimates.

The success of improving the precision of small area estimates greatly depends on the availability of correlated auxiliary information. From the available data, we selected the most correlated variables with the target variables, i.e. the HCR and PG direct-estimates.

3 Application

3.1 Data

EU-SILC 2013 was the primary data source for this study. EU-SILC is the European Union Statistics on Income and Living Conditions survey which is conducted annually in Italy by ISTAT since 2004. The EU-SILC 2013 sample size is about 18 000 households representative of all Italian households. Data are collected on the basis of a two-stage sample design where the first stage are municipalities and the second stage are households. In Italy, the EU-SILC survey adopts a rotational sampling design consisting of 4 rotational sub-samples where each sub-sample is followed-up for 4 years.

As previously specified, the measures targeted in our study are the HCR – poverty incidence – and the PG – poverty intensity. The HCR indicator is a widely used measure of poverty. The popularity of this indicator is due to the straightforwardness of its construction and interpretation. At the same time, this indicator also assumes all poor households to be in the same situation. For example, the easiest way of reducing the HCR is by targeting benefits to people who are just below the poverty line, as they are the ones who are the cheapest to move across the line. Hence, policies based on the HCR might be sub-optimal and for this reason, we also consider the PG indicator. The PG can be interpreted as the average shortfall of poor households. It shows the amounts which would have to be transferred to the poor to rise their income to the poverty line.

Although the regional level represents the most disaggregated level of estimation for which the EU-SILC usually allows for reliable estimates, ISTAT includes geographical information on the province and on the municipality of residence of the households involved in the survey. The provincial level is the smallest level at which almost all the areas have at least two or more sampled units, hence, SAE methods are applicable¹. Small

¹If an area has two or more sampled units, it is possible to compute a direct estimate, which

area estimates at the municipality level are impractical, as about 90% of the municipalities are out-of-sample (lacking sampled units in the areas). Moreover, it is worth noting that according to the sample-design of the survey, the sample size for each province is not sufficient to obtain reliable direct estimates of HCR and PG for each province. Indeed, the sample sizes for provinces range from 10 to 2018, with 50% of provinces having a sample size which is smaller than 274 units. For this reason, we used the SAE methods (described in detail in Section 2) in order to obtain reliable estimates at this level of disaggregation.

As explained in section 2, small area methods require the availability of auxiliary variables from administrative databases or censuses at the provincial level related to the phenomenon under study, as they are necessary in the estimation process to reduce the variability of the small area estimates with respect to that of direct estimates. To this purpose, we referred to two official sources of information, which provided us with suitable data at the provincial level: the Italian Ministry of Economy and Finance archive and the Italian population and housing census.

The auxiliary variables were selected using the AIC criterion, together with the significance of the regression parameters, separately for HCR and PG. This selection method resulted in an equal model for both the targets (HCR and PG). The model includes province level information on the number of households, the average household size, the share of household who own their house and the average taxable per capita income. In what follows we describe the sources of the auxiliary variables more clearly.

Firstly, the Italian Ministry of Economy and Finance published the taxable income (in Euros) for 2012 at the municipal level. Data regarding taxable income are available at the municipal level. Due to the hierarchical administrative division characterizing Italy (i.e. regions, provinces and municipalities), each municipality is included in a specific province. The availability of the total number of taxpayers (contributors) as well as the total amount of taxable incomes for each municipality enabled us to calculate the average value of taxable income per capita in each province. It is worth noting that the data processed and published by the Ministry of Economy and Finance are those declared by the taxpayer, not yet validated by the offices in charge and therefore subject to the presence of possible inconsistencies. Further methodological issues concerning the data collected can be found at http://www1.finanze.gov.it/pagina_dichiarazioni/dichiarazioni.html. There is evidence of a strong

will possibly be unreliable. With the availability of direct estimates and auxiliary variables, SAE methods can be used to possibly obtain reliable estimates at the small area level.

Table 1: Summary statistics of 110 Italian province value for taxable income (th.), share of households who owned their house (%), number of households (th.), average household size

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
x_1 Tax. inc.	13.35	15.97	18.39	18.02	19.89	26.17
x_2 House own.	56.65	71.06	73.25	73.29	75.64	83.37
x_3 N. of hhs	24.63	100.10	152.70	223.70	243.70	1743.00
x_4 Hh size	1.98	2.29	2.40	2.40	2.51	2.89

linear correlation between the taxable income per capita and the target variables HCR and PG direct estimates at the province level. The average value of annual taxable incomes at this level, x_{1i} , is equal to 18 017 Euros (std. dev. 2695.93) with the highest value recorded for the province of Milan, in Lombardy, (26 167 Euros) and the lowest observed for Crotona in Calabria (13 351 Euros). There is a high level of heterogeneity within the regions of northern and central Italy (above all in Tuscany, Lazio and Emilia Romagna) concerning income distribution at the provincial level, while a low level of heterogeneity is observed in the southern regions, especially in Calabria (see table 1).

Secondly, we selected home ownership (x_2), in terms of the share of households who owned their house, available at the provincial level from ISTAT official data (2011 Population Census and web-site <http://www.demo.istat.it>). The share of households that owned their house ranges from 56.65% to 83.37% with an interquartile range of 4.58% (see table 1). From the same source, we also selected the number of households at the provincial level (x_3) and the average household size (x_4), computed as the total number of persons divided by the total number of households at the provincial level. The number of households ranges from 24 630 in the province of Ogliastra (Sardinia island) to 1 743 000 in the province of Rome. The interquartile range is 143 600. The average household size ranges from 1.98 to 2.89, for the province of Trieste (in the north-east) and Naples (south), respectively, with an interquartile range of 0.22.

It is worth noting that other potential auxiliary variables at the provincial level were selected from the 2011 Population Census and from the Census on Social Actions and Services on Single and Associated Municipalities and were tested as potential auxiliary variables to be used in the estimation of the area-level model (i.e. the share of households with a female as the main family provider, the public expenditure at the provincial level to assist families, the elderly, disabled persons, immigrants, the poor and other

forms of monetary support). However, the related estimated coefficients were not statistically significant and therefore the other variables were not included in the area level models.

3.2 Results

In this section, we present the results we obtained by applying the spatial small area-level (11) to estimate the HCR and PG in the 110 Italian provinces.

We first focus on the model for the HCR. In this model, the response variable values are provided by the HCR direct estimates, obtained with the Horvitz and Thompson (1952) expansion estimator, while the auxiliary variables are those already presented in section 3.1. In table 2, the regression coefficients of the model and their significance are reported. The sign and magnitude of the coefficients seems reasonable, giving us a positive feedback on the model we used.

Table 2: Regression parameters estimates for the HCR (left) and PG (right) small area models.

	HCR		PG	
	$\hat{\beta}$	p -value	$\hat{\beta}$	p -value
Intercept	0.32759121	0.00211	0.74181872	0.00406
N. of hhs	0.00000002	0.05744	0.00000006	0.05154
Tax. inc.	-0.00000832	0.00001	-0.00002383	0.00000
Hh size	0.05647464	0.02588	0.13576023	0.02310
House own.	-0.00352420	0.00003	-0.00640856	0.00138
	$\hat{\sigma}_u = 0.054$		$\hat{\sigma}_u = 0.023$	
	$\hat{\rho} = 0.519$		$\hat{\rho} = 0.461$	

To evaluate the model assumptions on area-level errors, Figure 1 presents a non-parametric density estimate with 95% confidence interval bands (obtained according to Bowman et al. (1998)) of the area-level errors $\hat{u}_i$, with a superimposed Normal density obtained from the data. The Normal curve falls within the confidence bands, giving us evidence that the Normality assumption is reasonable.

Figure 2 reports the scatterplot of the HCR direct estimates versus the model-based ones: a high correlation of these estimates is a desirable characteristics of area-level small area models. In this case, the correlation between the estimates is rather high: the value of the linear correlation coefficient is 0.95.

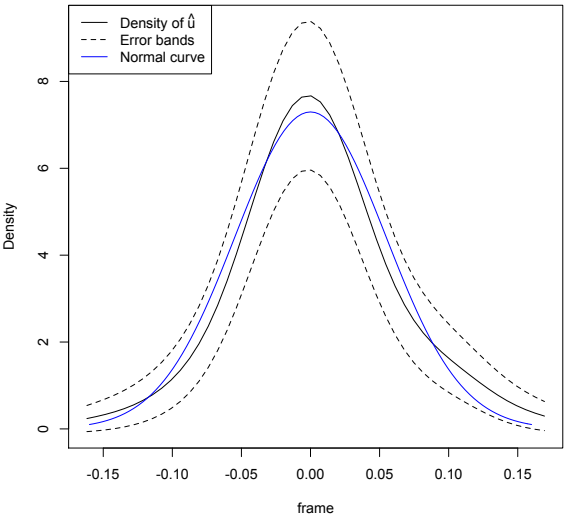


Figure 1: HCR model diagnostics: non-parametric density estimate with 95% confidence interval bands of $\hat{u}_i$ with a superimposed Normal density.

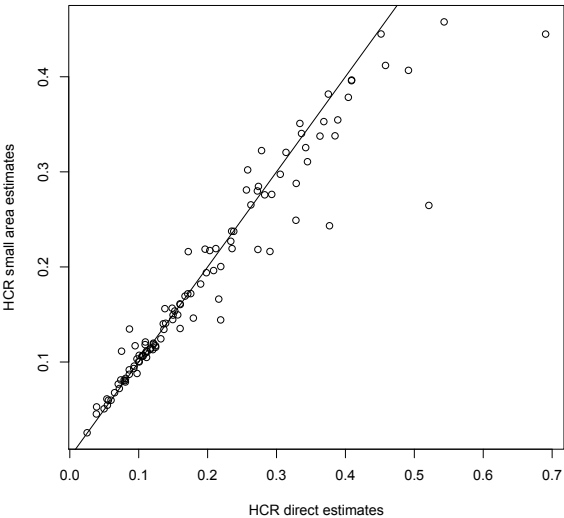


Figure 2: Scatterplot of direct estimates versus model based estimates.

As expected, the variability of small area estimates is lower than the variability of the direct ones (see figure 3). Moreover, when the sample size increases, the *rmse*s of direct estimates (white points in figure 3) and that of SEBLUPs (black points in figure 3) tend to be similar; on the contrary, when the sample size is small, the *rmse*s of SEBLUPs are remarkably smaller.

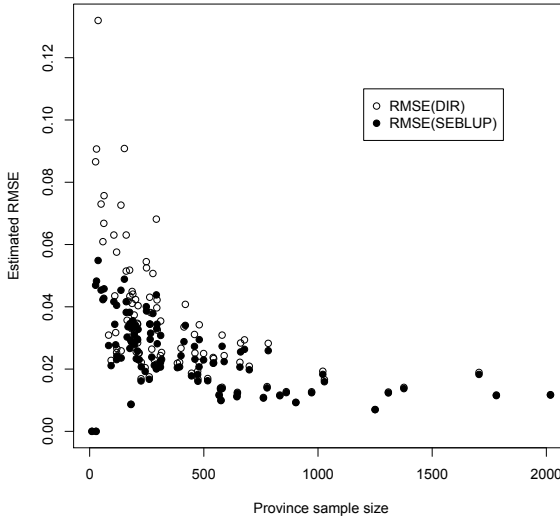


Figure 3: Root Mean Square Errors of direct (DIR) and model-based (SEBLUP) HCR estimates, by area sample size.

The distribution of estimated HCRs for the 110 Italian provinces is shown in table 3. The distribution of direct estimates and of SEBLUPs is similar up to the third quartile. The upper part of the distribution is different, but results of SEBLUPs are more realistic. Table 3 also shows the relative efficiency of the SEBLUPs (measured by the ratio $rmse(\hat{F}_{0i}^{SEBLUP})/rmse(\hat{F}_{0i}^{dir})$), which ranges between 41.6% and 99.4%. On average, the gain is about 14% and it is greater than about 20% for 25% of the provinces.

As we can see from table 3, there is a wide difference among the provinces, with a point estimate of 2.6% in Monza-Brianza up to 45.8% in Caserta. The difference between these two extremes is statistically significant; instead, when HCR point estimates are similar, a specific statistical test on the significance of the difference should be carried out. The incidence of poverty is mapped in figure 4, therefore, we can observe

Table 3: Distribution of the HCR and PG estimates and percentage efficiency of SAE in 110 Italian provinces.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
HCR						
$\hat{F}_{0i}^{dir}$	2.51	10.07	15.83	19.92	27.49	69.05
$\hat{F}_{0i}^{SEBLUP}$	2.56	10.67	15.16	18.82	26.79	45.77
$rmse(\hat{F}_{0i}^{SEBLUP})/rmse(\hat{F}_{0i}^{dir})$	41.59	80.77	90.81	86.11	94.56	99.44
PG						
$\hat{F}_{1i}^{dir}$	0.65	3.13	4.54	6.85	8.90	32.47
$\hat{F}_{1i}^{SEBLUP}$	0.73	3.20	4.59	6.25	8.52	18.75
$rmse(\hat{F}_{1i}^{SEBLUP})/rmse(\hat{F}_{1i}^{dir})$	40.19	81.73	90.54	86.98	96.00	99.32

existing differences among northern and central provinces of Italy and those in the South. Looking at the map in figure 4, we can see that in the southern Italy, the HCRs are substantially homogeneous within the regions, while in the North and Center, province details give us the opportunity to observe heterogeneity within the regions. Unfortunately, it is very challenging to increase the detail of the map with estimates at the municipality level to check for the heterogeneity which is likely to be present within the province.

The PG, which measures the intensity of poverty, is estimated with the same model used for the HCR, but the response values are the PG direct estimates. The estimated model parameters show a magnitude and a sign which seem reasonable and in accord with those of the model for the HCR (table 2). The model assumption concerning the normality of the area-level random effects, informally tested for the HCR model, as shown above, seems respected, according to the plot on figure 5. The SEBLUP point estimates are highly correlated with the direct ones (linear correlation equal 0.94), a desirable property, as shown in figure 6.

From figure 7 we can see that as the province sample size increases, the $rmse$ s of the SEBLUP converge to the $rmse$ s of the direct estimator. Moreover, when the sample size is small (less than 500) the $rmse$ s of the SEBLUP are smaller. The reduced $rmse$ s of the SEBLUP are also evident by looking at the distribution of the relative efficiency ($rmse(\hat{F}_{1i}^{SEBLUP})/rmse(\hat{F}_{1i}^{dir})$) in table 3. The gain in efficiency is about 14% in mean and, in 25% of the provinces, the gain is about 18%. In some extreme cases, the gain reaches 60%. PG point estimates range from 0.7% in the provinces of Livorno and Milano-Brianza, reaching about 18% in the provinces of Palermo and Caserta. We observe that the SEBLUPs are similar to the direct estimates in

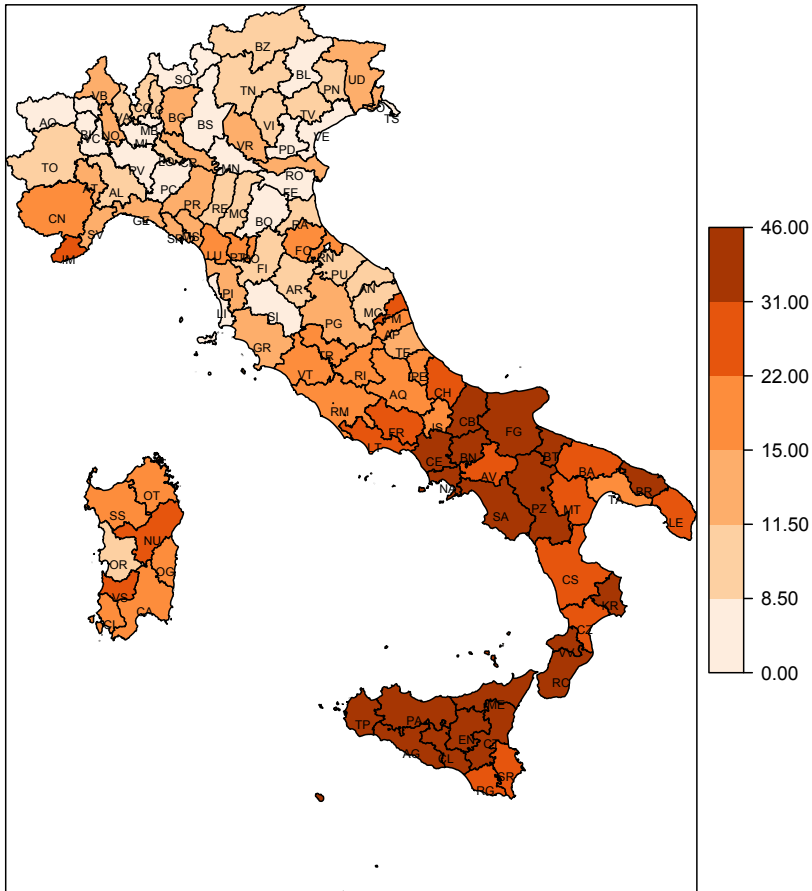


Figure 4: Map of the model-based HCR estimates for the 110 Italian provinces.

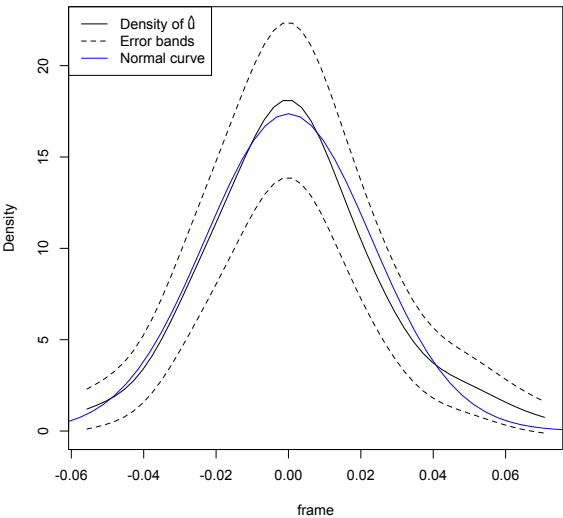


Figure 5: PG model diagnostics: non-parametric density estimate with 95% confidence interval bands of $\hat{u}_i$ with a superimposed Normal density.

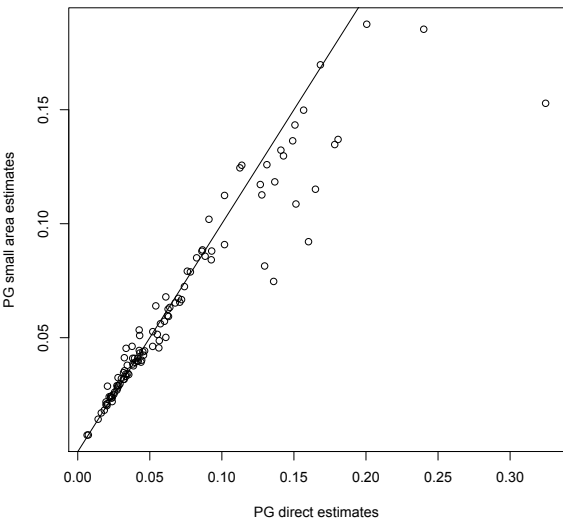


Figure 6: Scatterplot of direct estimates versus model based estimates.

the first 75% of the distribution among provinces, and differ increasingly in the upper quartile of the distribution, where PG direct estimates seem to be unreasonable. On the contrary, the SEBLUP values seem to be more realistic.

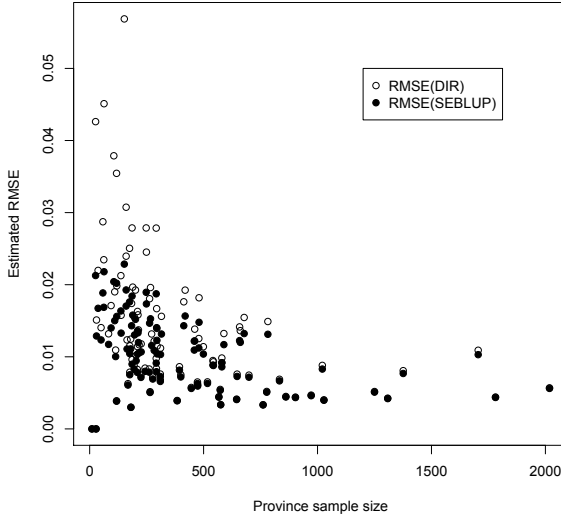


Figure 7: Root Mean Square Errors of direct (DIR) and model-based (SEBLUP) PG estimates, by area sample size.

PG spatial distribution in Italy is depicted in figure 8. This map shows a clear and well known North and Center vs. South divide. Heterogeneity within regions is present in North and Central Italy and nearly absent in the South. Thus, the PG point estimates show the same spatial behaviour of the HCR estimates. Indeed, the correlation between PG and HCR point estimates is 0.95. On average, a 1% increase in the HCR corresponds to an increase of the PG of 0.34%. However, a more in deep investigation concerning the income distribution of the (relatively) poor sections of the population is necessary.

4 Conclusions

In this chapter, we used EU-SILC data, Income Tax Office data and Population Census data to estimate the incidence (Head Count Ratio, HCR) and the intensity of poverty (Poverty Gap, PG), in 2012, at the provincial level

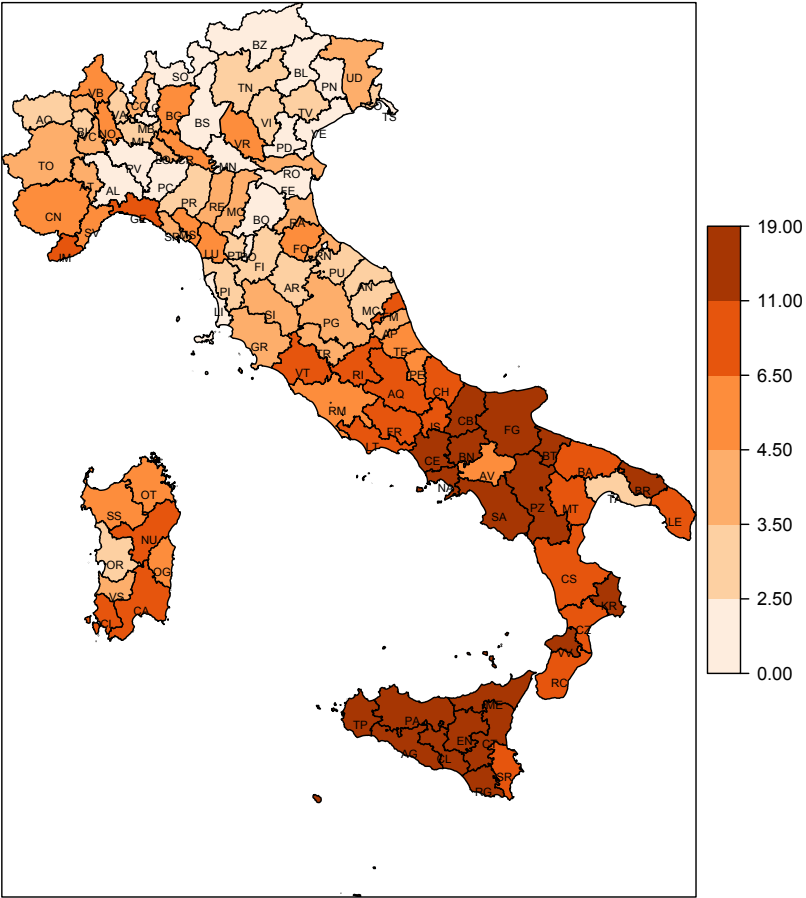


Figure 8: Map of the model-based PG estimates for the 110 Italian provinces.

in Italy by means of small area estimation methods. Given the presence of spatial correlation, we used a spatial estimator, the spatial Fay-Herriot estimator, to improve the precision of the area-level direct estimates of the two target indicators.

Our findings suggest a good fitting of the models and a good predictive power of the selected covariates. Moreover, through these models we were able to reduce the variability of the direct estimates computed at the provincial level using only EU-SILC data. The HCR and PG estimates we obtained, show a spatial distribution that would have been lost if the analysis had been conducted at a more aggregated geographical level (e.g. regional or for macro-areas), especially in the northern and central parts of Italy. The results suggest a high correlation for the HCR and PG in the 110 provinces: i.e., poverty incidence is higher where there is higher poverty intensity. From a geographical point of view, the higher estimates refer to provinces located in the southern Italy, including the region of Sicily.

Using the same models and data sources, indicators referring to the effectiveness of social protection interventions, such as the HCR before and after social transfers could also be computed. In this respect, the provincial dimension chosen in this study is particularly relevant, given that in Italy most welfare interventions are locally managed. Thus, the small area estimation techniques used in this study could also bridge the gap of good quality statistics on the impact of social transfers available at the local level.

Finally, in this chapter we analyzed data referring to the year 2012. Given that EU-SILC data are temporally comparable, we could try to evaluate the impact of the crisis, albeit indirectly, by repeating the analysis using older and more recent data (e.g. referring to years from 2007 to 2014). Unfortunately, the province variable is not included in the EU-SILC standard micro-data file and direct estimates at the provincial level are not available, so it has not been possible for us to repeat the analysis in different years. We thank the ISTAT office of Tuscany for providing the EU-SILC based direct estimates of HCR and PG at the province level for the year 2013 (income 2012).

References

- Bowman, A., Hall, P., and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85:799–808.
- Chandra, H., Salvati, N., and Chambers, R. (2015). A spatially nonstationary

- fay-herriot model for small area estimation. *Journal of survey statistics and methodology*, 3(2):109–135.
- Cressie, N. (1993). *Statistics for spatial data*. Wiley, New York.
- Datta, G. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10:613–627.
- Datta, G. S., Rao, J. N. K., and Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92(1):183–196.
- EC (2012). *Social protection performance monitor (SPPM) – methodological report by the indicators sub-group of the social protection committee*. Technical report, European Commission - Social Protection Committee.
- Eurostat (2016). *Sustainable development in the european union*. Technical report, European Union.
- Fay, R. and Herriot, R. (1979). Estimation of income from small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74:269–77.
- Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52:761–766.
- Giusti, C., Marchetti, S., Pratesi, M., and Salvati, N. (2012). Robust small area estimation and oversampling in the estimation of poverty indicators. *Survey Research Methods*, 6(3):155–163.
- Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–85.
- Molina, I., Salvati, N., and Pratesi, M. (2009). Bootstrap for estimating the mse of the spatial eblup. *Computational Statistics*, 24:441–458.
- Petrucchi, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:169–182.
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statist. Sci.*, 28(1):40–68.

- Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85:163–171.
- Rao, J. and Molina, I. (2015). *Small Area Estimation*. Wiley Series in Survey Methodology. Wiley.
- Singh, B., Shukla, G., and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, 31:183–195.
- Whelan, C., Nolan, B., and Maitre, B. (2014). Multidimensional poverty measurement in europe: An application of the adjusted headcount approach. *Journal of European Social Policy*, 24:183–197.
- Zimmerman, D. and Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann Inst Stat Math*, 44:27–43.