Session Number: Parallel Session 7C: Sub-National and Regional Estimation
Time: FRIDAY, AUGUST 29, MORNING

Paper Prepared for the30th General Conference of
The International Association for Research in Income and Wealth

Portoroz, Slovenia, August 24-30, 2008

Estimation of Poverty Rates for the Italian Population classified by Household Type and Administrative Region

Enrico Fabrizi, Maria Rosaria Ferrante, Silvia Pacei

For additional information please contact:

Name: Enrico Fabrizi
Affiliation: DISES, Facoltà di Economia, Università Cattolica, Piacenza
Email: enrico.fabrizi@unicatt.it

This paper is posted on the following website: http://www.iariw.org

Estimation of Poverty Rates for the Italian Population classified by Household Type and Administrative Region

Enrico Fabrizi [1], *Maria Rosaria Ferrante*[2], *Silvia Pacei*[3]

Summary

The aim of this work is to illustrate a methodology for calculating estimates of poverty rates based on different thresholds at the level of subsets of the Italian population (domains) obtained by cross-classification by household typology and Administrative Region. Estimates are based on data from the 2005 wave of the EU-SILC survey. As the domains of interest are much smaller that those for which the EU-SILC survey provides samples large enough for reliable estimation Small Area methods are employed. In particular we introduce a hierarchical multivariate Logistic-Normal model that is helpful in including auxiliary information into the estimation process and improving the efficiency of estimators.
We adopt a Hierarchical Bayesian approach to estimation where posterior distributions are approximated by means of MCMC computation methods.

1. Introduction

Poverty and social exclusion are unevenly distributed both geographically and across social groups. As a consequence design, implementation and monitoring of effective anti-poverty policies requires data at the level of the relevant or target sub-populations.

Many studies have shown a strong correlation between poverty and some characteristics of the household, namely its composition, with some of the household types markedly more exposed to the risk of poverty and social exclusion than others (Christopher *et. al.*, 2002; Eurostat, 2005a; 2005b). With reference to Italy the disparities among household types interact with those among the

[1] DISES, Facoltà di Economia, Università Cattolica, Piacenza, e-mail: enrico.fabrizi@unicatt.it
[2] Università di Bologna, e-mail: maria.ferrante@unibo.it
[3]Università di Bologna, e-mail: silvia.pacei@unibo.it

different regions of the country which is characterized by a low degree of regional cohesion (European Commission, 2005a), big differences in regional employement and unemployment rates and high concentration of industrial districts in some geographical areas. See also section 3 below.

We focus on the estimation of three different poverty rates for domains (sub-populations) defined cross-classifying the Italian population by household type and Administrative Region. In Italy there are 20 Administrative Regions of very different demographic size (ranging from 0.3 to 9 millions inhabitants). We consider 9 household types, thus defining 180 domains of estimation.

Estimators will be based on the data collected for Italy by the "European Union – Statistics on Income and Living Conditions" survey (EU-SILC – 2nd wave, year 2005). The three poverty rates are based on increasing poverty thresholds that are all defined as fractions of the national median of the equivalized disposable income, and are aimed at distinguishing between very poor people, poor people and people who are at risk of becoming poor (Istat, 2007).

The EU-SILC survey is designed to provide reliable estimates of main parameters of interest for areas, the Administrative Regions or group of Administrative regions that are much bigger domains than those we target. Moreover our domains are not planned (they are not survey design strata or union of strata), so no minimum sample size in the these domains is guaranteed. The number of units sampled from a large number of the domains we consider is too low, to obtain reliable estimates by direct estimation, that is applying standard design-consistent estimators to the domain-specific portion of the sample. We do not have domain specific sample sizes equal to 0 (this happened for the same domains in 1st wave of the EU-SILC, 2004). But in some cases we observe a number of 'poor' households equal to 0, that given the small sizes of the samples may not mean that 0 is a sensible estimate of the poverty rate in the domain.

To solve these problems a small area estimation (SAE) strategy is advisable. Most of SAE strategies rely on the explicit assumption of a model linking of values observed in different areas in order to improve estimation of area descriptive quantities. If relevant auxiliary information is available for each unit in the population, the models are usually specified at the unit level, in our context the household or the individual. See Rao (2003, ch 7 for an introduction, or Elbers *et al.* 2003 for an important application). In

many cases, the information at the unit level, whose sources may be Censuses or other administrative archives, is not updated or cannot be used because of confidentiality constraints.

In this application we consider area level models (see Rao, 2003, ch 7), that is we assume a model linking the estimates obtained by means of direct estimation methods. In particular, since we are interested in estimating poverty rates based on different, increasing thresholds we consider multivariate models that exploit the sampling correlation between the different rates. Multivariate SAE model often rely on the assuming normality for the direct estimators and underlying area parameters (see Ghosh et al. 1996, Datta et al., 1998). By the way normality may be inappropriate when the support of parameters to be estimated is restricted to the range [0,1] as in the case of rates, especially when the true parameter value is close to 0 or 1.

We propose a multivariate hierarchical Logistic-Normal model. With respect to more popular multivariate Normal-Normal models the Logistic-Normal warrants that the estimates of poverty rates associated to monotonically increasing thresholds are also monotonically increasing. Our proposal is to some extent similar to the one of Molina et al. (2007) for small area estimation in the analysis of the labour market; but we incorporate survey weights into the estimation process as they may protect against selection bias due to non-response and the effects of unequal selection probability.

The model we propose make use of auxiliary information, that is of information known from sources independent of the survey and that may be helpful in estimating area descriptive quantities. In particular, we use the per capita GDP at the Adiministrative Region level as obtained from Italian National System of Accounts. This variable has been selected among the many initially considered in a selection process that will be described below.

As far as estimation is concerned, we adopt a Hierarchical Bayesian approach implemented by means of MCMC computation methods. It is preferred to the frequentist prediction approach since it allows the handling of complex models such as the hierarchical multivariate non-normal model we consider in a simpler way; moreover we may use posterior variances as natural measures of uncertainty associated to point estimates (posterior summaries), while frequentist MSEs will be, for our model, very difficult to obtain. Note that posterior variances may be, under regularity conditions and careful choice of the prior distributions

for the parameters, good approximations of standard frequentist measures of variability such as the MSE (see Ganesh and Lahiri, 2008).

The results obtained allow us to compare the incidence of poverty by household type in the different Italian administrative regions. The suggested approach may be extended to the estimation of other indicators and could be used with data collected by the EU-SILC in other countries.

The paper is organized as follows. In Section 2 we briefly review EU-SILC survey and illustrate the auxiliary information considered in the application. In Section 3 we derive directs estimators and evaluate their reliability. Section 4 introduces the suggested small area models and Section 5 is devoted to the evaluation of the performance of the associated Small Area estimators. Conclusions and possible future developmentss are sketched in Section 6.

2. The data

2.1 "European Union – Statistics on Income and Living Conditions" survey: sampling design

The EU-SILC, European Union - Statistics on Income and living Conditions (European Parliament and Council Regulation, 2003; Eurostat, 2005b) is a rotating panel survey based on consistent methodology and definitions across most member of the European Union (EU). The survey is conducted in each country by National Institutes of Statistics (in Italy, by ISTAT) coordinated by Eurostat, the Statistical Bureau of the EU. In Italy, the first "official" wave of EU-SILC survey was launched in 2004.

In this paper we analyse data from the 2005 wave. The income reference period is 2004.

The EU-SILC is based on a stratified two-stage sampling design. First stage units are given by municipalities, stratified according to Administrative Province and demographic size (288 strata). Among municipalities those with at least 30,000 inhabitants are considered self-representative and form a take-all stratum. Secondary sampling units are given by households.

The effective sample of the 2005 wave of the survey contains 22,032 households and a total of 56,105 individuals. In Italy, the survey is designed to obtain reliable estimates at the level of

Administrative Regions (NUTS2 according to the EU "Nomenclature of Units for Territorial Statistics"; see http://europa.eu.int/comm/eurostat/ramon/). See Istat, 2007. Since our domains of interest are obtained subdividing the population of Administrative Regions by household typology, the sample in many of these domains is too small to obtain estimators reliable enough for meaningful analyses. In terms of households the domain specific sample sizes range from a minimum of 4 to a maximum of 600; 25^{th}, 50^{th} and 75^{th} percentiles are respectively 45, 97 and 165.

Eventually, we note that while Administrative regions are planned domains (i.e. they can be obtained as a union of strata), the domains of interest in this application are not.

2.2 Definition of target variables, domains of interest and poverty rates

The aim of the EU-SILC is to collect timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions.

The main variables, such as the total household gross and disposable income and the different income components, are defined trying to follow as closely as possible the international recommendations of the UN 'Canberra Manual' (Eurostat, 2004, 2005b).

Personal equivalent disposable income is obtained by dividing total disposable household income (see Appendix 1 for details) by equivalent household size calculated according to the OECD scale commonly used by the Eurostat. This formula gives a weight of 1.0 to the first adult, of 0.5 to the other persons aged 14 or over in the household and of 0.3 to children under the age of 14). The same equivalent disposable income is assigned to each person in the household.

The domains of interest are 180, obtained cross-classifying the population of the 20 Italian administrative regions by the 9 household typologies considered in the EU-SILC survey. These typologies are defined by simultaneously considering the household size, the presence of children and the age of components. They are defined as follows: 1. One person households; 2. Two adults, no dependent children, both adults under 65 years; 3. Two adults, no dependent children, at least one adult 65 years or more; 4. Other households without dependent children; 5. Single parent

household, one or more dependent children; 6. Two adults, one dependent child; 7. Two adults, two dependent children; 8. Two adults, three or more dependent children; 9. Other households with dependent children.

For each domain of interest we target the following poverty rates:

1. The 'poverty rate' (PR) defined the share of persons with an equivalent disposable income below the 60% of national median of personal equivalent income (standard poverty threshold).

2. The 'severe poverty rate' (PR80) defined as the share of persons with an equivalent disposable income below the 80% of the standard poverty threshold.

3. The 'at risk of poverty rate' (PR120) defined as the share of persons with an equivalent disposable income below the 120% of the standard poverty threshold.

2.3 Auxiliary information

The models described in section 4 make use of auxiliary information, that is of information on the domains of interest available from sources, such as Censuses or Administrative archives independent of the EU-SILC survey and that may be used to improve the estimation of area-specific poverty rates.

Many area level Small Area models use auxiliary information at the domain of interest level, but in principle auxiliary information at an higher level of aggregation may also be used. In our case it is not easy to obtain reliable information for the Italian population cross-classified by Administrative Region and household typology.

Analyzing a similar data set, in which domains were given by Administrative Regions (Fabrizi et al., 2008), we found evidence that poverty rates are strongly correlated with the unemployment rate. Although not routinely calculated and published, ISTAT kindly provided to us the estimates of the annual average unemployment rates for our domains in year 2004 (the income reference period). These estimates are based on the Italian Labour Force Survey (ILFS; ISTAT, 2003). In fact, the correlation is rather high (around 0.7). By the way, although calculated on a much bigger sample (the ILFS has an overall annual sample of around 300,000 households) estimates at the level of disaggregation we are interested in are characterized by a considerable level of uncertainty, in particular for typologies for which the rate of

participation to the labour market is low. This uncertainty has to be accounted for in the analysis. As we adopt an Hiearchical Bayes approach, this does not represent a technical problem, but in practice it is likely to reduce the power of this variable when used in the Small Area models.

We also considered auxiliary information at the Administrative Region level of aggregation using the regional section of the National System of Accounts, the ILFS and other administrative archives as data sources. The following variables have been taken into consideration: per-capita consumption of the household sector, per-capita GDP, per-capita employee income, per-capita expenditure for leisure and culture, per-capita taxable income, share of workers/value added in the manifacutring industry, school abandonment rate, annual average unemployment rate.

Note that all these variables being estimated at the Administrative Region level are characterized by a level of uncertainty that may be overlooked in the implementation of the models.

3. Direct estimators and estimation their variances

Since the domain we consider are not planned we modified the official final weights published in the EU-SILC data set in order to have weights calibrated on the distribution of the Italian population by Administrative Region and household typology. Final published weights are obtained by a double calibration correction of basic weights that are defined as the inverse of inclusion probabilities The first step adjusts basic weights for non-response, while the second step modifies these intermediate weights to calibrate them to known totals as suggested in the EUROSTAT guidelines for the EU-SILC survey (Istat, 2006). In particular the distribution of population by gender, age class and geographical region is considered.

To obtain weights calibrated on the distribution of the population in the domain of interest (i.e. administrative region by household type) we start from the survey intermediate weights and re-make the second step, considering the following calibration variables: Administrative Region of residence; household type; gender; age (5 classes). More precisely the weights are calibrated to the population of Administrative Regions classified by household typology and to the same population classified by age classes.

Totals are obtained from the same data sources used in the derivation of final official weights for all variables except the distribution of the population by household type within administrative regions which has been obtained as average of the quarterly Labor Force Survey results in 2005. In the calculation of the calibration weights, the log distance, leading to raking-ratio weights is used: it has the advantage of producing always positive weights (see Deville and Sarndal, 1992 for more details).

To evaluate the reliability of direct estimators we basically need to estimate their variances, and, to apply the small area multivariate models presented in the next section, we need also to estimate the covariances between estimators of different rates obtained for the same domain. Evaluating the variance and covariance of the direct estimators is in this case a complicated task, as i) the considered poverty rates are non-linear functions of data; ii) the underlying design is complex; ii) the weights used in their computation incorporate, as it has been previously described, two stages of calibration corrections.

In keeping with other work in this field (Verma and Betti, 2005), we opt for a solution based on re-sampling algorithms and in particular we propose a bootstrap estimation strategy. Boostrap variance estimators have been proposed and analyzed for sampling designs as general as multi-stage designs with stratification of primary units. See Rao (1999) for more details. By the way, these estimators rely on the assumptions that the number of strata is large and that few primary units (but at least two) are sampled from each stratum, so that the sampling fraction at the first stage is negligible. This latter assumption is not met in our case as there is a take all stratum of primary units.

For this reason, we propose a bootstrap algorithm in which any bootstrap sample is the union of two sub-samples, one taken resampling the population in the non self-representative strata and the other drawn from the stratum of self-representative municipalities, where the sampling design is actually single stage. After it is drawn, each bootstrap sample undergoes the same calibration adjustment of weights to known totals process applied to the original sample. The algorithm has been tested by means of simulation exercises and, found to provide estimates close to those obtained using the linearization method for simpler parameters (i.e. averages) for which this latter method may be applied.

Variances cannot be estimated in this way for domains in which there is no ‘poor’ households in the sample. In fact we would have

an estimate of 0 in all bootstrap samples and 0 estimates of the variance (and covariances) For the moment we decided to discard these domains (8) from the estimation process. Nonetheless note that a model based prediction of poverty rates for these domains may be obtained using the methodology illustrated in section 4.

As the number of domains is too high to present results obtained for each of them, we present, summary measures, that is, indicators allowing us to evaluate i) the variability of estimates between the domains; ii) their reliability.

Let $\hat{\theta}_{ijk}$ be the direct estimator of poverty rate θ_{ijk}, in the i,j -th domain, where i denotes the Administrative Region $(i=1,...,20)$ and j the household type $(j=1,...,9)$ and $k = PR80, PR, PR120$. In table 1 values for minimum, maximum, average, median for $\hat{\theta}_{ijk}$ and also minimum, maximum and average of their coefficient of variation are reported.

Table 1. *Summary of results obtained for the direct estimates*

	Parameter (in %)		
	PR80	PR	PR120
$\min\left(\hat{\theta}_{ijk}\right)$	0.000	0.000	0.000
$\max\left(\hat{\theta}_{ijk}\right)$	0.638	0.704	0.816
$\text{avg}\left(\hat{\theta}_{ijk}\right)$	0.122	0.213	0.315
$\text{median}\left(\hat{\theta}_{ijk}\right)$	0.091	0.184	0.301
$\min\left[CV\left(\hat{\theta}_{ijk}\right)\right]$	0.097	0.072	0.060
$\max\left[CV\left(\hat{\theta}_{ijk}\right)\right]$	1.786	1.405	1.405
$\text{avg}\left[CV\left(\hat{\theta}_{ijk}\right)\right]$	0.466	0.337	0.248

From Table 1 we may note that there are big differences among the domains in terms of poverty rates, as well as in terms or reliability of estimators.

As regards the reliability of direct estimators, the coefficient of variations are, on average, too high to consider the direct

estimators sufficiently reliable, even the case of PR120 for which the average is about 25%. This motivates the need for a Small Area estimation strategy.

In Table 2 the average correlations between the three set of rates are displayed: as expected they are positive and far from 0: this justifies the recourse to multivariate models.

Table 2. *Estimated correlation matrix (averaged over the domains)*

	PR80	PR	PR120
PR80	1	0.7	0.51
PR		1	0.73
PR120			1

Before concluding this section we add some comments on the values of estimates of poverty rates by Administrative Regions and household typology that may be helpful in understanding why there is interest in estimating poverty rates for the domains defined above.

Italy is characterized by large economic disparities: the North is rich, close to full employment, while Southern regions and Islands experience high unemployment rates and much lower levels of per-capita GDP. This 'North-South' divide is also apparent when looking at poverty rates. Table 3 reports the estimates of the three poverty rates we consider for NUTS1 macro-regions:

Table 3. *Estimated poverty reates by NUTS1 macro-regions*

Region	*PR80*	*PR*	*PR120*
North – West	0.052	0.106	0.176
North – East	0.046	0.099	0.175
Center	0.070	0.135	0.210
South	0.201	0.323	0.459
Islands	0.217	0.351	0.476

We observe big disparities also among the 9 considered household typologies; national estimates of poverty rates are reported in Table 4.

Table 4. *Estimated poverty reates by household typology*

Typology	*PR80*	*PR*	*PR120*
1	0.142	0.281	0.371
2	0.055	0.097	0.168
3	0.070	0.195	0.320
4	0.055	0.094	0.151
5	0.270	0.370	0.472
6	0.096	0.150	0.237
7	0.129	0.217	0.336
8	0.249	0.360	0.475
9	0.112	0.204	0.314

From Table 4 we may note how household typologies 1 (one person household), 5 (single parent household with dependent children) and 8 (two parents, three or more dependent children) are the most exposed to the risk of poverty.

Since regions within the country are characterized not only by different level of affluence but also by different economic and social structures, labor market participation rate, one interesting research question is too see whether the distribution of the poor in the household typologies is the same throughout the country . The limited evidence provided by direct estimators favors the answer no to this question. It seems that in Northern Administrative Regions there are more poor in typology 1,2,3 than expected under the assumption of constant distribution, and less in household typologies 7,8. The opposite seems to be true in Southern regions. A more accurate analysis of this issue cannot be conducted unless reliable estimates for the domains we defined are available.

4. The models

A Small Area area level model consists of two parts, a "sampling model" formalizing the assumptions on direct estimators and their relationships with underlying area parameters and a "linking model" that relates these parameters to area specific auxiliary information.

Let $\boldsymbol{\theta}_{ij} = \left(\theta_{ij1}, \theta_{ij2}, ..., \theta_{ijK}\right)^T$ be the $K = 3$ vector of the unknown poverty rates based on the increasing, median related thresholds introduced in section 2, calculated for the *ij*-th domain ($i = 1, ..., m = 20$ denoting the Administrative Region and

$j = 1,...,J = 9$ the household typology); let also $\hat{\boldsymbol{\theta}}_{ij}$ be the corresponding vector of direct estimators. $\boldsymbol{\theta}_{ij}$ and $\hat{\boldsymbol{\theta}}_{ij}$ are linked by the following sampling model:

$$\hat{\boldsymbol{\theta}}_{ij} \mid \boldsymbol{\theta}_{ij} \sim N_K\left(\boldsymbol{\theta}_{ij}, \boldsymbol{\Psi}_{ij}\right) \tag{4.1}$$

where the $K \times K$ positive definite $\boldsymbol{\Psi}_{ij} = V\left(\hat{\boldsymbol{\theta}}_{ij} \mid \boldsymbol{\theta}_{ij}\right)$ is assumed to be known and equal to the estimate obtained according to the bootstrap method illustrated in previous section (see Rao 2003, p. 76).

A popular and simple linking model often considered in the literature is based on assumption of normality:

$$\boldsymbol{\theta}_{ij} \mid \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_v \sim N_K\left(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_v\right) \tag{4.2}$$

where $\boldsymbol{\Sigma}_v$ is an assumed positive definite $K \times K$ prior variance matrix, and $\boldsymbol{\mu}_{ij}$ is a K-dimensional vector.

The matching of (4.1) and (4.2) produces a linear mixed model. We refer to this as to a Multivariate Normal-Normal model (M-NN). Similar models are considered in Datta *et al*. (1999), in Ghosh *et al*. (1996) and in Fabrizi *et al*. (2007). Normal-Normal models owes their popularity to the fact that posterior means of $\boldsymbol{\theta}_{ij}$ conditional on the hyperparameters may be expressed as weighted averages of direct estimators and model predictions.

Unfortunately the normality assumption does not guarantee that the estimates of rates fall into the [0,1] interval. Moreover the M-NN model does not take into consideration that $\theta_{ij1} \le \theta_{ij2} \le \theta_{ij3}$. As a consequence predictors from this model may not be monotonic in the poverty threshold. For these reasons we consider the M-NN model basically as a benchmark to evaluate the performances of the multivariate Logistic-Normal we propose.

To introduce this model let's consider that if the domain specific samples could be treated as simple random samples, then it would have been sensible to specify a multinomial likelihood for the data. In fact, as we consider poverty rates defined as function of an increasing threshold, each individual in the sample may belong to one and only one class defined according to his/her equivalent income, the boundaries between the classed being defined by the poverty thresholds. This approach is followed in obtaining small

area estimates for the Labour market by Medina *et al.* (2007). But domain specific samples cannot be treated as simple random samples as they are selected by a complex, clustered sampling design and affected by non-response.

We prefer to include sampling weights in the estimation process as they are, by construction, designed to protect against the potential bias due to non-response; moreover their use yields estimators with nice design-based properties for domains with a large number of observations. We include weights by keeping the sampling model (4.1).

Let $\xi_{ijk} = \theta_{ijk} - \theta_{ij,k-1}$, with $k = 1,2,3$, $\theta_{ij0} = 0$ and $\varsigma_{ijk} = \exp\left(\xi_{ijk}\right) / \left[1 + \sum_{k=1}^{K=3} \exp\left(\xi_{ijk}\right)\right]$. We assume the following linking model:

$$\boldsymbol{\varsigma}_{ij} \mid \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_v^* \sim N_K\left(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_v^*\right) \tag{4.3}$$

with $\boldsymbol{\varsigma}_{ij} = \left(\varsigma_{ijk}\right)_{k=1,\ldots,K}$, $\boldsymbol{\Sigma}_v^*$ is an assumed positive definite $K \times K$ prior variance matrix, and $\boldsymbol{\mu}_{ij}$ is a K-dimensional vector. Assumption (4.3) means that the ξ_{ijk} are supposed to follow Logistic-Normal distribution. The use of the Logistic-Normal as alternative to the Dirichlet distribution for the modeling of variates in the K-dimensional positive simplex is discussed in Aitchinson and Shen (1980). Even though the Dirichlet, because of its nice mathematical properties is sometimes regarded as the reference distribution in this context, the Logistic-Normal is more flexible (richer parametrization), may approximate the Dirichlet very well and the moments of its log transformation can be very easily modeled.

As the sampling and the linking model cannot be combined into a single expression, we say that this model is unmatched in the sense of You and Rao (2002). The logit linking model (4.3) has already been considered in the SAE context (see for instance Farrel *et al.,* 1997; Liu *et al.*, 2007) but only in univariate models.

We refer to the SAE model based on (4.2) and (4.3) as to the Multivariate Logistic-Normal model (M-LN).

Although models (4.2) and (4.3) are different, the vector $\boldsymbol{\mu}_{ij} = \left(\mu_{ijk}\right)_{k=1,2,3}$ is defined in the same way with

$$\mu_{ijk} = \alpha_{jk} + x_i \beta_{jk} \tag{4.4}$$

where α_{jk}, β_{jk} are intercept and slope parameters, x_i is the per-capita GDP of Administrative Region i (see section 2.3). Note that although x_i is constant for all household typologies within the same Region, slopes and intercepts are typology specific. The choice of this specification has been driven by the evidence that the relationships between the ς_{ijk} (θ_{ijk} in the M-NN case) and per-capita GDP are different for different household typologies.

The selection of per-capita GDP as the only regressor has been based on the following procedure. First we approximated the unobservable ς_{ijk} with $\hat{\varsigma}_{ijk} = \exp\left(\hat{\xi}_{ijk}\right) / \left[1 + \sum_{k=1}^{K=3} \exp\left(\hat{\xi}_{ijk}\right)\right]$ where $\hat{\xi}_{ijk} = \hat{\theta}_{ijk} - \hat{\theta}_{ij,k-1}$. Then separately for each k we selected the best subset of regressors for $\hat{\varsigma}_{ijk}$ among the variables described in section 2.3, using standard linear regression and combining forward and backward stepwise selection methods. The model with per-capita GDP as the only regressor turned out to be the best solutions in all cases. Since the M-NN model is introduced mostly for comparative purposes the same specification has been adopted.

As regards the prior specification needed to complete the Bayesian specification of models we assume:

$\alpha_{jk} \sim N\left(0, A_{jk}\right)$, $\beta_{jk} \sim N\left(0, B_{jk}\right)$, $\boldsymbol{\Sigma}_v^{-1} \sim Wishart\left(\mathbf{I}_K, K\right)$, $\left(\boldsymbol{\Sigma}_v^{*}\right)^{-1} \sim Wishart\left(\mathbf{I}_K, K\right)$

Constants A_{jk}, B_{jk} are set equal to 100, that is they are "big" with respect to the order of magnitude of the parameters. This choice implies diffuse, mildly informative, but proper and nice behaving from an MCMC point of view, prior distributions. The same criteria (approximation of non-informativeness and simplification of MCMC computation) drives the choice of *Wishart* priors for $\boldsymbol{\Sigma}_v^{-1}, \left(\boldsymbol{\Sigma}_v^{*}\right)^{-1}$.

Estimates of the poverty rates in the domains are obtained as summaries of the posterior distributions $p\left(\theta_{ijk} \mid \hat{\theta}_{ijk}, \boldsymbol{\Psi}_{ij}\right)$. Assuming, consistently with most applications a quadratic loss function we define $\hat{\theta}_{ijk}^{B} = E\left(\theta_{ijk} \mid \hat{\theta}_{ijk}, \boldsymbol{\Psi}_{ij}\right)$. $\hat{\theta}_{ijk}^{B}$ are approximated using MCMC algorithms that allow to generate samples from $p\left(\theta_{ijk} \mid \hat{\theta}_{ijk}, \boldsymbol{\Psi}_{ij}\right)$.

To implement MCMC calculations we use the OpenBugs software (Thomas *et al.,* 2006, Spiegelhalter *et al.*, 2003) which is very widely used in applied hierarchical modeling. More in detail, we run three parallel chains of $R = 25{,}000$ draws each, the starting point of which is taken from an over-dispersed distribution, and we monitor convergence by visual inspection of the chains plots. Moreover, the Gelman and Rubin statistic is also computed (Gelman and Rubin, 1992) and the autocorrelation diagrams analyzed. Although all the chains involved in our model display converge quickly, as a precaution we conservatively discard the first 5,000 iterations from each chain.

5. Model checking and analysis of the performances of Small Area estimators.

In this section we evaluate the adequacy of the proposed models and the gains in efficiency allowed by the associated estimators with respect to the direct estimators.

According to most Bayesian literature we check the fit of the models discussed in previous section following the posterior predictive approach: new observations are generated according to the posterior distribution of the given model; if the fit is adequate, then the generated observations should be similar to the observed data, otherwise the discrepancy may be summarized by some suitable measure. Among the many possible discrepancy measures we consider the following one proposed in Datta *et al.* 1999:

$$d\left(\hat{\boldsymbol{\theta}}^{B},\boldsymbol{\theta}\right)=\sum_{i=1}^{m}\sum_{j=1}^{J}\left(\hat{\boldsymbol{\theta}}_{ij}^{B}-\boldsymbol{\theta}_{ij}\right)^{T}\boldsymbol{\Psi}_{ij}^{-1}\left(\hat{\boldsymbol{\theta}}_{ij}^{B}-\boldsymbol{\theta}_{ij}\right) \qquad (5.1)$$

where $\boldsymbol{\theta}=\left(\theta_{ijk}\right)_{i=1,\ldots m;j=1,..J;k=1,2,3}$. On the basis of this discrepancy measure we can obtain the posterior predictive *p-values* as the probability that the discrepancy measure calculated for the generated new data is larger than that obtained for the observed data, given the observed data. The posterior predictive *p-value*, is expected to be near 0.5 if the model adequately fits the data.

The suggested models are then compared on the basis of the deviance information criterion (DIC), a generalization of the AIC (Akaike Information Criterion) for hierarchical models (Spiegelhalter *et al.*, 2002). It is particularly useful in Bayesian model selection problems where the posterior distributions are obtained by MCMC computational methods. The model with the

smallest DIC is assumed to be the model that would best predict a replicate dataset which has the same structure as the one currently observed. In Table 5 values obtained for such measures for the considered models are reported.

Table 5. *Bayesian measures of model fit*

	M-NN	M-LN
Posterior predictive p value	0.83	0.75
DIC	-1710	-1905

According to the chosen discrepancy measure the fit results adequate for both models, while the values of the *DIC* statistic show that the multivariate Logistic-Normal model is better than the multivariate Normal-Normal model.

As regards the evaluation of the performances of the associated estimators of poverty rates for the domains of interest, we measure the gains in efficiency using: *i*) the average percentage reduction of the Coefficient of Variation of the small area estimators versus the direct ones; *ii*) the average percentage reduction of the width of credible intervals with respect to confidence intervals of direct estimators.

Let $\hat{\theta}_{ijk}^{h}$ be equal to direct estimator of θ_{ijk} and the Bayes predictor associated to the multivariate Logistic-Normal and Normal-Normal models when $h = 0,1,2$ respectively. The average reduction of the Coefficient of Variation is defined as

$$ACVR_{kh} = 100 - \frac{ACV_{kh}}{ACV_{k0}}100\,,\ h = 1,2$$

where $ACV_{kh} = \frac{1}{mJ}\sum_{i=1}^{m}\sum_{j=1}^{J}\frac{\sqrt{\mathrm{var}\left(\hat{\theta}_{ijk}^{h}\right)}}{\hat{\theta}_{ijk}^{h}}$

Note that $\mathrm{var}\left(\hat{\theta}_{ijk}^{h}\right)$ is the posterior variance.

Moreover let $ARWCI_{h}$ be defined as

$$ARWCI_{kh} = 100 - \frac{AWCI_{kh}}{AWCI_{k0}}100\,,\ h = 1,2$$

with $AWCI_{kh} = \frac{1}{mJ}\sum_{i=1}^{m}\sum_{j=1}^{J} WCI\left(\hat{\theta}_{ijk}^{h}\right)$. Here $WCI\left(\hat{\theta}_{ijk}^{h}\right)$ indicates the posterior probability (credible) interval associated to estimator $\hat{\theta}_{ijk}^{h}$. Credible intervals of the predictors obtained by Hierarchical Bayes methods are determined using the percentiles calculated from the MCMC samples drawn from the posterior distributions, while confidence intervals for the direct estimators are calculated using the output of the bootstrap algorithm described in section 3. In all cases the probability level is set to 0.95.

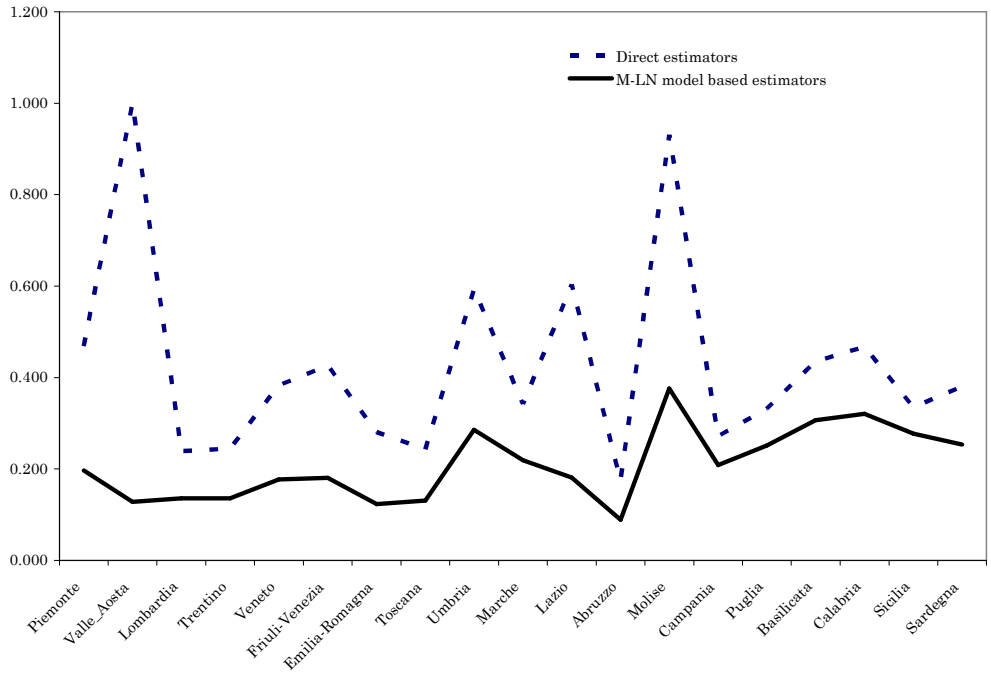
Table 6. *Performances of the Small Area estimators in terms of average reduction of coefficient of variation and width of credible intervals.*

	PR80	PR	PR120
	ACVR		
M-NN	15.0	18.0	18.3
M-LN	32.8	31.4	26.1
	ARWCI		
M-NN	20.5	23.5	21.7
M-LN	37.5	37.3	32.6

The results are summarized in Table 6. From it, it is clear that both models improve considerably the performances of direct estimators and that, consistently with the model comparisons of Table 5, the multivariate Logistic-Normal model performs better than the multivariate Normal-Normal.

The consideration of average reductions (in CVs and width of credible intervals) masks the fact that these reductions are bigger for domains of smaller size. For instance if we restrict our attention to typology 8 (two adults, three or more dependent children) which is a small subset of the households in all Administrative Regions and is characterized by much higher than the average poverty rates, the average reductions of the of the coefficient of variations and width of confidence intervals for the rate PR (Logistic-Normal model) are respectively 43.5 and 51.5. The width of probability intervals associated to M-LN model compared with the ones of direct estimators (rate PR) for this household typology and all Administrative Regions is plotted in Figure 1.

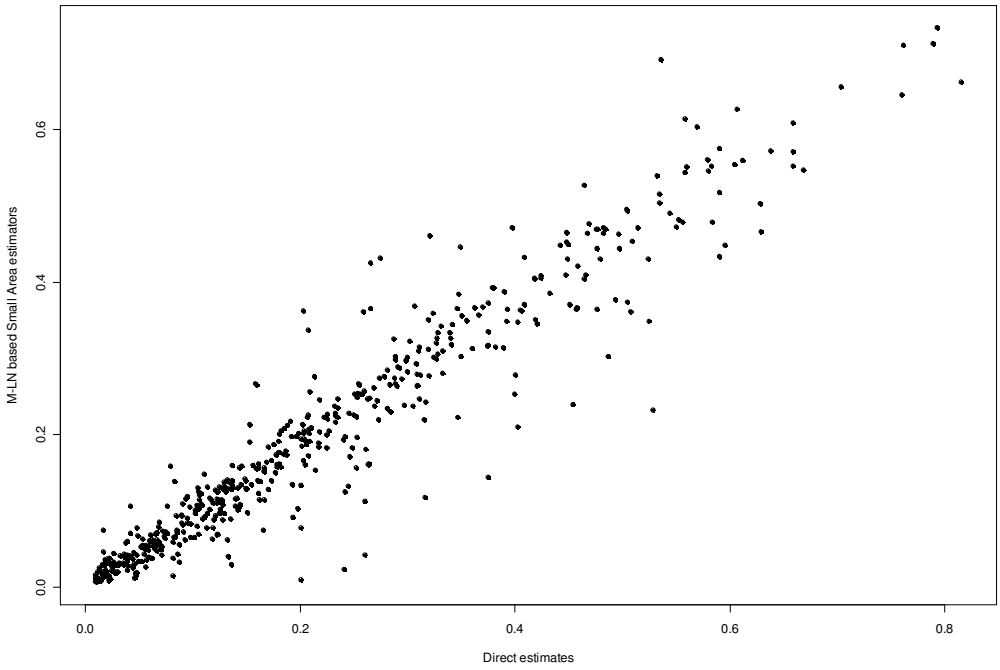
Figure 1. *Width of the credible intervals for PR – household typology 8, estimators associated to M-LN model compared with the width of confidence intervals of direct estimators*



When looking at Figure 1 consider that the biggest reduction in the width of the intervals are associated to smallest domains (Valle d'Aosta, Molise); while for big Regions (as Lombardia), the Small Area estimator remains more efficient then the direct one but the improvement is smaller, as the sample available for direct estimation is rather big.

Direct estimators and those based on Small Area models need not to be equal. For the Normal-Normal models we know that model predictors are "weighted averages" of direct and model predictions, with weights depending on the precision of direct estimators and the fit of the model. In general, the more precise is the direct estimator, the closer to it will be the small area predictor. For the Logistic-Normal and other non-matched, non linear models, the situation is more complex but estimators behave approximately in the same way. Small area estimators associated to M-LN model and direct estimators for the poverty rate PR are compared in Figure 2.

Figure 2. *Direct and Small Area estimators associated to M-LN model for parameter PR*



Even if the aim of this paper is to illustrate the methodology rather than analyzing the results, Small Area estimators associated to Logistic-Normal model for the rate *PR* and their standard deviations are reported in Appendix 2. Estimates of *PR80* and *PR120* are omitted for brevity.

6. Concluding remarks and future work

The aim of this paper was to propose a consistent methodology for the analysis of the Italian section of the EU-SILC data in order to provide reliable estimates of poverty rates of interest to policy makers and researchers.

The multivariate Logistic-Normal model improves considerably the performances of the more 'usual' models based exclusively on Normality. By the way, there are several problem still open for future research.

The assumption of normality of the direct estimators in (4.1) is not satisfying, especially when we deal with rates close to 0, whose distribution is likely to be skewed and, of course, left truncated at 0. Normality may be justified invoking the Central Limit Theorem

(when the sample size is large enough) or simply by mathematical simplicity. In the univariate context we tried models characterized by Beta sampling models. Their extension to multivariate models is a problem we are still working on.

Also in the computation of direct estimators there are open problems. In some domains direct estimates are equal to zero, this giving sampling variances equal to zero that does not necessarily imply a high degree of accuracy of the estimates (Elazar, 2004; Ghosh and Maiti, 2004). Moreover the bootstrap estimation of the variance-covariance matrix does not guarantee that it is positive definite. A solution to both problems may be represented by the introduction of smoothed estimators of the variance-covariance matrix using for instance a Generalized Covariance Function approach.

Eventually the obtained small area estimates may be used in statistical and economic analyses. We already outlined an open research problem at the end of section 3, but many others may be faced using this information.

References

Aitchinson J., Shen M.S. (1980) Logistic-Normal distributions: some properties and uses, *Biometrika,* 67, 261-272.

Christopher K., England P., Smeeding T.M., Phillips K.R. (2002), "The Gender Gap in Poverty in Modern Nations: Single Motherhood, The Market, and The State", Sociological Perspective, 45, 3, 219-242.

Datta G.S.., Day B., Maiti T. (1998), Multivariate Bayesian Small Area Estimation: An Application to Survey and Satellite Data, *Sankhya*, Serie A, 60, 1-19.

Datta G.S., Lahiri P., Maiti T. Lu (1999) Hierarchical Bayes estimation of unemployement rates for the states of the US, *Journal of the American Statistical Association*, 94, 1074-1082.

Deville J.C., Sarndal C.E. (1992), Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association,* 87, 376-382.

Elazar D. (2004), Small Area Estimation of Disability in Australia, *Statistics in Transition,* 6(5), 667-684.

Elbers C., Lanjouw J.O., Lanjouw P. (2003) Micro-level estimation of poverty and inequality, *Econometrica,* 71, 355-364.

European Commission (2005a), *Regional Indicators to reflect social exclusion and poverty*, Report prepared for Employment and Social Affairs DG - with contribution of Gianni Betti, Achille Lemmi (Project Director), Anna Mulas, Michela Natilli, Laura Neri, Nicola Salvati and Vijay Verma (Research Director), Brussels

European Commission (2005b), *Third progress report on cohesion - Towards a new partnership for growth, jobs and cohesion*, Brussels.

European Parliament and Council Regulation (EC) No 1177/2003 of 16 June 2003 concerning Community statistics on income and living conditions (EU-SILC). (3.7.2003 L 165/1 Official Journal of the European Union).

Eurostat (2004), *Description of Target Variables: Cross-sectional and Longitudinal*, Doc. EU-SILC 065/2004, 2004.

Eurostat (2005a), *The continuity of indicators during the transition between ECHP and EU-SILC*, Working papers and studies, 2005 Edition.

Eurostat (2005b), "Income Poverty and Social Exclusion in the EU25", *Statistics in Focus – Population and Social Conditions*, n. 13, A.C. Guio.

Eurostat (2005c), "Material Deprivation in the EU", *Statistics in Focus – Population and Social Conditions*, n. 21, A.C. Guio.

Fabrizi E., Ferrante M.R., Pacei S. (2005) Estimation of poverty indicators at the sub-national level using multivariate small area models, *Statistics in Transition*, 7, n. 3, pp 587-608.

Fabrizi E., Ferrante M.R., Pacei S. (2007) Small area estimation of average household income based on panel data, *Survey Methodology*, 33, 187-198.

Fabrizi E., Ferrante M.R., Pacei S. (2008), Measuring Sub-National Financial Poverty by using a Small Area Multivariate approach, *The Review of Income and Wealth,* forthcoming.

Farrell P.J., MacGibbon B., Tomberlin T.J. (1997), Empirical Bayes Estimators of Small Area Proportions in Multistage Design, *Statistica Sinica*, 7, 1065-1083.

Gelman A., Rubin D.B. (1992) Inference from iterative simulation using multiple sequences, *Statistical Science*, 7, 457-72.

Ganesh N. and Lahiri P. (2008), A new class of average moment matching prior, *Biometrika*, 95, 514-520.

Ghosh M., Nangia N. and Kim D. (1996), Estimation of Median Income of Four-Person Families: a Bayesian Time Series

Approach, *Journal of the American Statistical Association*, 91, 1423-1431.

Istat, Forze di lavoro – Media 2002, Annuari, 2003 (in Italian)

Istat (2007), La povertà relativa in Italia nel 2006, Statistiche in breve (in Italian).

Jiang J. and Lahiri P. (2006), Estimation of finite population domain means; a model assisted empirical best approach, *Journal of the American Statistical Association*, 101, 301-311

Liu B., Lahiri P. and Kalton G. (2007), Hieararchical Bayes modeling of survey weighted small area proportions, Proceedings of the Survey Research Methods Section, ASA, 3181-3186.

Molina I., Saei A., José Lombardia M. (2007), Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model, *Journal of Royal Statistical Society*, ser. A, 170, 975-1000.

Rao J.N.K. (1999), Some current trends in sample survey theory and methods, *Sankhya*, Serie B, 61, 1-55.

Rao J.N.K. (2003), Small Area Estimation, *Wiley series in Survey Mehtodology*, John Wiley and Sons.

Spiegelhalter D.J. Best N.G., Carlin B.P., Van der Linde A. (2002), Bayesian Measure of Model Complexity and Fit (with discussion), *Journal of the Royal Statistical Society*, *Serie B*, 64(4), p. 583-616.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., (2003), WinBUGS User Manual Version 1.4 , downlodable at "http://www.mrc-bsu.cam.ac.uk/bugs".

Thomas A., O' Hara B., Ligges U., Sturz S. (2006) Making BUGS Open, *R News*, 6, 12-17.

You Y. and Rao J.N.K. (2002), Small area estimation unsing unmatched sampling and linking models, *Canadian Journal of Statistics*, 30, 3-15.

Verma V. and Betti G. (2005), Sampling Errors for Measures of Inequality and Poverty", Classification and Data Analysis 2005, Book of Short Papers, CLADAG, 175-178.

Appendix 1 – Definition of total disposable household income in the EU-SILC survey

Total disposable household income can be computed as follow:

The **sum** for all household members of gross personal income components (gross employee cash or near cash income; gross non-cash employee income; employers' social insurance contributions; gross cash benefits or losses from self-employment (including royalties); value of goods produced for own consumption; unemployment benefits; old-age benefits; survivor' benefits, sickness benefits; disability benefits and education-related allowances plus gross income components at household level (imputed rent; income from rental of a property or land; family/children related allowances; social exclusion not elsewhere classified; housing allowances; regular inter-household cash transfers received; interests, dividends, profit from capital investments in unincorporated business; income received by people aged under 16

minus

employer's social insurance contributions; interest paid on mortgage; regular taxes on wealth; regular inter-household cash transfer paid; tax on income and social insurance contributions.

Appendix 2

Small area estimates of poverty rate PR associated to Logistic-Normal model. Posterior means

Administrative Region	Type								
	1	2	3	4	5	6	7	8	9
Piemonte	0.223	0.054	0.125	0.028	0.223	0.070	0.097	0.144	0.051
Valle_Aosta	0.156	0.031	0.079		0.179	0.028		0.078	
Lombardia	0.226	0.027	0.130	0.018	0.238	0.047	0.065	0.089	0.027
Trentino A. Adige	0.264	0.028		0.019	0.201	0.031	0.061	0.097	
Veneto	0.253	0.038	0.130	0.028	0.208	0.074	0.083	0.125	0.043
Friuli V. Giulia	0.227	0.054	0.115		0.313	0.071	0.080	0.128	
Liguria	0.274	0.072	0.138	0.033	0.276	0.102	0.134		0.074
Emilia Romagna	0.209	0.042	0.104	0.021	0.237	0.065	0.060	0.084	0.020
Toscana	0.223	0.042	0.142	0.029	0.213	0.060	0.117	0.116	0.050
Umbria	0.288	0.092	0.173	0.051	0.371	0.109	0.121	0.239	0.074
Marche	0.293	0.089	0.166	0.043	0.362	0.068	0.131	0.193	0.072
Lazio	0.253	0.065	0.157	0.026	0.315	0.103	0.097	0.118	0.040
Abruzzo	0.405	0.088	0.264	0.093	0.361	0.134	0.204	0.106	0.101
Molise	0.342	0.222	0.303	0.094	0.461	0.219	0.298	0.466	0.162
Campania	0.366	0.192	0.264	0.174	0.656	0.309	0.344	0.465	0.266
Puglia	0.355	0.195	0.297	0.205	0.552	0.316	0.334	0.493	0.363
Basilicata	0.404	0.157	0.348	0.180	0.517	0.204	0.287	0.482	0.279
Calabria	0.367	0.224	0.301	0.197	0.571	0.247	0.420	0.559	0.348
Sicilia	0.463	0.235	0.371	0.247	0.431	0.254	0.444	0.540	0.364
Sardegna	0.246	0.138	0.226	0.103	0.365	0.173	0.299	0.326	0.200

Small area estimates of poverty rate PR associated to Logistic-Normal model. Posterior standard deviations

Administrative Region	Typology								
	1	2	3	4	5	6	7	8	9
Piemonte	0.022	0.014	0.022	0.008	0.046	0.015	0.020	0.051	0.021
Valle_Aosta	0.029	0.011	0.022		0.055	0.009		0.033	
Lombardia	0.020	0.006	0.020	0.007	0.054	0.015	0.017	0.035	0.014
Trentino A. Adige	0.036	0.009		0.008	0.052	0.009	0.016	0.035	
Veneto	0.023	0.010	0.023	0.009	0.042	0.016	0.021	0.045	0.019
Friuli V. Giulia	0.024	0.015	0.021		0.064	0.020	0.020	0.046	
Liguria	0.026	0.014	0.021	0.010	0.059	0.024	0.031		0.031
Emilia Romagna	0.020	0.014	0.016	0.007	0.053	0.020	0.014	0.032	0.007
Toscana	0.024	0.009	0.025	0.008	0.046	0.013	0.025	0.034	0.018
Umbria	0.031	0.023	0.032	0.013	0.074	0.026	0.023	0.073	0.021
Marche	0.031	0.024	0.027	0.011	0.065	0.016	0.027	0.056	0.025
Lazio	0.020	0.019	0.026	0.010	0.054	0.033	0.022	0.046	0.021
Abruzzo	0.047	0.023	0.045	0.022	0.072	0.031	0.036	0.023	0.032
Molise	0.038	0.061	0.053	0.021	0.093	0.051	0.062	0.096	0.042
Campania	0.044	0.032	0.033	0.030	0.050	0.040	0.033	0.054	0.040
Puglia	0.026	0.037	0.042	0.036	0.084	0.057	0.033	0.064	0.049
Basilicata	0.046	0.044	0.051	0.046	0.101	0.049	0.043	0.079	0.058
Calabria	0.049	0.046	0.045	0.046	0.090	0.051	0.053	0.082	0.064
Sicilia	0.032	0.039	0.047	0.053	0.080	0.045	0.046	0.071	0.056
Sardegna	0.035	0.034	0.040	0.018	0.077	0.034	0.051	0.065	0.046