Small area estimation: recent methodological developments for the estimation of poverty measures

Caterina Giusti Monica Pratesi Nicola Salvati

Department of Statistics and Mathematics Applied to Economics, University of Pisa

Workshop
"Metodi quantitativi per l'analisi delle condizioni di vita:
nuove concettualizzazioni, stime statistiche e procedure operative
Modena - 30 gennaio 2009

SAMPLE
Small Area Methods for Poverty and Living Condition Estimates

SEVENTH FRAMEWORK PROGRAMME

SAMPLE project

Small Area Methods for Poverty and Living condition Estimates
EU-FP7- SSH-2007-1- Grant Agreement 217565

- Starting date: 1^{st} March 2008
- Partners: University of Pisa (Coordinator), University of Siena, University of Manchester, Universidad Carlos III de Madrid , Universidad Miguel Hernandez de Helce, Warsaw School of Economics, Province of Pisa, Simurg Ricerche, Glowny Urzad Statystyczny
- Web-site: www.sample-project.eu

SAMPLE
Small Area Methods for Poverty and Living Condition Estimates

SEVENTH FRAMEWORK PROGRAMME

SAMPLE project: the goal

The aim of the SAMPLE project is

- to identify and develop new indicators and models that will help the understanding of inequality and poverty with special attention to social exclusion and deprivation
- to develop models and implement procedures for estimating these indicators and their corresponding accuracy measures at the level of small area (NUTS3 and LAU 1 and 2 level).

SAMPLE
Small Area Methods for Poverty and Living Condition Estimates

SEVENTH FRAMEWORK PROGRAMME

SAMPLE project: structure of the project

The project is structured in six parts corresponding to six main areas of research or development. Each part consists of a group of tasks (called Work Package - WP) and will be carried out by a set of participant entities.

WP 1 New indicators and models for inequality and poverty with attention to social exclusion, vulnerability and deprivation (CRIDIRE / WSE / GUS / PP / UNIPI-DSMAE / SR)

WP 2 Small area estimation of poverty and inequality indicators (UNIPI-DSMAE / CCSR / UC3M / UMH)

WP 3 Integration of EU-SILC data with administrative data (PP / SR / UNIPI-DSMAE)

WP 4 Standardisation and application development - Software for living conditions estimates (SR)

WP 5 Management (UNIPI-DSMAE / ALL)

WP 6 Information, dissemination of results (SR / ALL)

Structure of the Presentation

1. Poverty indicators
2. Review mixed effects models for small area estimation
3. Extend M-quantile models to perform small area estimation
4. Nonparametric extension of M-quantile regression models
5. Define a general framework for small area estimation under which target parameters are defined as a functional of the small area distribution function
6. Present results from an application of small area models to poverty mapping

Part II

Poverty Indicators

Poverty indicators

- *At risk of poverty rate:* the share of persons with an equivalised total net income below 60% national median income
- *S80/S20 quintile share ratio:* ratio of total income received by the 20% of the country's population with the highest income (top quintile) to that received by the 20% of the country's population with the lowest income (lowest quintile)
- *Per capita income:* how much each individual receives, in monetary terms, of the yearly income that is generated in their domain through productive activities

Household income: equivalised income measures (OECD scale)

Part III

Review mixed effects models for small area estimation

Introduction to Small Area Estimation

- Increasing demand from official and private institutions of statistical data at small domains (geographical areas or population subgroups)
- National survey (e.g. EUSILC) are planned to provide reliable estimates at certain geographical level (e.g. Regioni in Italy)
- At small areas (e.g. SEL, Local Economic Systems, AR, Agrarian Regions, Provinces in Italy), there is lack of survey data and direct estimators have too large variances
- **Small Area/Domain:** Geographical area/domain where direct estimators do not reach a minimum level of precision

The Small Area Estimation Problem

- One problem is the lack of sufficient data to perform **direct** estimation at the domain level
- **Small area estimation** techniques are employed when sample data are insufficient for acceptably precise direct estimation in domains of interest
- Current methodology focuses mainly on estimating means. Nevertheless, a more complete picture is obtained by estimating the distribution of the variable of interest in a small area

Methods for Small Area Estimation

- Modern small area estimation is based on methods that are more commonly known as **model-based methods**
- The idea is to use statistical models to link the variable of interest with covariate information that is also known for units not in the sample
- A class of models suitable for small area estimation is **multilevel (mixed/random effects)** models
- A novel approach to small area estimation is based on **quantile/M-quantile models**

The Industry Standard for Small Area Estimation: Mixed Effects Models that Include Random Area Effects

Concept

Include random area-specific effects to account for the between area variation beyond that explained by the variation in model covariates

Notation: (j =area, i =individual)

- Variable of interest: y_{ij}
- Focus on unit level covariate information: $\mathbf{x}_{ij}$
- Area level random effect: γ_j - (hp: normal distribution)
- Random error: ϵ_{ij} - (hp: normal distribution)

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \gamma_j + \epsilon_{ij}, \;\; i = 1, ..., n_j, \;\; j = 1, ...d$$

Estimator of Small Area Mean

$$\hat{m}_j = N_j^{-1} \left(\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T \hat{\gamma}_j \right)$$

Part IV

Extend M-quantile models to perform small area estimation

M-quantile models

- With regression models we model the mean of the variable of interest (y) given the covariates ($\mathbf{x}$)
- A more complete picture is offered, however, by modeling not only the mean of (y) given ($\mathbf{x}$) but also other quantiles. Examples include the median, the 25th, 75th percentiles. This is known as quantile regression
- An M-quantile regression model for quantile q

$$Q_q = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\psi(q)$$

Main features of these models

- Relax the hypothesis of normal distribution
- Robust methods (influence function of the M-quantile regression)

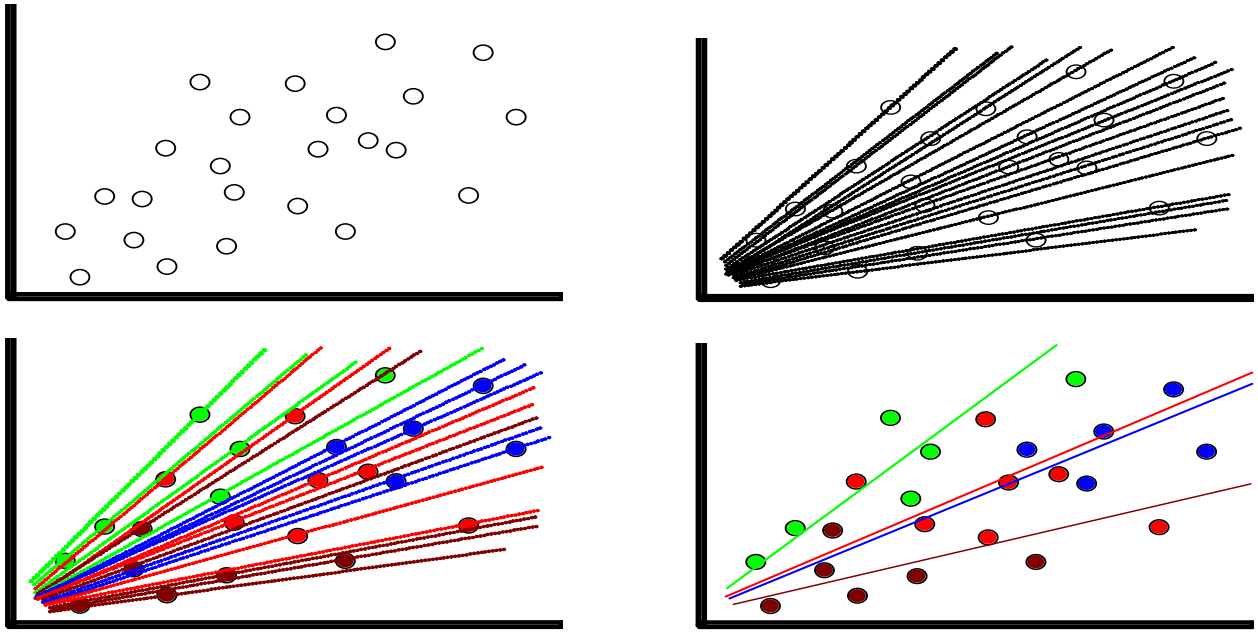
Using M-quantile Models to Measure Area Effects

- Individual level data on y and $\mathbf{x}$
- Each sample pair (x, y) lies on one and only one M-quantile line
- The q-value of this line = **M-quantile coefficient** or q value of the corresponding sample unit
- Calculate an *M-quantile coefficient* for each area j by suitably **averaging** the q values of each sampled individual in that area. Denote this average **area-specific q-value** by θ_j
- Estimate the area specific target parameter by fitting an M- quantile model for each area at $\hat{\theta}_j$

$$Q_q = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\psi(\hat{\theta}_j)$$

- A mixed effects model uses random effects γ_j to capture the dissimilarity between groups. M-quantile models attempt to capture this dissimilarity via the group-specific M-quantile coefficients $\hat{\theta}_j$

Regression M-quantile modeling of multilevel data



Part V

Nonparametric extension of M-quantile regression models

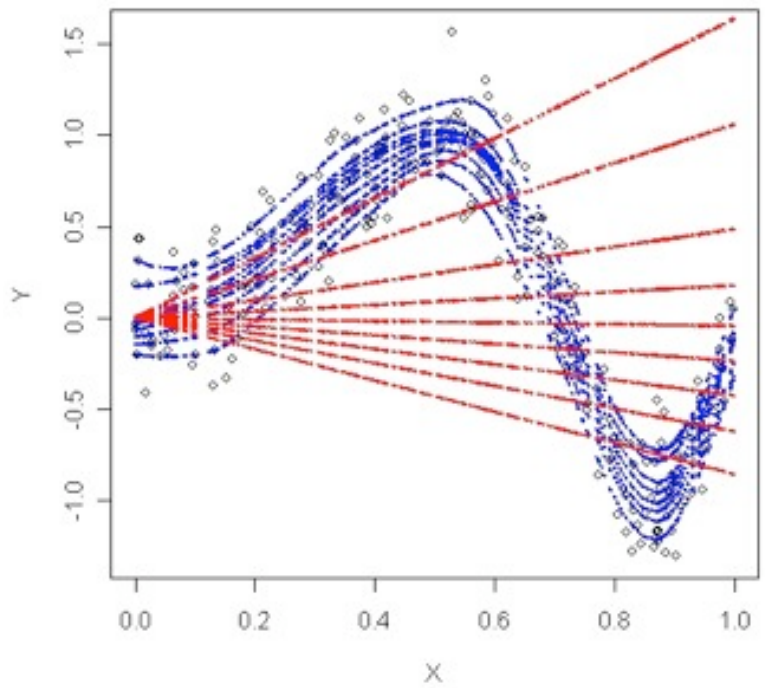
Nonparametric M-quantile Regression (Pratesi et al., 2008)

Extension of the M-quantile regression models

- the relationship between y and x is described by a semiparametric regression model (Ruppert et al., 2003)
- the $q-th$ M-quantile of y given x can be described as an additive model in which some covariates enter the model parametrically and some others nonparametrically. For the latter, the relationship is left unspecified and learnt form the data through penalized splines

A toy example of model misspecification ...

straight lines $\longrightarrow$ ordinary M-quantile regression
curves $\longrightarrow$ nonparametric M-quantile regression via p-splines (Pratesi, Ranalli & Salvati, 2008)



Nonparametric M-quantile Regression
Pratesi, Ranalli & Salvati (2008)

The assumption of linearity is relaxed. Possible models:

- $Q_q(y|\mathbf{x}) = f_\psi(x_1; q)$, univariate case
- $Q_q(y|\mathbf{x}) = f_\psi(x_1, x_2; q)$, bivariate case
- $Q_q(y|\mathbf{x}) = f_\psi(x_1, x_2; q) + x_3\boldsymbol{\beta}_\psi(q)$, additive case

Part VI

Define a general framework for small area estimation under which target parameters are defined as a functional of the small area distribution function

A General Framework for Small Area Estimation

- The target parameter in a small area can be derived from a distribution function. We start by defining the empirical distribution function, which for a small area j is

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

- The population small area mean for example is given by

$$m_j = N_j^{-1}\left(\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} y_{ij}\right)$$

- However, the y values for non-sampled units are not known and need to be predicted

Approach 1: The *Naïve* Approach

- Under mixed model

$$\hat{F}_j(t) = N_j^{-1}\Big[\sum_{i\in s_j} I(y_{ij} \leqslant t) + \sum_{i\in r_j} I(\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\hat{\gamma}_j \leqslant t)\Big]$$

- Under M-quantile model

$$\hat{F}_j(t) = N_j^{-1}\Big[\sum_{i\in s_j} I(y_{ij} \leqslant t) + \sum_{i\in r_j} I(\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(\hat{\theta}_j) \leqslant t)\Big]$$

Approach 1: The *Naïve* Approach (Cont´d)

- The corresponding estimate of the mean in small area j under the mixed effects model is given by

$$\hat{m}_j = \int_{-\infty}^{\infty} t d\hat{F}_j(t) = N_j^{-1}\Big[\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\hat{\gamma}_j\Big]$$

- The corresponding estimate of the mean in small area j under the M-quantile model is given by

$$\hat{m}_j = \int_{-\infty}^{\infty} t d\hat{F}_j(t) = N_j^{-1}\Big[\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(\hat{\theta}_j)\Big]$$

The Problem

- When the naïve approach is combined with the M-quantile model for estimating the small area mean, the estimates exhibit large bias (Chambers & Tzavidis 2006)
- A bias correction is needed. Such a correction is provided by using the Chambers-Dunstan (1986) estimator of the population distribution function

Approach 2: The Chambers-Dunstan Estimator

- Chambers and Dunstan (1986) (hereafter CD) proposed an alternative estimator of the population distribution function

$$\hat{F}_{CD,j}(t) = N_j^{-1}\Big\{\sum_{i\in s_j} I(y_{ij} \leqslant t) + n_j^{-1}\sum_{i\in r_j}\sum_{k\in s_j} I\Big\{[\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + (y_{kj} - \hat{y}_{kj})] \leqslant t\Big\}\Big\}$$

- The corresponding estimate of the mean in small area j, $\hat{m}_j^{MQ/CD}$, is given by

$$\int_{-\infty}^{\infty} t d\hat{F}_{CD,j}(t) = N_j^{-1}\Big\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(\hat{\theta}_j) + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}[y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(\hat{\theta}_j)]\Big\}$$

- Estimates of other small area quantiles are given by numerically integrating the CD distribution function

MSE Estimation for the M-quantile CD Estimator

- Given that this estimator can be expressed in a linear form, standard methods of robust mean squared error estimation for linear predictors of population quantities (Royall & Cumberland 1978) can be used
- A MSE estimator for M-quantile estimators has been proposed in Chambers, Chandra and Tzavidis (2007)

Note: The proposed MSE estimator is a first-order approximation because $\hat{\theta}_j$ are treated as fixed and not random

Part VII

Present results from an application of small area models to poverty mapping

An Application: Estimating the Head Count Ratio, the Mean and the Quantiles of Household Equivalised Income in Tuscany Provinces

- Data on the equivalised income in 2003 for 1751 households in the 10 Tuscany Provinces are available from the EU-SILC survey 2004
- A set of explanatory variables is available for each unit in the population from the Population Census 2001
- We employ linear mixed models and M-quantile models (parametric and nonparametric specifications) for estimating:
  (a) the *at risk of poverty rate* (HCR)
  (b) the mean of household income in each Province
  (c) the quantiles of household income in each Province
- The Municipality of Firenze, with 178 units out of 545 in the Province, is considered as a stand-alone area

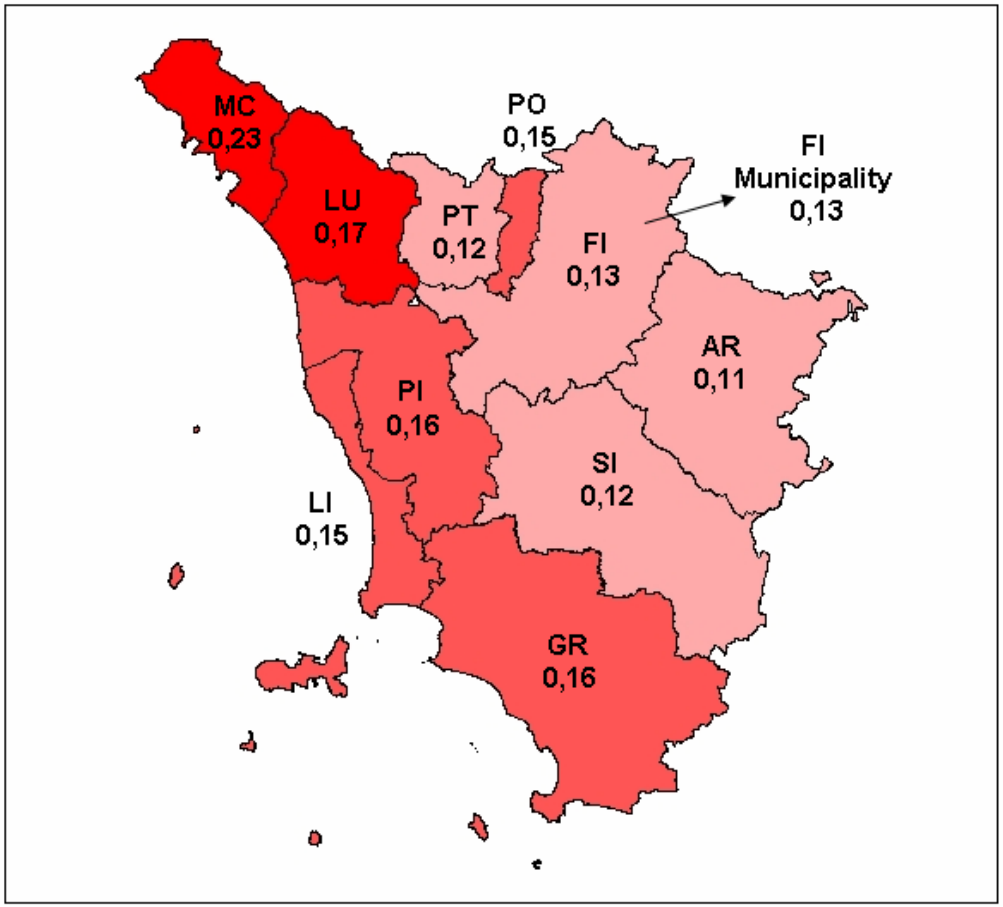
Model Specifications

- The selection of covariates to fit the small area models relies on prior studies of poverty assessment
- The following covariates have been selected:
  - household size (integer value)
  - ownership of dwelling (owner/tenant)
  - age of the head of the household (integer value)
  - years of education of the head of the household (integer value)
  - working position of the head of the household (employed / unemployed in the previous week)
- For the HCR the poverty line is set at 9188,16 Euros (60% of the median of household equivalised income)

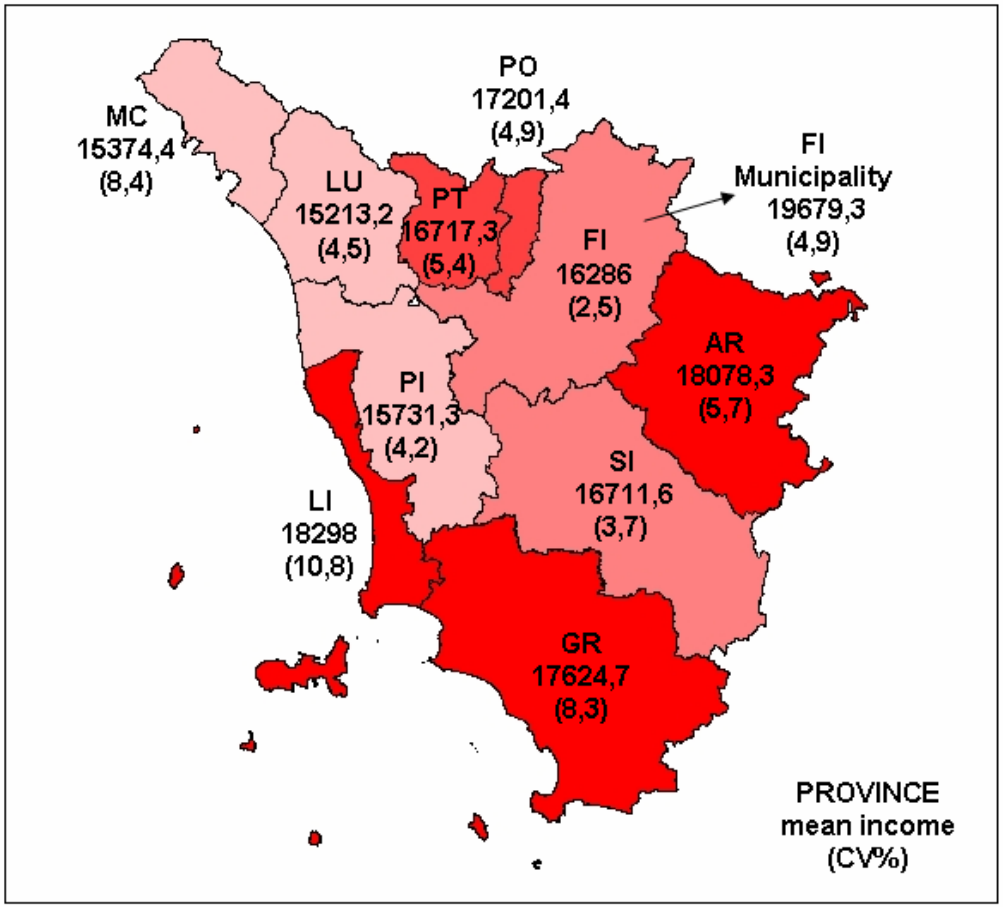
Results: a comparison of the estimators

- All the estimators give the same indication about the monetary poverty in the small areas
- Massa Province has the highest percentage of poor individuals (more than 20%)
- Arezzo Province has the lowest percentage of poor individuals (10-14%)
- For the estimation of the mean, the EBLUP estimator of the linear mixed model:
  - over-shrink the distribution of the small area estimates
  - the precision of the estimates is smoothed among areas
- The linear models (mixed and M-quantile) seem able to handle the slight non-linearity in the relationship of the income with the age of the head of the household
- Finally, we note that the Gaussian assumption of the mixed models seems not to be met: an M-quantile model with a bounded influence function seems more reasonable for these data

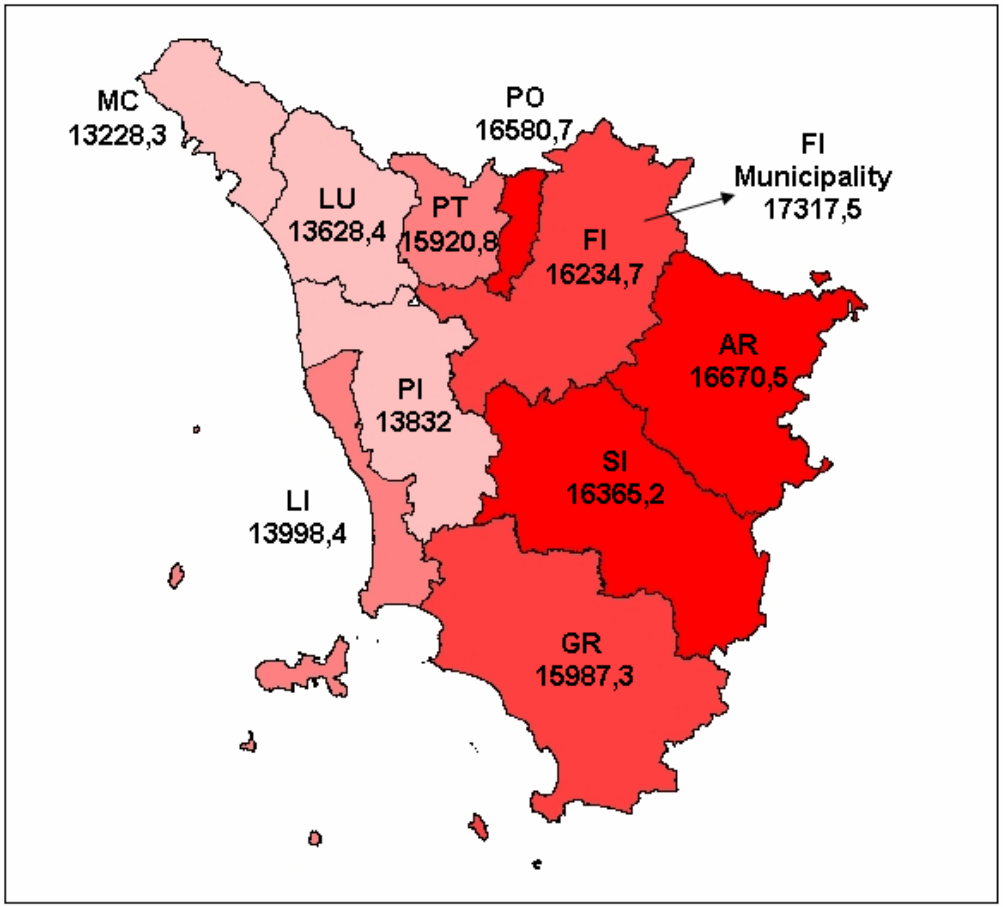
Estimate of the Head Count Ratio (% of individuals below the poverty line) - MQCD Estimator



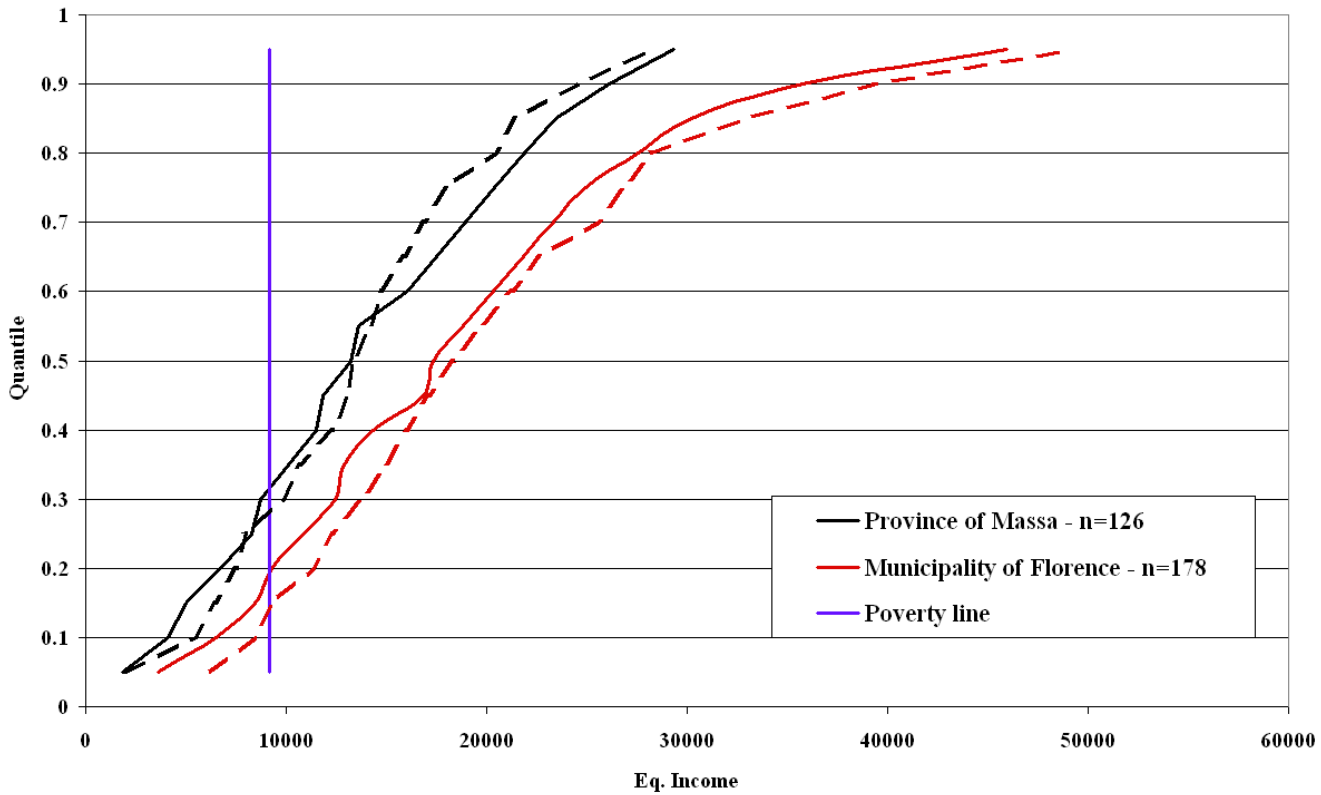
Estimate of the mean of the household equivalised income - MQCD Estimator



Estimate of the median of the household equivalised income - MQCD Estimator



Estimate of Cumulative Distribution Function of the household equivalised income - MQCD Estimator



Part VIII

Concluding remarks

Concluding remarks and ongoing research

Main results

- Small area methods play a crucial role in providing poverty measures at local level
- It comes from our application that, although the estimates are similar, M-quantile estimators seems to better track the the differences in precision across the areas (estimation of the mean)
- Moreover, M-quantile estimators out-perform the other methods when the Gaussian hypothesis is not met

Future research

- Consider non-monetary measures of poverty (Cheli and Lemmi, 1995)
- Enhance the fitting of the models, considering some transformation for the income
- Develop an MSE estimator for the cumulative distribution function
- The next step in Italy is to obtain estimates for Municipalities
- Over-sampling for the EU-SILC survey (ISTAT) in the Province of Pisa (SAMPLE project)

Essential bibliography

- Breckling, J. and Chambers, R. (1988). M -quantiles. *Biometrika*, **75**, 761–771.
- Chambers, R. and Dunstan, R. (1986). Estimating distribution function from survey data, *Biometrika*. **73**, 597–604.
- Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**, 255–268.
- Chambers, R., Chandra, H. and Tzavidis, N. (2007). On robust mean squared error estimation for linear predictors for domains. CCSR Working paper 2007-10, University of Manchester.
- Cheli B. and Lemmi, A. (1995). A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. *Economic Notes*, 24, 115-134.
- Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2006). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, 70, 265-286.
- Pratesi, M., Ranalli M.G. and Salvati N. (2008) Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US, *Environmetrics*, 19, 687-701.
- Royall, R. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, 73, 351-358.
- Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2007). M-quantile models for poverty mapping. *Statistical Methods & Applications*, 17, 393-411.