Recent Development in Small Area Estimation Methodology

Prof. Monica Pratesi

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

Valmiera, 24-28 August 2012

Structure of the Presentation

1. The small area estimation problem
2. Recent development in Small Area Estimation Methodology
3. The Elbers, Lanjouw and Lanjouw approach to SAE
4. The Empirical Best Predictor for SAE
5. The M-quantile approach to SAE
   - Simulations
6. Small area estimation methods that use spatial information
7. Concluding remarks

Part I

A short introduction to the small area estimation problem

Introduction to Small Area Estimation

- Problem: demand from official and private institutions of statistical data referred to a given population of interest
- Possible solutions:
  - Census
  - Sample survey

Sample surveys have been recognized as cost-effectiveness means of obtaining information on wide-ranging topics of interest at frequent interval over time

Introduction to Small Area Estimation

- Population of interest (or target population): population for which the survey is designed

  $\rightarrow$*direct estimators* should be reliable for the target population
- Domain: sub-population of the population of interest, they could be planned or not in the survey design
  - Geographic areas (e.g. Regions, Provinces, Municipalities, Health Service Area)
  - Socio-demographic groups (e.g. Sex, Age, Race within a large geographic area)
  - Other sub-populations (e.g. the set of firms belonging to a industry subdivision)

  $\rightarrow$we don't know the reliability of *direct estimators* for the domains that have not been planned in the survey design

Introduction to Small Area Estimation

- Often *direct estimators* are not reliable for some domains of interest
- In these cases we have two choices:
  - oversampling over that domains
  - applying statistical techniques that allow for reliable estimates in that domains

Small Domain or Small Area

Geographical area or domain where direct estimators do not reach a minimum level of precision

Small Area Estimator (SAE)

An estimator created to obtain reliable estimate in a Small Area

Introduction to Small Area Estimation: Example

- Target population: households who live in an Italian Region
- Variable of interest: Income or other poverty measures
- Survey sample: EUSILC (European Union Statistics on Income and Living Conditions), designed to obtain reliable estimate at Regional level in Italy
  - planned design domains: Regions
  - unplanned design domains: e.g. Provinces, Municipalities
- EUSILC sample size in Tuscany: 1751 households
  - Pisa province 158 households $\longrightarrow$ need SAE (or an oversample)
  - Grosseto province 70 households $\longrightarrow$ need SAE (or an oversample)

Introduction to Small Area Estimation: Example

- US sample sizes with an equal probability of selection method sample of 10,000 persons

State	1994 Population (thousands)	Sample size
California	31,431	1207
Texas	18,378	706
New York	18,169	698
⋮	⋮	⋮
DC	570	22
Wyoming	476	18

- Suppose to measure customer satisfaction for a government service:
  - California 24.86% $\rightarrow$ leads to a confidence interval of 22.4%-27.3% (reliable)
  - Wyoming 33.33% $\rightarrow$ leads to a confidence interval of 10.9%-55.7% (unreliable)

Part II

Recent development in Small Area Estimation Methodology

Motivation

- Research in Small Area Estimation field is mainly focused on estimating small area means and totals
- In some application fields there is need to estimate non-linear statistics at small area level
- This fact leads to attempt to create new estimation methods in the small area context

One application field that need non-linear statics estimation at small area level is the so called *poverty mapping*

Motivation - Poverty mapping

- Wikipedia definition of *Poverty mapping*

Methodology for providing a detailed description of the spatial distribution of poverty and inequality within a country. It combines individual and household (micro) survey data and population (macro) census data with the objective of estimating welfare indicators for specific geographic area as small as village or hamlet.

Motivation - Poverty Mapping - Welfare Indicators

- The poverty is a complex phenomena and should not be synthesized with a unique indicator
- We think that to try to understand poverty from a quantitative point of view is important to estimate quantiles and poverty indicators, such as head count ratio, poverty gap and inequality indexes as well as means
- In many developed country (relative) poverty occurs locally, for small domains
- ....and small domains are unplanned domains in sample survey on individuals and households

To have a local picture of poverty there is need of small area estimators for income quantiles and poverty - welfare indicators

Poverty Indexes

- Among poverty indicators the so called Laeken indicators are very often used to target poverty and inequalities
- Laeken indicators are a core set of statistical indicators on poverty and social exclusion agreed by the European Council in December 2001, in the Brussels suburb of Laeken, Belgium
- They include measures of the incidence of poverty, such as the Head Count Ratio (also known as at-risk-of-poverty-rate - HCR) and the intensity of poverty, such as the Poverty Gap (PG)
- These two poverty indicators are part of the generalized measures of poverty introduced by Foster et al. (1984) (FGT poverty measures hereafter)

Poverty Indexes

Foster, Greer and Thorbecke (FGT) (1984) define a measure of poverty based on the poverty line t and on a welfare variable y. For N units their poverty measure is

$$Z(\alpha, t) = \left(\frac{t - y_i}{t}\right)^{\alpha} I(y_i \leq t) \quad i = 1, \ldots, N \, .$$

- Setting $\alpha = 0$ we define the Head Count Ratio (or At-Risk-of-Poverty-Rate) index. The HCR is a measure of incidence of the poverty
- Setting $\alpha = 1$ we define the Poverty Gap index. The PG is a measure of intensity of the poverty
- Setting $\alpha = 2$ the measure is called poverty severity. This measure squares, and large values of $Z(2, t)$ point out to areas with severe level of poverty
- The poverty line t is generally computed as $0.6 \cdot \mathit{median}(y)$ and in this presentation is treated as a known value

Motivation - Poverty Mapping - Estimation approaches

Recent experiences

Despite the great importance of the disposal of poverty estimates for policy makers, the estimation of poverty in small areas has been studied only recently

- SAIPE, Small Area Income and Poverty Estimates: Fay and Herriot approach to SAE http://www.census.gov/hhes/www/saipe (Bell, 1997; Maiti and Slud 2002)
- World Bank applications of the method by Elbers, Lanjouw and Lanjouw (2003)
- AMELI project, Advanced Methodology for the European Laeken Indicators: design-based and model-based approach to SAE via calibration methods
- SAMPLE project, Small Area Methods for Poverty and Living Conditions Indicators: model based approach to SAE via mixed models and M-quantile models

Motivation - Poverty Mapping - Estimation approaches

Available Solutions

To estimates means, quantiles and poverty indicators for small areas and provide also an estimator of the corresponding mean squared errors, we mention here three possible approaches: ELL, EBP, MQ

- ELL: the Elbers, Lanjouw and Lanjouw approach to SAE
- EBP: The Empirical Best Predictor approach by Molina and Rao
- MQ: M-quantile approach by Chambers, Tzavidis and SAMPLE team

Notation

- m: is the number of small areas of interest
- i: is the subscript for the small areas, $i = 1, \ldots, m$
- j: is the subscript for the units in a small area, $j = 1, \ldots, n_i$
- n_i: is the sample size in area i
- n: is the total sample size, $\sum_{i=1}^{m} n_i = n$
- N is the population size, while N_i is the population size in area i
- s: is the set of the sampled units and s_i is the set of sampled units in area i
- r: is the set of the non sampled units and r_i is the set of non sampled units in area i
- y_{ij}: is the study variable for the unit j in the area i
- $\mathbf{x}_{ij}$: is the vector of the p auxiliary variables for the unit j in area i (this vector is known for all the units in the population)

The ELL approach (or World Bank approach)

- This approach is mainly due to the work of Elbers, Lanjouw and Lanjouw (2003)
- This method is widely used to estimate poverty indicators and it is also known as World Bank method
- In what follow we show a short description of the ELL method as stated in Molina and Rao (2010)

The ELL approach (or WB approach)

- Suppose the population of interest has L clusters, $l = 1, \ldots, L$
- Suppose that there is a one-to-one transformation $t_{lj} = T(y_{lj})$, such that the vector $\mathbf{t}$ containing the values t_{lj} for all the population units satisfies $\mathbf{t} \sim N(\boldsymbol{\mu}, \mathbf{V})$
- Super-population model $t_{lj} = \mathbf{x}_{lj}^T \boldsymbol{\beta} + u_l + \epsilon_{lj}$
- $u_l \sim \text{iid} N(0, \sigma_u^2)$, $\epsilon_{lj} \sim \text{iid} N(0, \sigma_\epsilon^2)$, $u_l \perp \epsilon_{lj}$
- With REML we obtain estimates of $\boldsymbol{\beta}$, σ_u and σ_ϵ
- Generate A ($a = 1, \ldots, A$) bootstrap populations: $t_{lj}^{*,a} = \mathbf{x}_{lj}^T \hat{\boldsymbol{\beta}} + u_l^* + \epsilon_{lj}^*$, where u_l^* and ϵ_{lj}^* have been drawn respectively from $N(0, \hat{\sigma}_u^2)$ and $N(0, \hat{\sigma}_\epsilon)$
- Compute any statistics $\theta_l^{*,a}$ on the $t_{lj}^{*,a}$ values
- Any statistics θ_l can be estimated as $\hat{\theta}_l = A^{-1} \sum_{a=1}^{A} \theta_l^{*,a}$
- The MSE of $\hat{\theta}_l$ is estimated as $A^{-1} \sum_{a=1}^{A} (\theta_l^{*,a} - \hat{\theta}_l)^2$

Remark: When we use the ELL method to estimate small areas statistics we assume that clusters correspond to small area

The EBP for SAE

- This method has been proposed by Molina and Rao (2003)
- Suppose that there is a one-to-one transformation $t_{ij} = T(y_{ij})$, such that the vector $\mathbf{t}$ containing the values t_{ij} for all the population units satisfies $\mathbf{t} \sim N(\boldsymbol{\mu}, \mathbf{V})$
- The target is to predict $\theta = h(\mathbf{t})$, where h is a real measurable function
- The predictor $\hat{\theta}$ of θ can be obtained minimizing the $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$
- The Best Predictor of θ is $\hat{\theta}^B = E_{\mathbf{t}_r}[\theta|\mathbf{t}_s]$
- The Empirical Best Predictor (EBP) of θ is $\hat{\theta} = E_{\mathbf{t}_r|\mathbf{t}_s}[\theta|\mathbf{t}_s]$ where the unknown parameters that determine the distribution of $\mathbf{t}$ are replaced by a proper estimator

The EBP for SAE Poverty Mapping

- Super-population model $t_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + \epsilon_{ij}$
- $u_i \sim \text{iid} N(0, \sigma_u^2)$, $\epsilon_{ij} \sim \text{iid} N(0, \sigma_\epsilon^2)$, $u_i \perp \epsilon_{ij}$
- With REML we obtain estimates of $\boldsymbol{\beta}$, σ_u and σ_ϵ
- The conditional small area predictor under this model is $\boldsymbol{\mu}_i = \mathbf{x}_{ij}^T \beta + \hat{u}_i$
- Generate A $(a = 1, \ldots, A)$ times the t value for the non-sampled units: $t_{ik}^{*,a} = \boldsymbol{\mu}_i + \nu_i^* + \epsilon_{ik}^*$, $k \in r_i$, where ν_i^* and ϵ_{ij}^* have been drawn respectively from $N(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i))$ and $N(0, \hat{\sigma}_\epsilon)$
- $\gamma_i = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/N_i)$ and $\hat{\gamma}_i = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/N_i)$
- Generate A bootstrap population such that $\mathbf{t}_i = [t_{ij}, j \in s_i, \cup t_{kj}^{*,a}, j \in r_i]$, where t_{ij} is the transformed value for sampled units
- Compute any statistics $\theta_i^{*,a}$ on the $\mathbf{t}_i$ values
- Any statistics θ_i can be estimated as $\hat{\theta}_i = A^{-1} \sum_{a=1}^{A} \theta_i^{*,a}$
- MSE of $\hat{\theta}_i$ can be obtained by a parametric bootstrap technique

The M-quantile approach to SAE

- The M-quantile approach to small area estimation has been proposed by Chambers and Tzavidis (2006)
- This method is based on the M-quantile regression model and it is an alternative to the methods that are based on the mixed effect models
- The M-quantile regression is a generalized robust model to handle the tail of a conditional distribution
- The estimators we present here are based on the M-quantile linear model with a squared loss function and with the Huber proposal 2 influence function

M-quantile

- Given q, with $q \in (0,1)$, the M-Quantile θ_q of a random variable X is defined as:

$$\int \psi_q(x - \theta_q) dF(x) = 0$$

where

$$\psi_q(u) = \begin{cases} (1-q)\psi(u) & u < 0 \\ q\psi(u) & u \geq 0 \end{cases}$$

and $\psi_q(u)$ is an opportunely chosen influence function

The M-Quantile is a generalization of the quantile concept and includes as particular cases quantile and expectile

M-quantile Regression

M-quantile Linear Regression

- Dependent variable $(y_1, \ldots, y_n)$
- Auxiliary variables for unit j: $\mathbf{x}_i = [x_{1_i}, \ldots, x_{p_i}]^T$
- $\theta_q(x) = \alpha_q + \mathbf{x}_i^T \boldsymbol{\beta}_q + \epsilon_i$
- The M-Quantile θ_q of order q, with $q \in (0, 1)$, of the conditional distribution $Y|\mathbf{X}$ is defined as: $\int \psi_q(y - \theta_q(\mathbf{x}))dF(y|\mathbf{x}) = 0$ with

$$\psi_q(u) = \begin{cases} (1-q)\psi(u) & u < 0 \\ q\psi(u) & u \geq 0 \end{cases}$$

- $\psi(u)$ is a continuous influence function

- M-Quantile regression is a unified model that includes quantile regression and expectile regression as particular cases

How to use M-quantile model to measure small area effects

Small area model-based estimators borrow strength from all the sample to capture random area effects, given the hierarchical structure of the data. M-quantile regression does not depend on a hierarchical structure. We can characterise conditional variability across the population of interest by the M-quantile coefficients of the population units

- Linear mixed effects model captures random area effects as differences in the conditional distribution of y given x between small areas
- M-Quantile model determines area effect with M-Quantile coefficients of the units belonging to the area

How to use M-quantile model to measure small area effects

- Assume that we have individual level data on y and $\mathbf{x}$
- Each sample value of $(\mathbf{x}, y)$ will lie on one and only one M-Quantile line
- We refer to the q-value of this line as the M-Quantile coefficient of the corresponding sample unit. So every sample unit will have an associate q-value
- In order to estimate these unit specific q-values, we define a fine grid of q-values (e.g. 0.001,...,0.999) that adequately covers the conditional distribution of y and $\mathbf{x}$.
- We fit an M-Quantile model for each q-value in the grid and use linear interpolation to estimate a unique q-value, q_j, for each individual j in the sample

How to use M-quantile model to measure small area effects

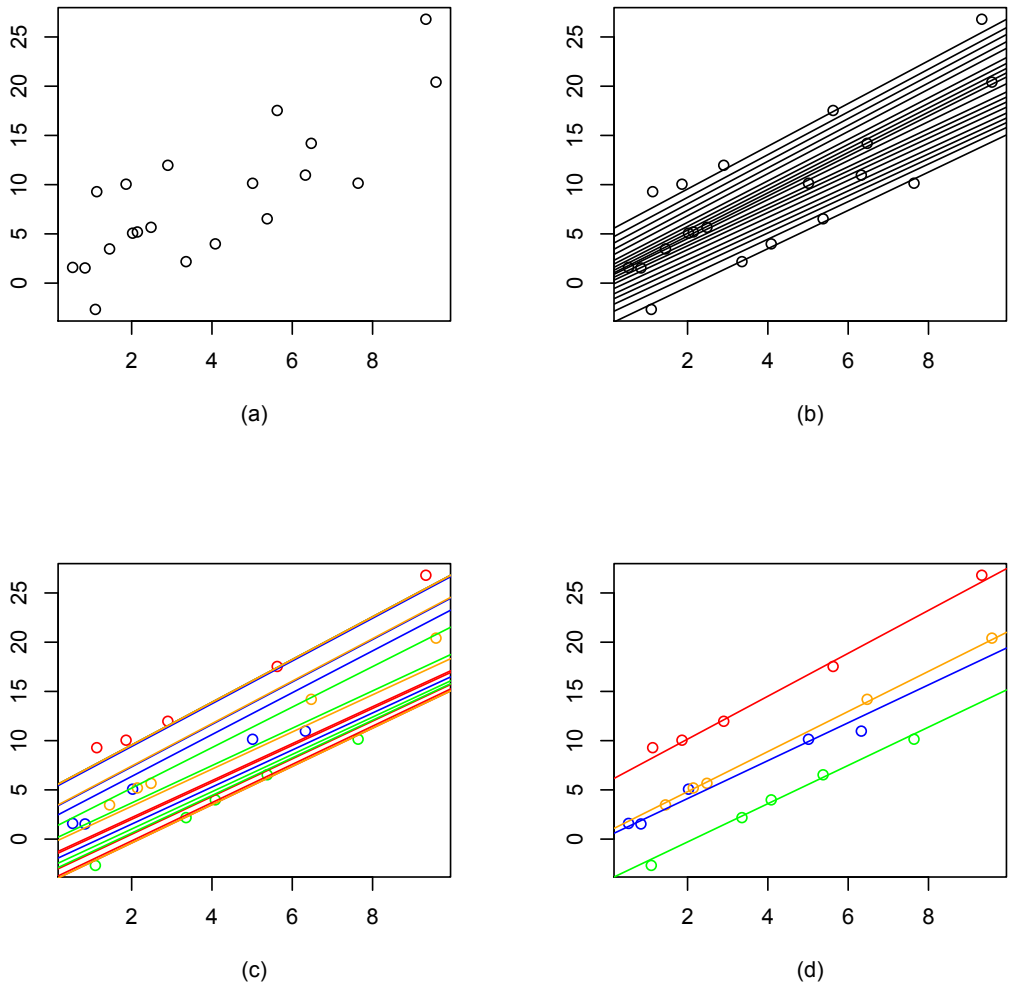


Figure: (a) Sample data, (b) M-quantile lines, (c) M-quantile lines associated to each unit, (d) M-quantile area lines

How to use M-quantile model to measure small area effects

- Calculate an M-Quantile coefficient for each area by suitably averaging the q-values of each sampled individual in that areas. Denote this area-specific q-value by $\hat{\theta}_i$

The M-Quantile small area model is

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\psi(\theta_i) + \varepsilon_{ij}$$

- $\boldsymbol{\beta}$ is the unknown regression vector
- θ_i is the unknown area specific coefficient
- ε_{ij} is an individual disturbance

The linear M-quantile small area model

- Linear M-quantile small area model: $y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\psi(\theta_i) + \varepsilon_{ij}$
- $\boldsymbol{\beta}_\psi$ is estimated using the iterative weighted least square
- θ_i is obtained by averaging the q-values of the sampled units belonging to area i
- $\psi(u) = u\, I(|u| \leq c) + \text{sgn}(u)\, c\, I(|u| > c)$ (Huber proposal 2 influence function)
- ε_{ij} has a non specified distribution
- The predictor for the target variable of the non sampled unit k in area i is

$$\hat{y}_{ki} = \mathbf{x}_{ki}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)$$

M-quantile Poverty Mapping

Denoting by t the poverty line and by y a measure of welfare, the Foster et al. (1984) poverty measures (FGT) for a small area i can be defined as

$$Z_i(\alpha, t) = N_i^{-1}\left[\sum_{j \in s_i} z_{ij}(\alpha, t) + \sum_{k \in r_i} z_{ik}(\alpha, t)\right]$$

where for a generic unit j in area i

$$z_{ij}(\alpha, t) = \left(\frac{t - y_{ij}}{t}\right)^{\alpha} \mathrm{I}(y_{ij} \leqslant t) \quad j = 1, \ldots, N_i$$

- $z_{ij}(\alpha, t)$ is known for $j \in s_i$
- $z_{ik}(\alpha, t)$ is unknown for $k \in r_i$ and should be predicted

Poverty Measures Estimator

Using a *smearing-type* predictor that follow the same idea of the Chambers and Dunstan (1986) distribution function estimator we can predict the $z_{ik}(\alpha, t)$ values

$$\hat{z}_{ik}(\alpha, t) = n_i^{-1} \sum_{j \in s_i} \left(\frac{t - \hat{y}_{ikj}}{t} \right)^{\alpha} \mathrm{I}(\hat{y}_{ikj} \leqslant t) \quad k \in r_i, j \in s_i$$

- $\hat{y}_{ikj} = \mathbf{x}_{ik}^T \boldsymbol{\beta}_{\psi}(\hat{\theta}_i) + e_{ij}$
- $e_{ij} = y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_{\psi}(\hat{\theta}_i)$

Finally, the small area estimator of FGT poverty measures is

$$\hat{Z}_i(\alpha, t) = N_i^{-1} \left[\sum_{j \in s_i} z_{ij}(\alpha, t) + \sum_{k \in r_i} \hat{z}_{ik}(\alpha, t) \right]$$

Setting $\alpha = 0$ defines the *Head Count Ratio* whereas setting $\alpha = 1$ defines the *Poverty Gap*.

A Mean Squared Error Estimator of the Poverty Measures Estimator

To estimate the mean squared error of the M-quantile poverty estimators we can use the bootstrap proposed by Tzavidis et al. (2010) and Marchetti et al. (2010).

- Let $b = (1, \ldots, B)$, where B is the number of bootstrap populations
- Let $r = (1, \ldots, R)$, where R is the number of bootstrap samples
- Let $\Omega = (y_k, \mathbf{x}_k)$, $k \in (1, \ldots, N)$, be the target population
- By $\cdot^*$ we denote bootstrap quantities
- $\hat{Z}_d(\alpha, t)$ denotes the FGT poverty measures estimator of the small area d
- Let y be the study variable that is known only for sampled units and let $\mathbf{x}$ be the vector of auxiliary variables that is known for all the population units
- Let $s = (1, \ldots, n)$ be a within area simple random sample of the finite population $\Omega = \{1, \ldots, N\}$

A Mean Squared Error Estimator of the Poverty Measures Estimator

- Fit the M-quantile regression model on sample s, $\hat{y}_{jd} = \mathbf{x}_{jd}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d)$
- Compute the residuals, $y_{jd} - \hat{y}_{jd} = e_{jd}$
- Generate B bootstrap populations of dimension N, Ω^{*b}
  1. $y_{kd}^* = \mathbf{x}_{kd}' \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d) + e_{kd}^*$, $k = (1, \ldots, N)$
  2. e_{kd}^* are obtained by sampling with replacement residuals e_{jd}
  3. residuals can be sampled from the empirical distribution function or from a smoothed distribution function
  4. we can consider all the residuals $(e_j, j = 1, \ldots, n)$, that is the unconditional approach or only area residuals $(e_{jd}, j = 1, \ldots, n_d)$, that is the conditional approach.
- From every bootstrap population draw R samples of size n without replacement

A Mean Squared Error Estimator of the Poverty Measures Estimator

- From the B bootstrap populations and from the R samples drawn from every bootstrap population estimate the mean squared error of the FGT estimator

Bias

$$\hat{E}\left[\hat{Z}(\alpha,t)^{*}-Z(\alpha,t)^{*}\right]=B^{-1}\sum_{b=1}^{B}R^{-1}\sum_{r=1}^{R}\left(\hat{Z}(\alpha,t)^{*br}-Z(\alpha,t)^{*b}\right)$$

Variance

$$\widehat{Var}\left[\hat{Z}(\alpha,t)^{*}-Z(\alpha,t)^{*}\right]=B^{-1}\sum_{b=1}^{B}R^{-1}\sum_{r=1}^{R}\left(\hat{Z}(\alpha,t)^{*br}-\bar{\hat{Z}}(\alpha,t)^{*br}\right)^{2}$$

where

- $Z(\alpha,t)^{*b}$ is the FGT of the bth bootstrap population
- $\hat{Z}(\alpha,t)^{*br}$ is the FGT estimate for $Z(\alpha,t)^{*b}$ estimated using the rth sample drown from the bth bootstrap population
- $\bar{\hat{Z}}(\alpha,t)^{*br}=R^{-1}\sum_{r=1}^{R}\hat{Z}(\alpha,t)^{*br}$

Small area quantiles estimators based on the M-quantile approach

- We start by defining the empirical distribution function, which for a small area i is

$$F_i(t) = N_i^{-1}\left(\sum_{j \in s_i} I(y_{ij} \leqslant t) + \sum_{j \in r_i} I(y_{ij} \leqslant t)\right)$$

- The qth quantile of small area i, $\tau(q)_i$ is given by

$$\int_{-\infty}^{\tau(q)_i} dF_i(t) = q$$

- The y values for non-sampled units are not known and need to be predicted

Small area quantiles estimators based on the M-quantile approach

- Chambers and Dunstan (1986) (hereafter CD) proposed a smearing type estimator of the population distribution function

$$\hat{F}_{CD,i}(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_{ij} \leqslant t) + n_i^{-1}\sum_{j\in r_i}\sum_{k\in s_i} I\left\{[\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + (y_{ij} - \hat{y}_{ik})] \leqslant t\right\}\right\}$$

- The corresponding estimate of the quantile q in small area j, $\hat{\tau}(q)_j$, is given by

$$\int_{-\infty}^{\hat{\tau}(q)_i} d\hat{F}_{CD,i}(t) = q$$

- To estimate the MSE of the small area quantiles estimator we used a bootstrap method proposed firstly by Tzavids et al. (2010) that is similar to the bootstrap described before

Simulations scenario

- Individuals, $j = 1, \ldots, N_i$, are clustered within 30 areas, $i = 1, \ldots, 30$
- The response variable, y_{ij}, which reflects a welfare indicator, is generated for each individual in the population for the nested error regression model
- A single covariate is drawn from a Normal distribution $x_{ij} \sim N(\mu_i, 1)$
- The mean, μ_i, vary across areas within the range $3 \leq \mu_i \leq 10$
- Intercept and slope terms take the values $\boldsymbol{\beta} = [3000, -150]^T$
- Area effects and individual errors are also drawn from Normal distributions
  - $u_i \sim N(0, 200^2)$, area effects
  - $\epsilon_{ij} \sim N(0, 800^2)$, individual errors

Simulations scenario

- Population values of the welfare variable are generated using $y_{ij} = 3000 - 150 \cdot x_{ij} + u_i + \epsilon_{ij}$ (super-population model)
- $N = 9580$ population units are generated
- A sample is taken from each of the populations generated so that the sample size of each area is 10% of its total size
- The super-population model is used to simulate $H = 500$ populations
- The simulation focused on FGT poverty indicators estimation
- Results are contrasted with direct FGT estimates

Measuring Estimators Performances

- True empirical values for the FGT poverty measures are calculated for each area from the corresponding Monte-Carlo population as

$$z_i(\alpha, t) = N_i^{-1} \sum_{j=1}^{N_i} z_{ij}(\alpha, t)$$

- Where the poverty status of an individual, $z_{ij}(\alpha, t)$, is calculated as

$$z_{ij}(\alpha, t) = \left(\frac{t - y_{ij}}{t} \right)^{\alpha} I(y_{ij} \leq t)$$

- The Bias and Root MSE (RMSE) of the estimates for each area are calculated over simulations using

$$Bias(\hat{z}_j(\alpha, t)) = H^{-1} \sum_{h=1}^{H} (\hat{z}_j(\alpha, t) - z_j(\alpha, t))$$

$$RMSE(\hat{z}_j(\alpha, t)) = \sqrt{H^{-1} \sum_{h=1}^{H} (\hat{z}_j(\alpha, t) - z_j(\alpha, t))^2}$$

Simulations results

Across area distribution of Bias and RMSE of estimates of HCR. Results are averaged over Monte-Carlo simulations

Bias - HCR	Min.	25th	Median	Mean	75th	Max.
MQ	-0.0091	-0.0066	-0.0056	-0.0055	-0.0045	0.0016
Direct	-0.0054	-0.0023	-0.0001	-0.0005	0.0017	0.0033

RMSE - HCR	Min.	25th	Median	Mean	75th	Max.
MQ	0.0232	0.0310	0.0354	0.0382	0.0443	0.0657
Direct	0.0404	0.0518	0.0541	0.0583	0.0647	0.0912

Simulations results

Across area distribution of Bias and RMSE of estimates of PG. Results are averaged over Monte-Carlo simulations

Bias - PG	Min.	25th	Median	Mean	75th	Max.
MQ	-0.0059	-0.0033	-0.0026	-0.0027	-0.0018	-0.0010
Direct	-0.0053	-0.0009	-0.00004	-0.0002	0.0007	0.0014

RMSE - PG	Min.	25th	Median	Mean	75th	Max.
MQ	0.0083	0.0122	0.0151	0.0182	0.0224	0.0446
Direct	0.0157	0.0217	0.0240	0.0275	0.0317	0.0559

Simulations results

- Examining the simulation results, we note that the M-quantile estimates of HCR and PG are more efficient than the corresponding Direct estimates
- This implies that using a small area model improves estimation in this case
- Comparisons between M-quantile, EBP and ELL approaches will be reported elsewhere (lecture 2)
- Simulation studies on RMSE estimation of the M-quantile HCR and PG small area estimators are available in Marchetti et. al 2012

Small Area Estimation by Borrowing Strength over Space

- In applications involving economic, environmental and epidemiological data observations that are spatially close may be more alike than observations that are further apart
- This creates a type of spatial dependency or spatial association in the data that invalidates the assumption of independent and identically distributed (iid) observations used by conventional regression models
- One approach to accounting for spatial correlation in the data is offered by specifying models with spatially correlated errors (Anselin 1992; Cressie 1993)
- Small area literature suggests that prediction of small area parameters may be improved by borrowing strength over space (Saei and Chambers 2003; Singh *et al.* 2005; Petrucci and Salvati 2006; Pratesi and Salvati 2007, 2009)

Global Vs. Local Models for Modeling Spatial Dependency

- Regression models with spatially correlated errors are global models i.e. they assume that the relationship we are modelling holds everywhere in the study area
- Another approach to modelling a spatially non-stationary process is offered via Geographically Weighted Regression (GWR) (Brunsdon *et al.* 1996; Fotheringham *et al.* 1997)
- GWR models attempt to capture the spatial association in the data by allowing local, rather than global parameters, to be estimated

GWR Models

- Assume that we have n observations on $(y_j, \mathbf{x}_j)$ at a set of Locations (u_j)
- A GWR model is defined as follows

$$y_j = \mathbf{x}_j^T \boldsymbol{\beta}(u_j) + \epsilon(u_j)$$

- GWR models allow for local rather than global parameters to be estimated and will produce estimated local surfaces of the relationship between y and x
- GWR models work by assuming that observed data near to location j will have a greater influence on the estimation of $\boldsymbol{\beta}(u_j)$ than observations farther from j
- Weighted Least Squares (WLS) is used for estimating the GWR parameters

M-quantile Geographically Weighted Models

- Salvati, Tzavidis, Pratesi and Chambers (2010) propose a robust GWR model namely an M-quantile GWR model. This is a locally robust to outliers model
- With this model the authors attempt to model locally the different quantiles of the conditional distribution accounting at the same time for the spatial non-stationarity in the data
- For estimating the parameters of the M-quantile GWR model an Iterative Weighted Least Squares algorithm is used

Estimation for M-quantile Geographically Weighted Models

- An M-quantile GWR model is defined as follows

$$y_j = \mathbf{x}_j^T \boldsymbol{\beta}(u_j; q) + \epsilon(u_j; q)$$

- The model parameters $\boldsymbol{\beta}(u_j; q)$ are estimated by solving

$$\sum_{l=1}^{L} w(u_l, u) \sum_{j=1}^{n_l} \psi_q \left\{ y_{jl} - x_{jl}^T \boldsymbol{\beta}(u; q) \right\} x_{jl} = 0$$

- Estimates of $\boldsymbol{\beta}(u_j; q)$'s are obtained via IWLS:

$$\hat{\boldsymbol{\beta}}(u_j, v_j; q) = (\mathbf{x}^T \mathbf{W}^* \mathbf{x})^{-1} \mathbf{x}^T \mathbf{W}^* \mathbf{y}$$

- $\mathbf{W}^*$ is an n by n diagonal matrix combining the spatial weights with the weights from the influence function and the modeled quantile

M-quantile GWR Models for Small Area Estimation

- Achieved via an extension to the algorithm used for estimating group effects with M-quantile models
  1. Estimate an M-quantile coefficient for each unit in the sample, $\hat{\theta}_{ij}$, using M-quantile GWR models. The $\hat{\theta}_{ij}$'s are now estimated accounting for the spatial structure in the data
  2. Recognize the hierarchical structure of the data and estimate a group specific M-quantile coefficient, $\hat{\theta}_i$, using the unit level M-quantile coefficients, $\hat{\theta}_{ij}$
  3. Estimate the area specific target parameter by fitting an M-quantile GWR model for each area at $\hat{\theta}_i$

$$y_{ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(u_j; \hat{\theta}_i) + \epsilon_{ij}(u_j; \hat{\theta}_i)$$

M-quantile GWR Small Area Estimators

- Under an M-quantile GWR model a 'naïve' small area estimator of the mean is

$$\hat{m}_i^{MQGWR} = N_i^{-1}\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(u_j;\hat{\theta}_i)\}$$

- A bias-corrected small area estimator derived under the CD or the RKM estimator $\hat{m}_j^{MQGWR/CD}$ of the distribution function is (Salvati et al., 2010)

$$N_i^{-1}\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(u_j;\hat{\theta}_i) + \frac{N_i - n_i}{n_i}\sum_{j\in s_i}[y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(u_j;\hat{\theta}_i)]\}$$

MSE Estimation

- MSE estimation of the small area mean is based on the ideas described in Chambers, Chandra and Tzavidis (2009) and in Salvati, Chandra, Ranalli and Chambers (2010)
- The MQGWR CD estimator can be expressed as a weighted sum of the sample y-values

$$\hat{m}_i^{MQGWR/CD} = N_i^{-1}\mathbf{w}_{s_i}^T\mathbf{y}_s$$

$$\mathbf{w}_{s_i} = \frac{N_i}{n_i}\mathbf{1}_{s_i} + \sum_{j\in r_i}\mathbf{H}_{ij}^T\mathbf{x}_j - \frac{N_i - n_i}{n_i}\sum_{j\in s_i}\mathbf{H}_{ij}^T\mathbf{x}_j$$

- Given the linear representation, an approximation to the MSE can be computed by applying the ideas of robust mean squared error estimation for linear predictors of population quantities (Royall and Cumberland, 1978)

$$\hat{V}(\hat{m}_i^{MQGWR/CD}) = \sum_{k:n_k>0}\sum_{j\in s_k}\lambda_{ijk}\left\{y_j - \hat{Q}_{\hat{\theta}_i}(x_j;\psi,u_j)\right\}^2$$

where $\lambda_{ijk} = \{(w_{ij}-1)^2 + (n_i-1)^{-1}(N_i-n_i)\}I(k=i) + w_{jk}^2 I(k\neq i)$

Estimation for Out of Sample Areas

- There are situations where we are interested in estimating small area characteristics for domains with no sample observations
- The conventional approach to estimating a small area characteristic in this case is synthetic estimation:

$$\hat{m}_i^{MX/SYNTH} = N_i^{-1} \sum_{j \in U_i} \mathbf{x}_j \hat{\boldsymbol{\beta}}$$

$$\hat{m}_i^{MQ/SYNTH} = N_i^{-1} \sum_{j \in U_i} \mathbf{x}_j \hat{\boldsymbol{\beta}}(0.5)$$

- One way of potentially improving the efficiency of synthetic estimation is by using the MQ GWR model. A synthetic-type mean predictor for out of sample area i is then

$$\hat{m}_i^{MQGWR/SYNTH} = N_i^{-1} \sum_{j \in U_i} \hat{Q}_{0.5}(\mathbf{x}_j; u_j)$$

Part III

Concluding remarks

Advantages and drawbacks of the M-Quantile approach with respect to the mixed model based approaches

Main advantages

1. Distributional assumptions on parameters are not needed
2. Assumptions on the hierarchical structure are not needed
3. M-Quantile model is robust against outliers
4. It is easy to implement non parametric M-Qauntile approach
5. Bootstrap approach to the estimate of MSE is faster than bootstrap for EBLUP (mixed linear model require double bootstrap techniques)

Main drawbacks

1. There is no specification if the response variable is multivariate
2. There is no specification if the response variable is binary (but work in progress)

Concluding remarks and ongoing research

Main results

- Focus on new methods for Poverty Mapping
- Small area methods play a crucial role in providing poverty measures at local level

Ongoing and future research

- Consider non-monetary measures of poverty (Cheli and Lemmi, 1995)
- Enhance the fitting of the models, considering non parametric models and spatial models
- Compare with alternative methods
- Take into account the survey weights

Essential bibliography

- Breckling, J. and Chambers, R. (1988). M -quantiles. *Biometrika*, **75**, 761–771.
- Brunsdon, C., Fotheringham, A.S. and Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**, 281–298.
- 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.L., Chandra, H., Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, **37**, 153–170.
- Cheli B. and Lemmi, A. (1995). A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. *Economic Notes*, **24**, 115–134.
- Elbers, C., Lanjouw, J. O., Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.
- Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). Geographically Weighted Regression. John Wiley and Sons, West Sussex.
- Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.
- Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68**, 2, 221–238.
- Lombardia M.J., Gonzalez-Manteiga W. and Prada-Sanchez J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of finite population distribution function. *Journal of Statistical Planning and Inference*, **116**, 367–388.
- Marchetti, S., Tzavidis, N. and Pratesi, P. (2012). Nonparametric Bootstrap Mean Squared Error Estimation for M-quantile Estimators for Small Area Averages, Quantiles and Poverty Indicators. *Computational Statistics and Data Analysis*, **56**, 2889–2902.
- Royall, R. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, **73**, 351–358.
- Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2010). Small Area Estimation Via M-quantile Geographically Weighted Regression. [Paper submitted for publication in TEST]
- Salvati, N., Chandra, H., Ranalli, M.G. and Chambers, R. (2010). Small Area Estimation Using a Nonparametric Model Based Direct Estimator. *Journal of Computational Statistics and Data Analysis*, **54**, 2159-2171.
- Tzavidis N., Marchetti S. and Chambers R. (2010). Robust estimation of small area means and quantiles. *Australian and New Zealand Journal of Statistics*, **52**, 2, 167–186.
- Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2007). M-quantile models for poverty mapping. *Statistical Methods & Applications*, **17**, 393–411.

Use of Small Area Estimators in Italy. Case Studies

Prof. Monica Pratesi

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

Valmiera, 24-28 August 2012

Structure of the Presentation

Part I

Poverty Mapping in Italy

Demand of statistics at local level

- During the last decade there has been a rising interest for what concern poverty mapping
- The measure of Poverty (absolute and relative) play a central role for the policy makers
- Often, poverty occurs locally while official statics are released only for large domain (e.g. NUTS 2)
- Provide a set of reliable poverty indicators at a local level can help to fight social exclusion and deprivation
- Small area poverty mapping try to fill the gap between official statistics and local request of data

Uses of poverty maps

- Guiding intervention mechanisms
- Formulating social and economic policies
- Allocation of government funds
- Regional planning
- Business decision making

Warning 1

maps should reveal intra-regional differences in the distribution of the indicator

Warning 2

integrating maps with other information in decision making process

Demand of statistics at local level

- Available data to measure poverty and living conditions in Italy come mainly from sample surveys, such as the Survey on Income and Living Conditions (EU-SILC)
- However, EU-SILC data can be used to produce accurate estimates only at the NUTS 2 level (that is, regional level)
- To satisfy the increasing demand from official and private institutions of statistical estimates on poverty and living conditions referring to smaller domains (LAU 1 and LAU 2 levels, that is Provinces and Municipalities), there is the need to resort to small area methodologies
- We focus on the estimation of poverty measures, i.e. quantiles, head count ratio, poverty gap and average, at the small area level. For this purpose we use data coming from the EU-SILC survey 2008 and from the Population Census 2001

Remark: Although the 2008 EU- SILC data were collected six years after the census (2008 EU-SILC data refers to 2007), the 2001-2007 period was one of relatively slow growth and low inflation in Italy, so it is reasonable to assume that there was relatively little change

Target population

- The case studies have as target population *Italian or foreign persons who are legally registered as persons established in the region of interest*
- Persons who are not legally living in Italy are excluded
- Persons who live effectively in another region than that of interest but are legally registered in the region of interest are included
- Persons who live in the region of interest but are legally registered to another region are excluded
- Persons who legally live in Italy but are homeless are excluded

Target variable

- Relative poverty measures are related to income or consumption
- Our estimate are based on the *household equivalised disposable income* (target variable)
- Averages, percentiles and poverty indicators are computed on the household equivalised disposable income
- The disposable household equivalised income is computed as

$$\frac{\text{[Disposable household income]} \cdot \text{[Within-household non-response inflation factor]}}{\text{Equivalised household size}}$$

Target variable

- Disposable household income:

  The sum for all household members of gross personal income components *plus* gross income components at household level *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

- Within-household non-response inflation factor:

  Factor by which it is necessary to multiply the total gross income, the total disposable income or the total disposable income before social transfers to compensate the non-response in individual questionnaires

Target variable

- Equivalised household size:
  - Let $HM14+$ be the number of household members aged 14 and over (at the end of income reference period)
  - Let HM_{13-} be the number of household members aged 13 or less(at the end of income reference period)

$$\text{Equivalised household size} = 1 + 0.5 \cdot (HM_{14+} - 1) + 0.3 \cdot HM_{13-}$$

Remark: by this way we take into account the economy of scale present in an household

EU-SILC data

- Data on the equivalised income in 2007 are available from the EU-SILC survey 2008 for 1495 households in the 10 Tuscany Provinces, for 1286 households in the 5 Campania Provinces and for 2274 households in the 11 Lombardia Provinces
- A set of explanatory variables is available for each unit in the population from the Population Census 2001
- We employ an M-quantile model to estimate
  - Head Count Ratio and Poverty Gap at a LAU 1 level (Provinces)
  - 20th Percentile, Median and 80th Percentile at a LAU 1 level (Provinces)
  - Mean at a LAU 1 level (Provinces)
- National poverty line: 9310.74 Euros (equivalised household income)

EU-SILC data

- Remark 1: it is important to underline that EU-SILC data are confidential. These data were provided by ISTAT, the Italian National Institute of Statistics, to the researchers of the SAMPLE project and were analyzed by respecting all confidentiality restrictions
- Remark 2: We chose the Campania, Lombardia and Toscana regions because they are representative respectively of the South, Center and North of Italy
- Remark 3: The choice of three representative regions for North, Center and South of Italy has been driven by the well known North-South divide
- Remark 4: the National poverty line has been computed as the 60% of the median of the household disposable equivalised income in Italy (21 regions)

Target variable statistics

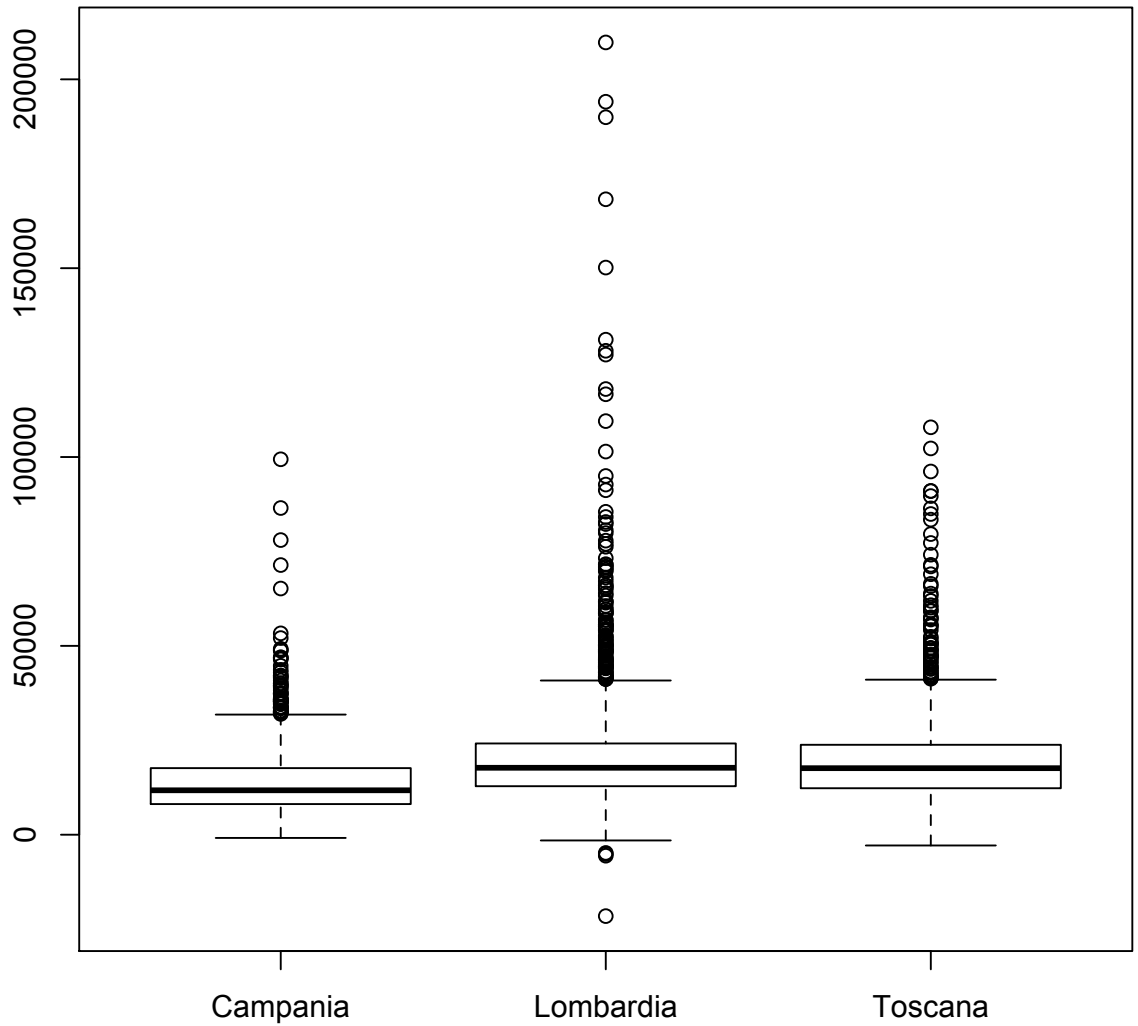


Figure: Boxplots of the disposable equivalised household income

Target variable statistics

- The boxplots show evidence of skew distribution of the household equivalised income with heavy tail on the right in all the three regions
- The boxplots shown evidence of outliers
- Evidence of outliers emerges also from summary statics obtained using the cross-sectional EU-SILC household weights:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Campania	−852.7	8073	11560	13550	17430	99400
Lombardia	−21550.0	12620	17670	20040	24000	209800
Toscana	−2849.0	12120	17230	19430	23570	107900

Table: Summary statistics for the household equivalised income

Census 2001 data

- Italian Census 2001 was collected by ISTAT
- Campania region accounts for 1,862,855 households
- Lombardia region accounts for 3,652,944 households
- Toscana region accounts for 1,388,252 households
- Available variables: 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), gender of the head of the household, civil status of the head of the household, latitude and longitude of the centroid of the household municipality

Remark: it is important to underline that Census data are confidential. These data were provided by ISTAT, the Italian National Institute of Statistics, to the researchers of the SAMPLE project and were analyzed by respecting all confidentiality restrictions

Census 2001 data. Campania

- Distribution of the households and the sampled households over provinces in the Campania region:

Province	Households	Sampled
CASERTA	279,684	128
BENEVENTO	102,441	53
NAPOLI	969,310	810
AVELLINO	152,340	97
SALERNO	359,080	198

Table: Households: number of household in 2001 Census, Sampled: number of households sampled in EU-SILC 2008 survey

Remark: some domains, e.g. the province of Napoli, have a sample dimension that allow for reliable direct estimates

Census 2001 data. Toscana

- Distribution of the households and the sampled households over provinces in the Toscana region:

Province	Households	Sampled
MASSA CARRARA	80,810	105
LUCCA	146,117	150
PISTOIA	104,466	136
FIRENZE	376,255	415
LIVORNO	133,729	105
PISA	150,259	149
AREZZO	123,880	143
SIENA	101,399	104
GROSSETO	87,720	65
PRATO	83,617	123

Table: Households: number of household in 2001 Census, Sampled: number of households sampled in EU-SILC 2008 survey

Census 2001 data. Lombardia

- Distribution of the households and the sampled households over provinces in the Lombardia region:

Province	Households	Sampled
VARESE	320,899	305
COMO	210,587	153
SONDRIO	69,817	29
MILANO	1,545,502	824
BERGAMO	375,778	219
BRESCIA	437,706	265
PAVIA	211,786	56
CREMONA	135,321	82
MANTOVA	146,249	168
LECCO	121,321	114
LODI	77,978	59

Table: Households: number of household in 2001 Census, Sampled: number of households sampled in EU-SILC 2008 survey

Part II

Poverty Mapping in Three Representative Italian Regions

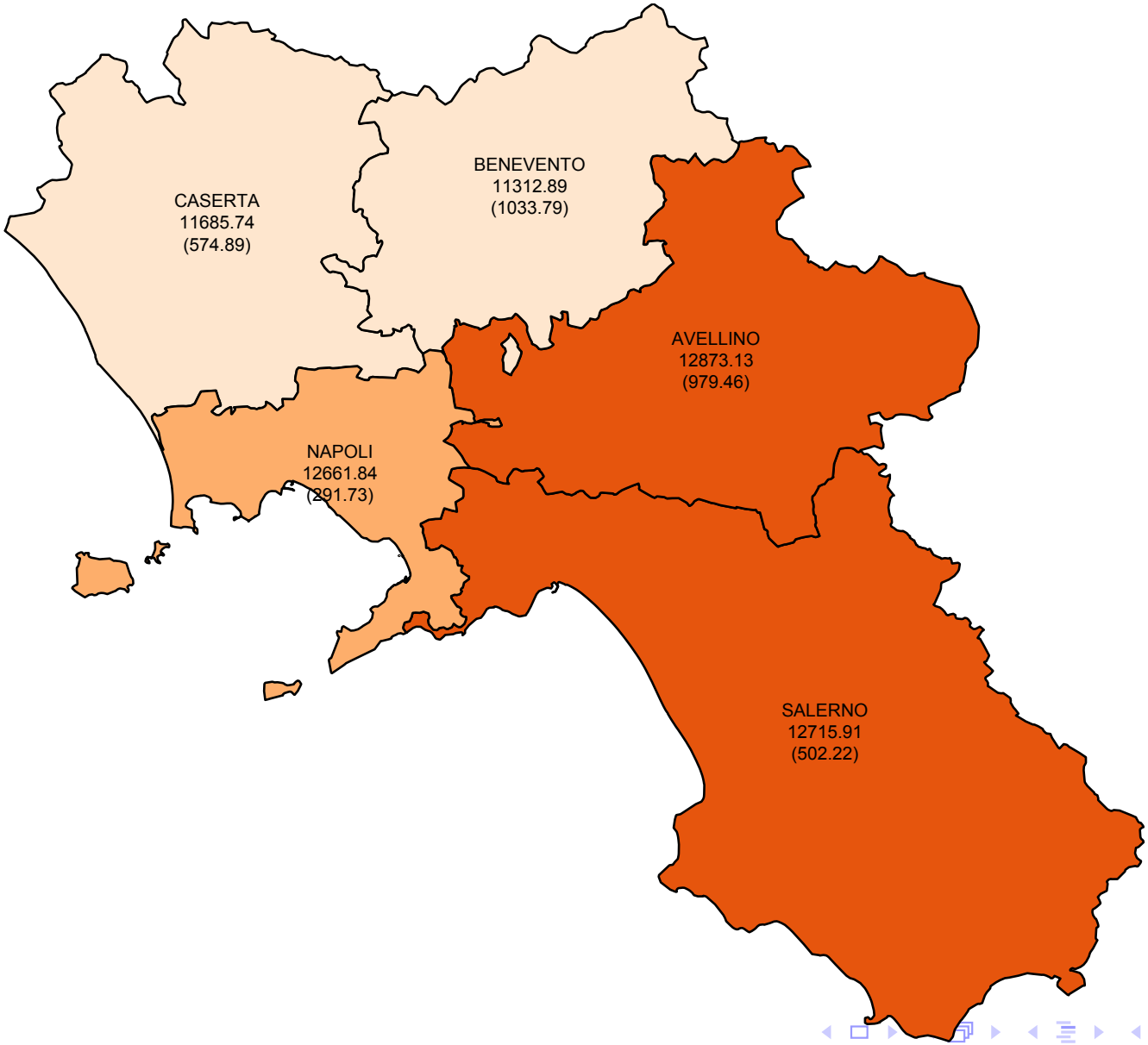
Model Specifications

- We used the M-quantile linear model to compute indicators at LAU 1 level (Provinces)
- 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)
  - gender of the head of the household

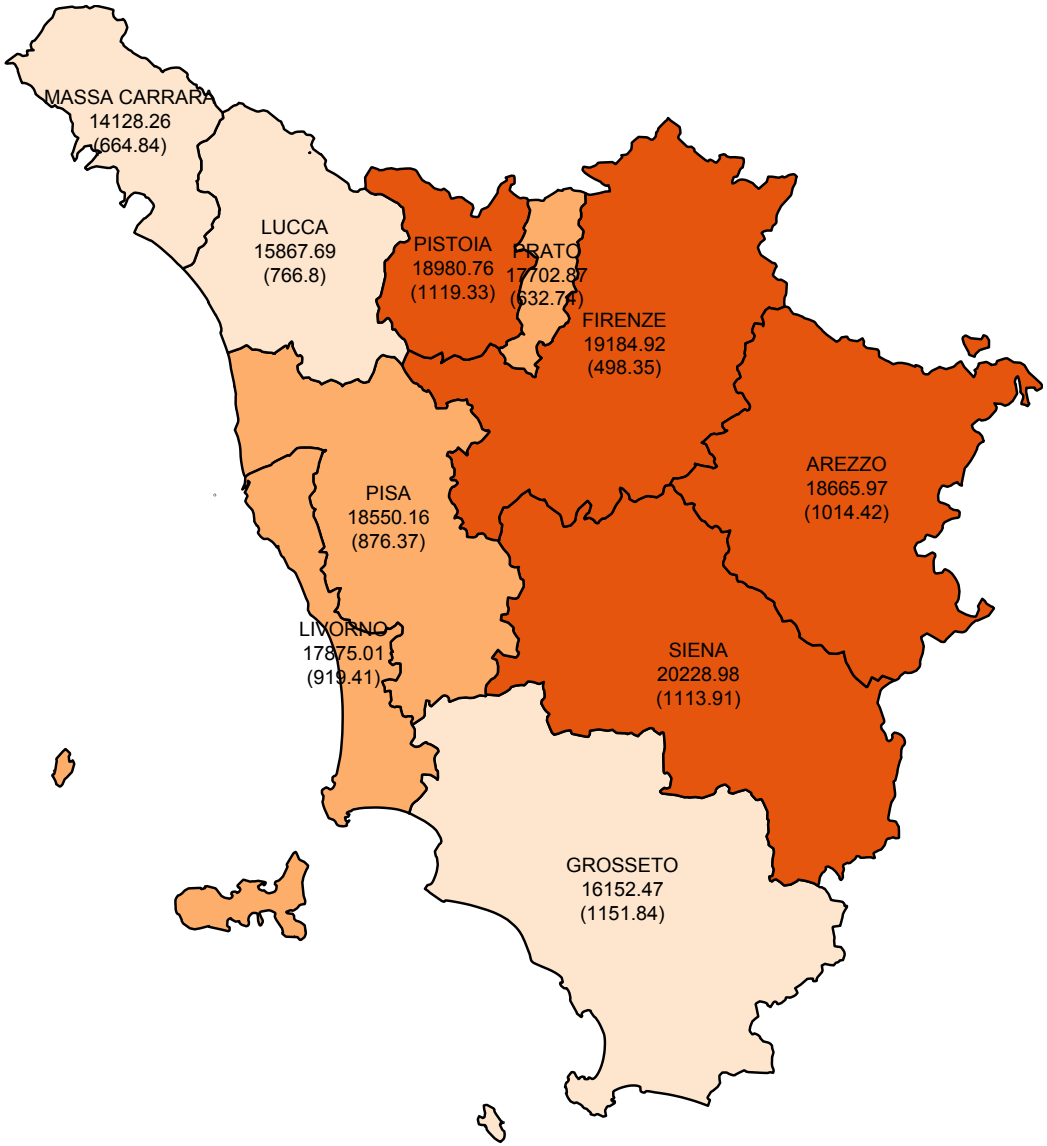
Preliminary Modelling of the Italian EU-SILC

- Working model: 2-level random effects (Province)
- Outcome variable: Equivalised income
- Covariates: Household and head of household variables
- Percentage of variability explained - 18 per cent
- Intracluster correlation coefficients - 4.5 per cent
- Normal QQ plots show departures from normality

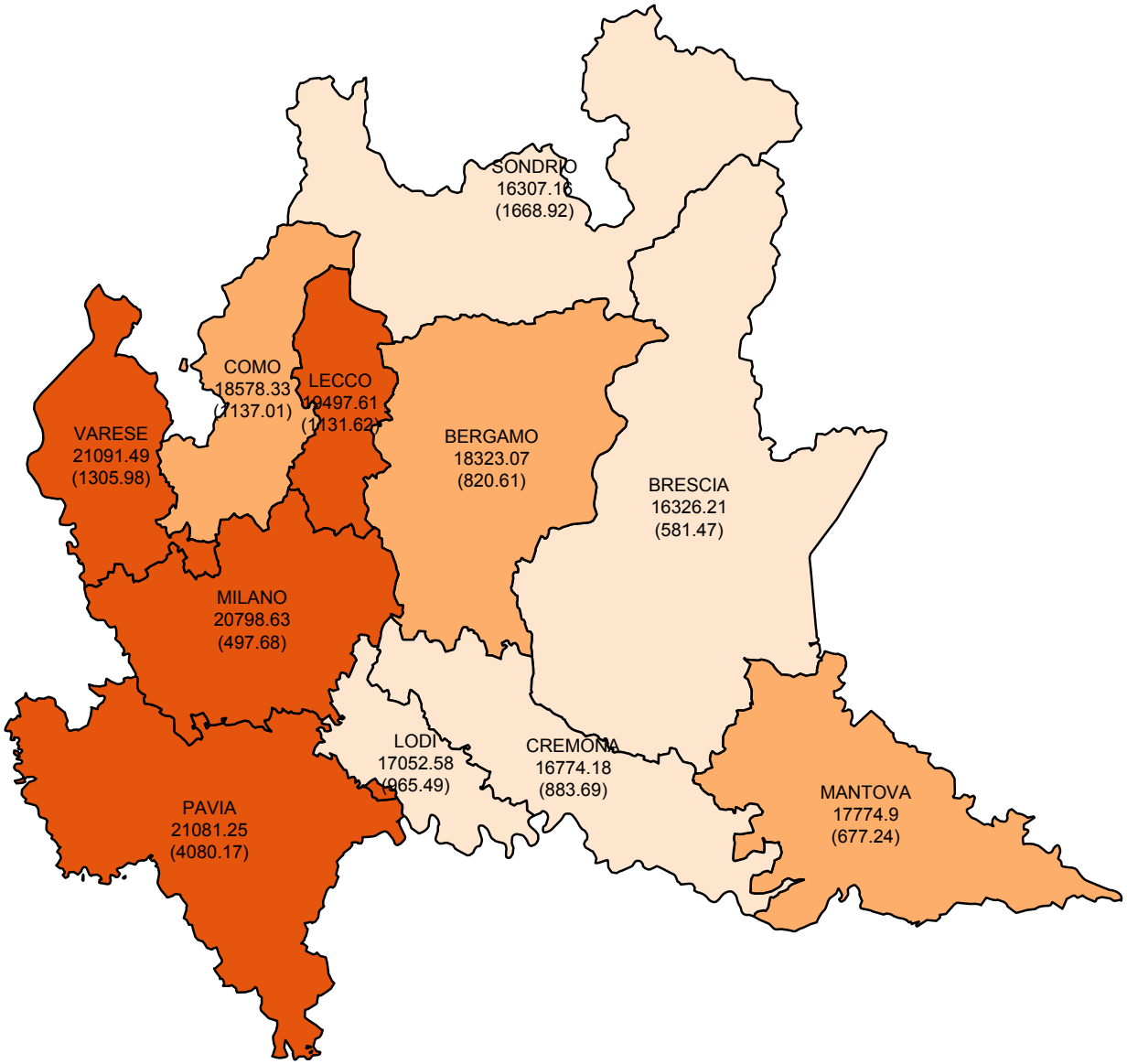
Estimates of the Income Average for Campania Provinces



Estimates of the Income Average for Toscana Provinces



Estimates of the Income Average for Lombardia Provinces



Income Averages in Provinces

The estimates of the average income for each province show that there are intra-regional differences

- Lombardia: the provinces of Milano, Pavia and Varese have the highest mean equivalised household income while the provinces of Sondrio, Cremona and Brescia have lower average income.
- Toscana: the provinces of Siena and Firenze appear to be as wealthy as the wealthier provinces of Lombardia whereas the provinces of Lucca and Massa-Cararra have lower average income
- Campania: compared to Caserta and Benevento, the provinces of Avellino, Salerno and Napoli have higher average income although the intra-regional differences in Campania are not so pronounced.

Income Averages in Provinces

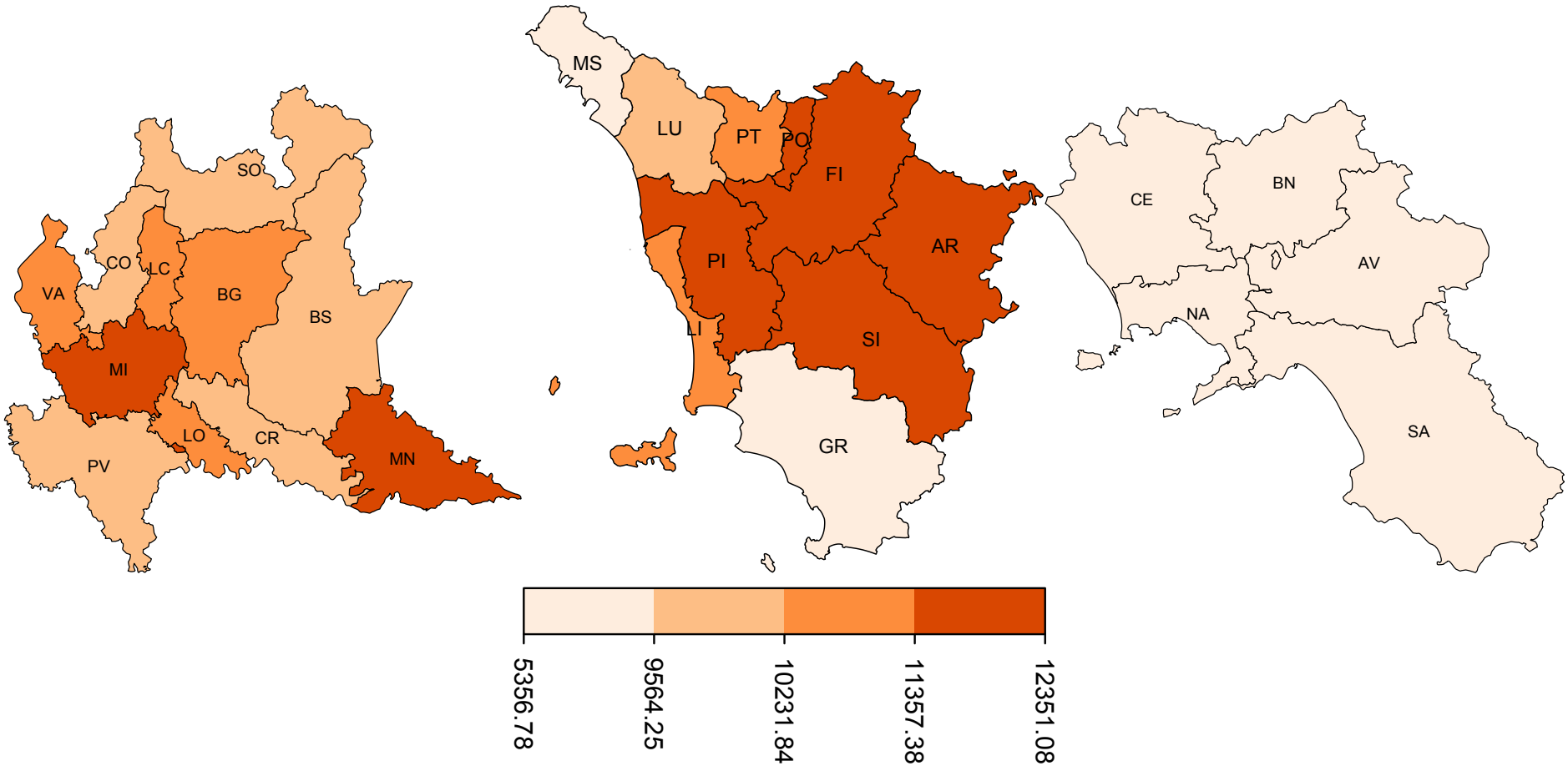
Result 1

Toscana and Lombardia have similar levels of average equivalised household income although, one may say that Lombardia is somewhat wealthier

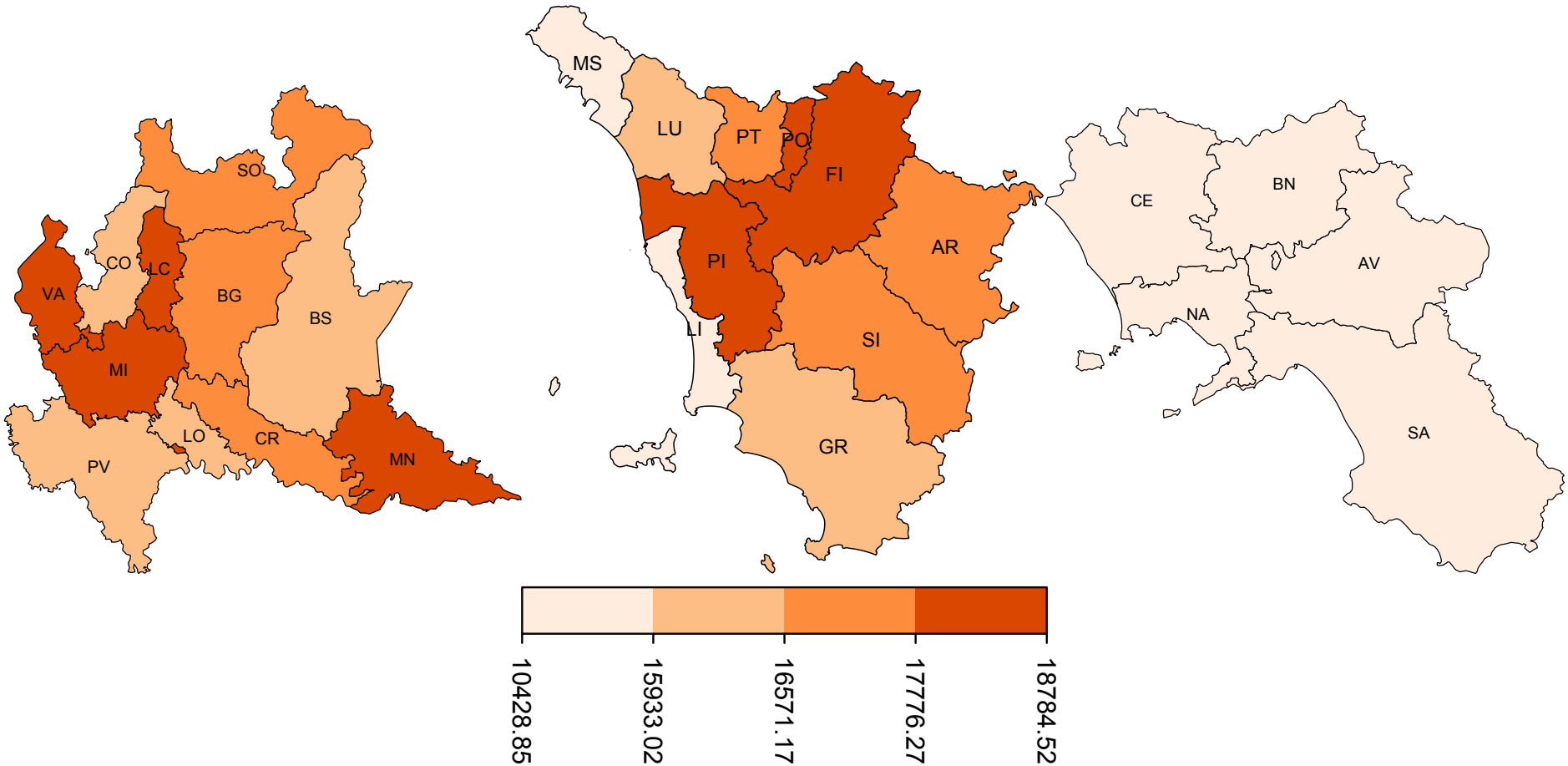
Result 2

Provinces in Campania have smaller average equivalised household income than provinces in Lombardia and Toscana

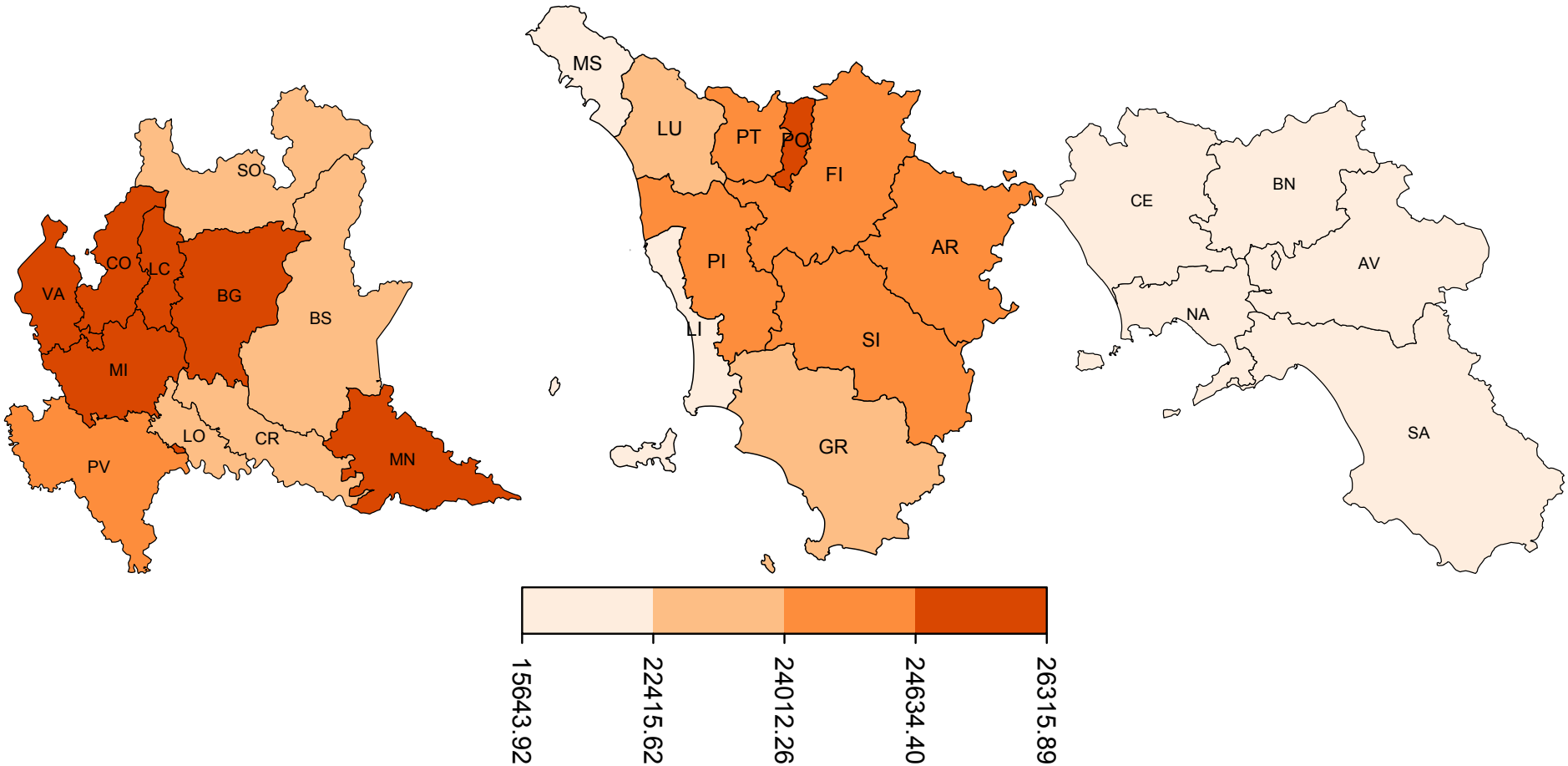
Estimate of the 20th Percentile at Provincial level Lombardia, Tuscany and Campania



Estimate of the Median at Provincial level Lombardia, Tuscany and Campania



Estimate of the 80th Percentile at Provincial level Lombardia, Tuscany and Campania



Income Percentiles in Provinces

Estimated median income in provinces is lower than the corresponding average income. This illustrates the asymmetry of the income distribution and motivates the estimation of small area income distribution functions

- the 20th income percentile in provinces of Lombardia is comparable to the median income of provinces in Campania
- certain provinces in Toscana appear to have higher gap between the 20th and 80th percentiles of income (Massa-Carrara, Lucca and Grosseto). They are similar to Sondrio in Lombardia.
- wealthier provinces: some provinces of Tuscany are comparable to the province of Milano
- Lombardia: very wealthy provinces in terms of average income have a wide gap between the 20th and 80th percentiles (Pavia and Lecco)

Income Averages in Provinces

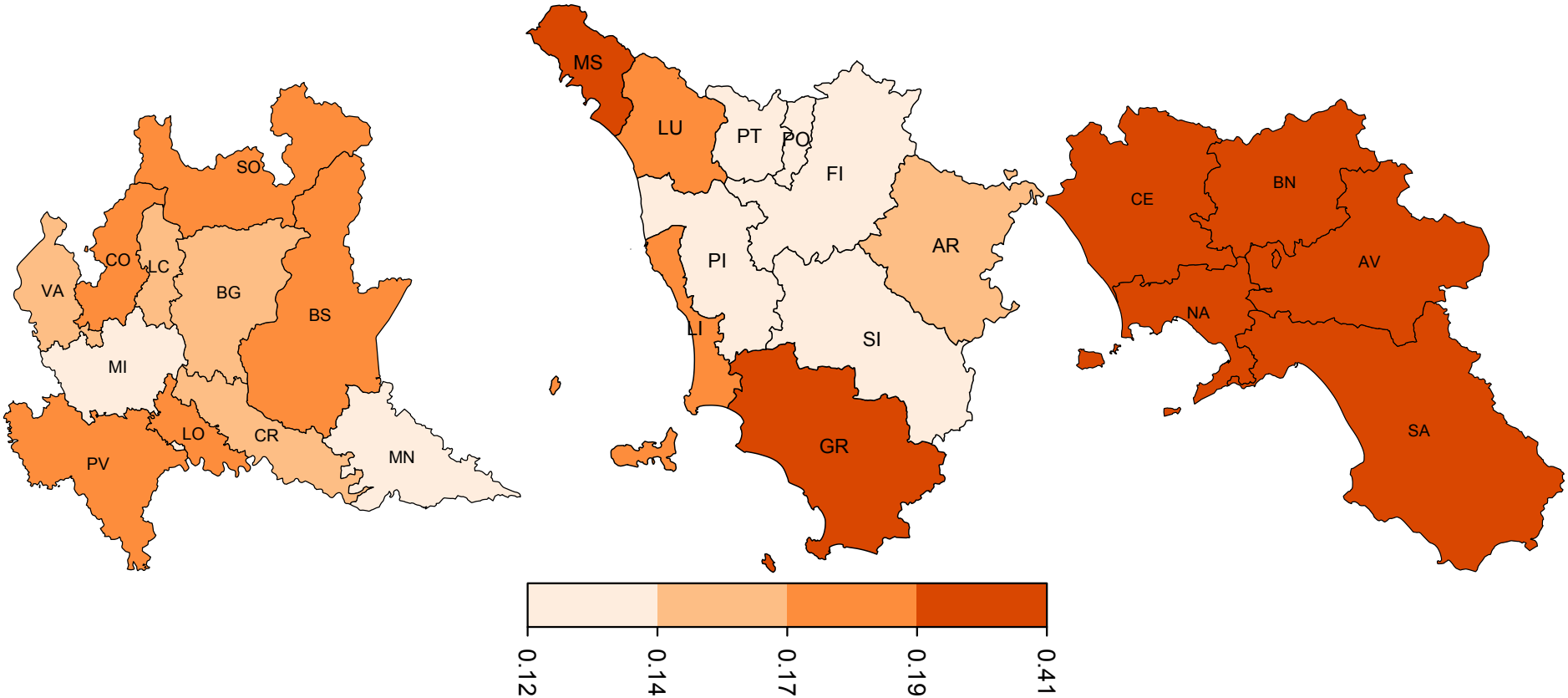
Result 1

The comparison of Lombardia, Toscana and Campania is easier using Percentiles and Averages

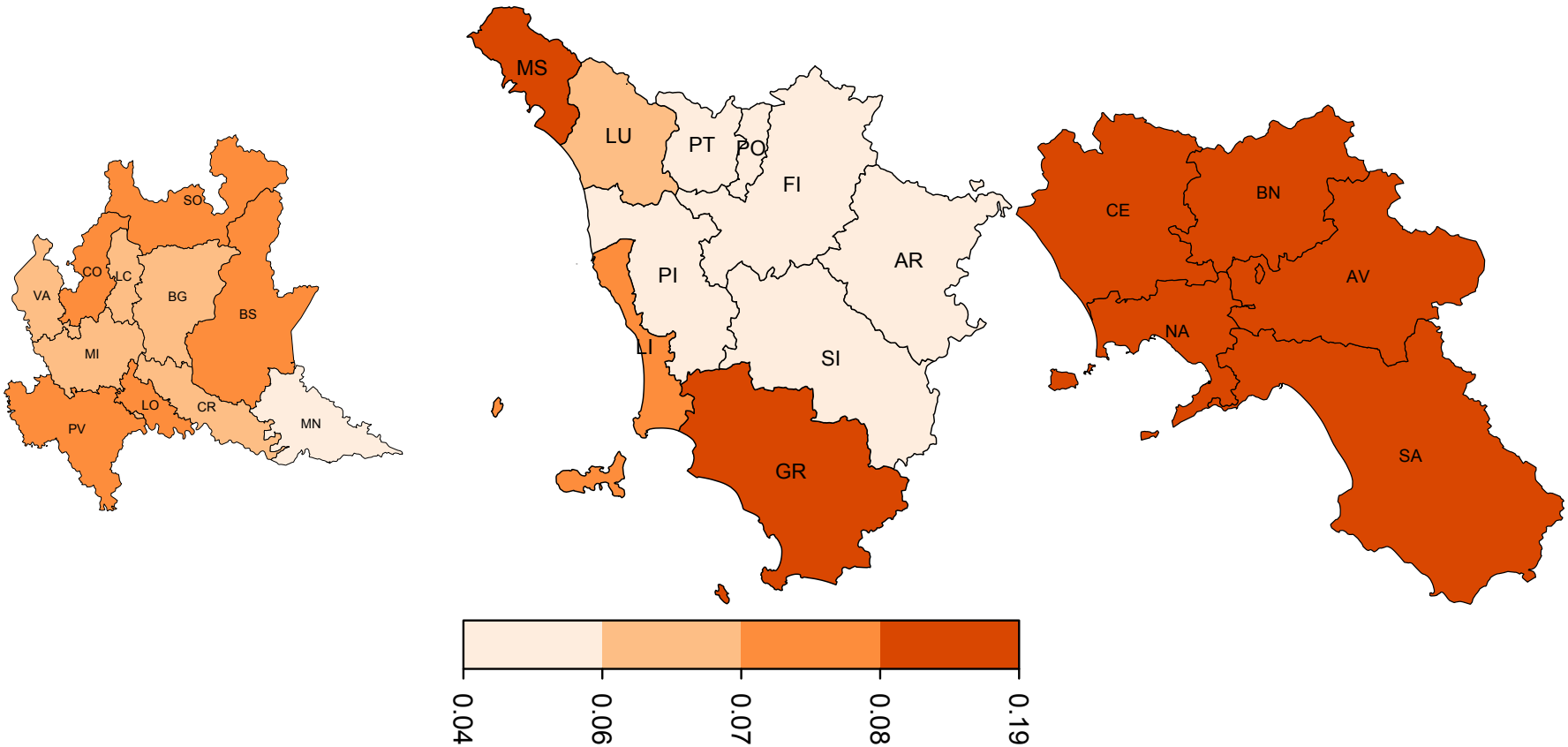
Result 2

Relying solely on estimates of average income does not always provide an accurate picture of the wealth of a small area

Estimate of the Head Count Ratio at Provincial level Lombardia, Tuscany and Campania



Estimate of the Poverty Gap at Provincial level Lombardia, Tuscany and Campania



Modelling EU-SILC data in Tuscany (Italy) to obtain LAU 2 estimates of the income average

Aim: estimate the mean of equivalised household income at Municipality level in Tuscany

- Data on the equivalised income in 2007 for 59 of the 287 Tuscany Municipalities are available from the EU-SILC survey 2008
- A set of explanatory variables is available for all the 287 Municipalities from the Population Census 2001
- We employ the M-Quantile GWR model for estimating the mean of household income in each of the 287 Municipalities (LAU 2)

Remark: *with the spatial information included in the model we can obtain estimates for the 228 out of sample areas*

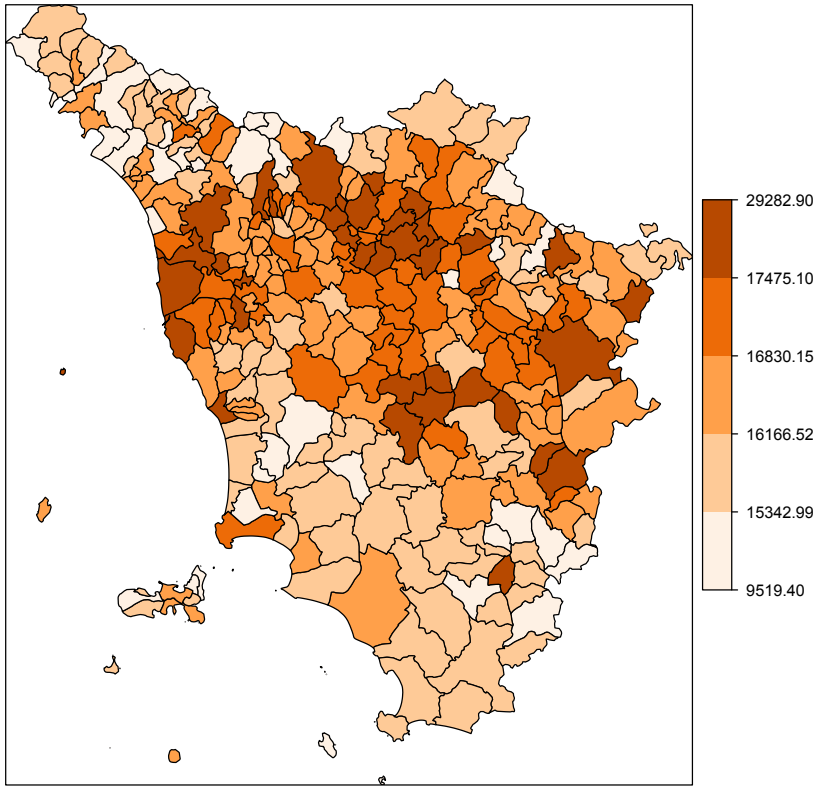
MQGWR Model Specifications

- The selection of covariates to fit the small area models relies on prior studies of poverty assessment and on the availability of data
- The following covariates have been selected:
  - age of householders
  - sex of householders
  - years of education of householders
  - employment status of householders
  - square meters of the house
  - property status of the house (owner or not owner)
  - household size
  - centroids of the Municipalities

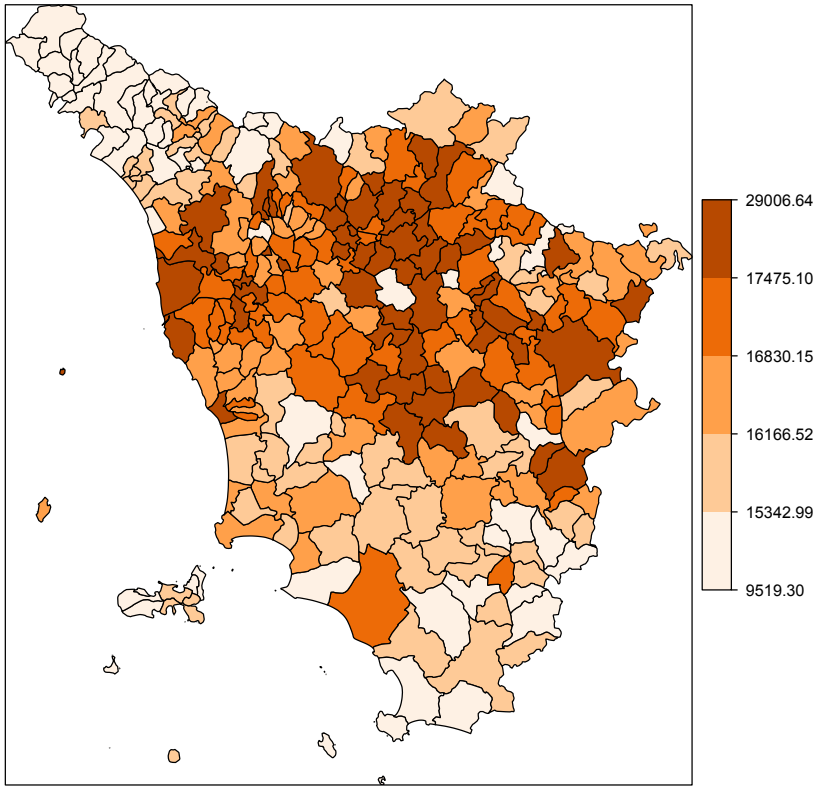
MQGWR Model contrasted with the M-quantile linear model

- The MQGWR model uses spatial information to estimate the target statistics in the out-of-sample areas
- Synthetic estimates in the out-of-sample areas can be obtained using the M-quantile linear model: this can be done letting the area representative quantile, θ_i, be equal to 0.5
- The next map shows the MQGWR estimates contrasted with the M-quantile estimates of the average of the equivalised household income at municipalities (LAU 2) level

Estimate of the Mean Equivalised Income at Municipality level in Tuscany



MQ



MQGWR

Figure: MQ: M-quantile estimates (obtained letting $\theta_i = 0.5, i \in$ out of sample areas), MQGWR: M-quantile GWR estimates (use of spatial information)

Income Averages in Municipalities

Result 1

From Provinces to Municipalities: detailed lower levels of average household income in the North-West and in the South-West of Tuscany

Result 2

detailed high estimates of average household income in some municipalities of the province of Florence, Arezzo and Siena. These results are consistent with the spatial distribution of the average values of household income produced by other nonparametric models (Giusti et al 2011)

Oversampling and Small Area Estimation: A Comparison

When direct estimates are unreliable there are two possible solutions:

- Increase the sample size in the domains of interest in such a way that direct estimates became reliable (oversampling solution)
- Use of small area methods (small area solution)

In order to make a comparison between these alternatives we can take the opportunity to use data referring to an EU-SILC 2008 oversampling of households for the Province of Pisa

- Sample size for the province of Pisa EU-SILC 2008: 149 households
- Sample size for the province of Pisa Oversample: 675 households (that include the 149 household of the EU-SILC survey)

Remark: Oversample has been managed by the ISTAT who warrantees the high quality of the data

Oversampling and Small Area Estimation: A Comparison

We estimate the Mean, the Median, the Head Count Ratio and the Poverty Gap with

- Direct estimators based on the EU-SILC survey data (149 observations), θ^{Dir}
- Direct estimators based on the Oversampling data (675 observations), θ^{Over}
- M-quantile small are estimators based on the EU-SILC survey data (149 observations for the Province of Pisa, 1495 observations for all the Tuscany region), θ^{MQ}

	Mean	Median	HCR	PG
θ^{Dir}	19472.92 (889.74)	18293.43 (915.35)	11.02 (2.73)	4.40 (1.52)
θ^{Over}	18819.62 (695.85)	16706.96 (564.74)	13.59 (1.72)	4.20 (0.97)
θ^{MQ}	19148.60 (842.12)	17937.92 (646.72)	13.27 (1.51)	5.05 (0.82)

Table: Comparison between direct estimates and small area estimates for the Province of Pisa

Scenarios for the Simulations

Scenario 1

Clusters coincide with small areas

- The ELL method uses the correct model
- However, the ELL provides a synthetic estimator

Scenario 2

Clusters do not coincide with small areas

- When true area effects exist, ELL does not account for between area variation

For both scenarios EBP and M-quantile methods account appropriately for between area variation

Scenarios for the Simulations

Targets

- Assess the properties of point and MSE estimators

Scenarios

- Scenario 1: $\log(y_{ij}) = 20 - x_{1ij} - 0.05 \cdot x_{2ij} + v_i + \epsilon_{ij}$
- $v_i \sim N(0, 0.8), \quad \epsilon_{ij} \sim N(0, 2)$
- $m = 30$, $\min(n_i) = 8$, $\max(n_i) = 34$, Monte Carlo runs 500

- Scenario 2: $y_{ij} = 3000 - 150x_{1ij} + \gamma_i + \epsilon_{ij}$
- $v_i \sim N(0, 200), \quad \epsilon_{ij} \sim (1-\gamma)N(0, 800) + \gamma N(0, 4000) \quad \gamma = 0.01$
- $m = 30$, $\min(n_i) = 23$, $\max(n_i) = 45$, Monte Carlo runs 500

Empirical RMSEs

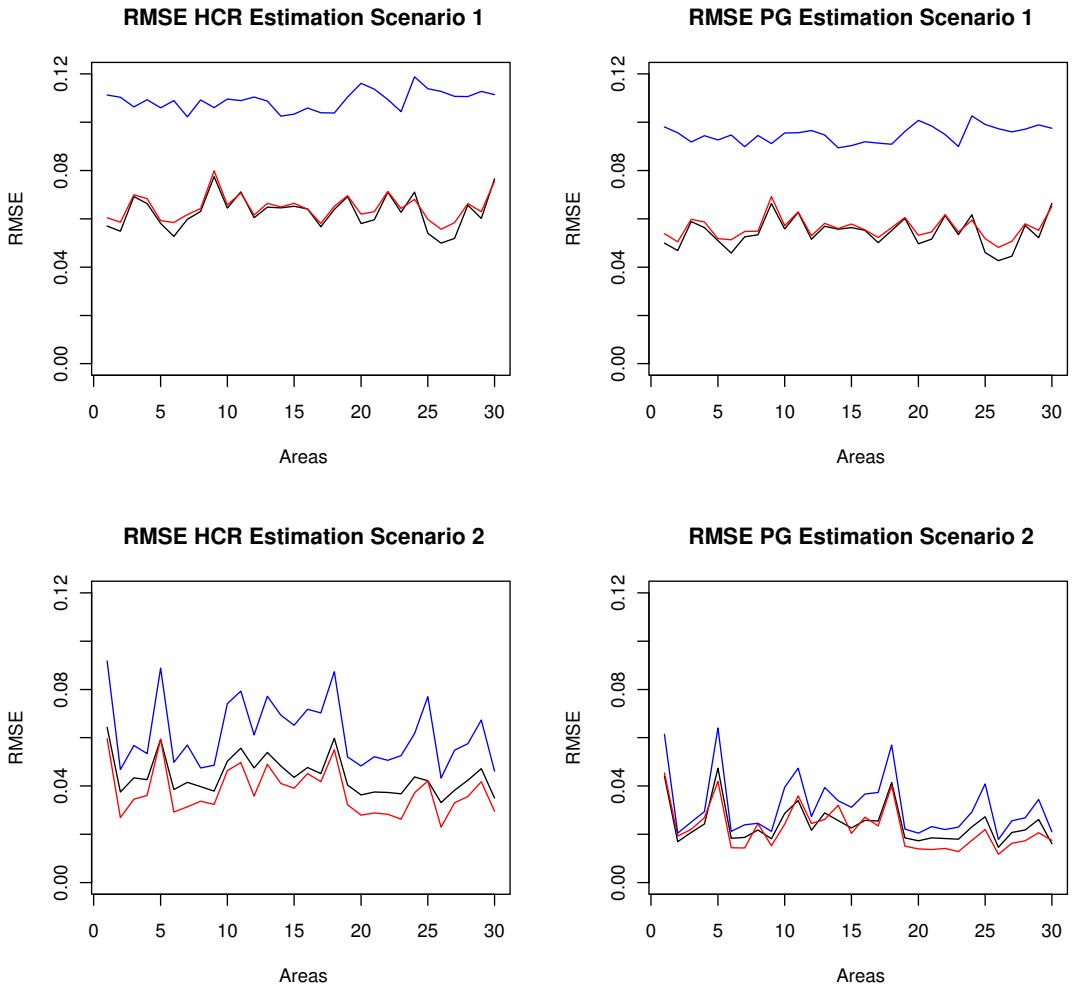


Figure: Simulations results. — = EBP, — = ELL, — = M-quantile

Scenario 1 - MSE Estimation

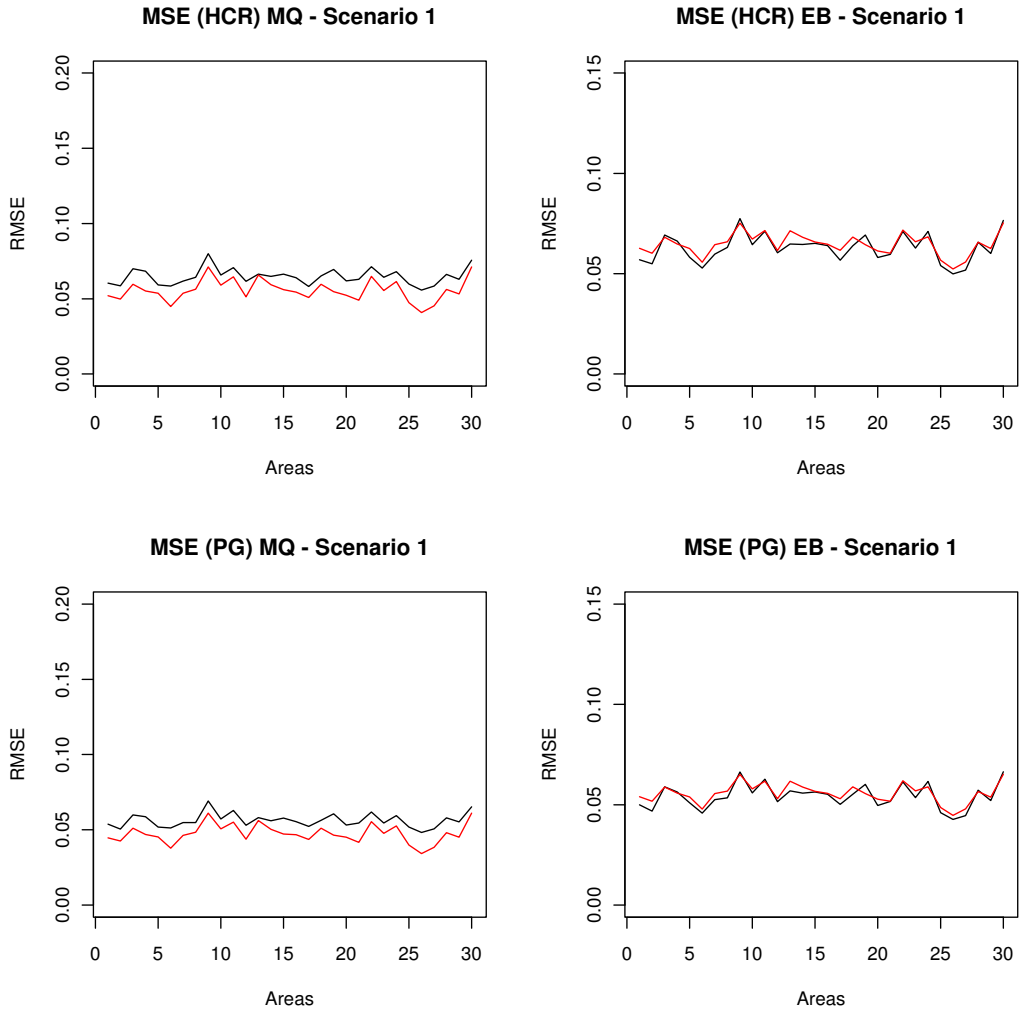


Figure: Simulations results. — = Empirical RMSE, — = Bootstrap RMSE estimates

Scenario 2 - MSE Estimation

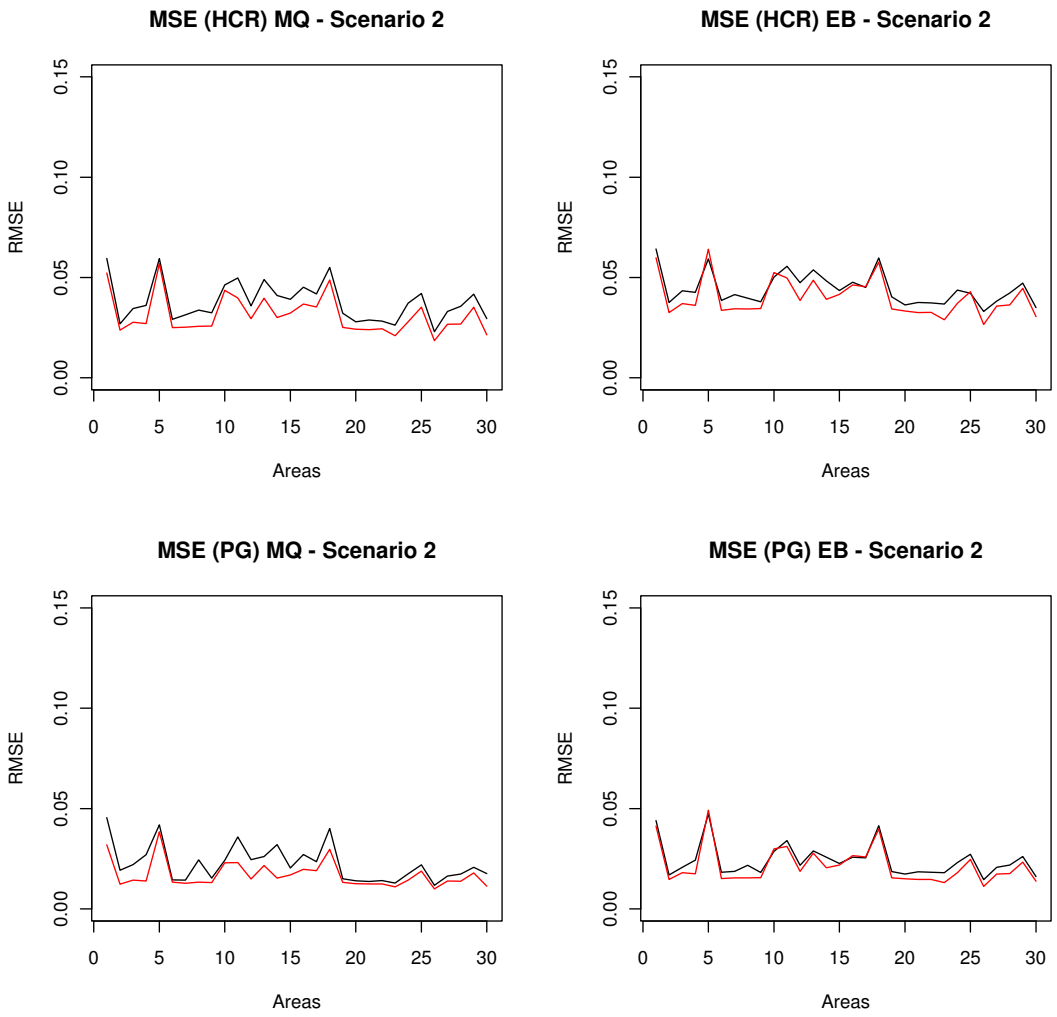


Figure: Simulations results. — = Empirical RMSE, — = Bootstrap RMSE estimates

Discussion

- More realistic evaluations. Current framework is unfair for the ELL approach
- PSUs cutting across areas. Complex covariance structures
- Ignorable sampling design (EBP and MQ). Availability of design (stratification/cluster) variables
- Multivariate extensions
- Comparisons of alternative robust approaches: 1. change the parametric assumptions of the model, 2. keep model assumptions and control for outliers

Part III

Concluding remarks

Concluding remarks

- Focus on poverty mapping in three representative Italian regions: Campania (South), Toscana (Center) and Lombardia (North)
- Domain of interest: provinces (LAU 1) and municipalities (LAU 2)
- Small area methods play a crucial role in providing poverty measures at local level
- Using small area techniques decision makers can have local poverty measures almost costless
- Increase sample size can be an alternative to small area estimation but it is extremely expansive and, in general, there are no resources in terms of time and money to undertake this solution

Concluding remarks

- Results show a remarkable difference in terms of poverty between Campania region and Toscana and Lombardia regions
- Toscana and Lombardia regions are similar with respect to poverty measures.
- In the Toscana region there are evidence of relevant (relative) poverty in the province of Massa-Carrara and Grosseto while in Lombardia there are no criticism like these
- Campania region is poorer than the other two region. This situation is well known in Italy. However if analyzed as stand alone region we can see dissimilarities between provinces
- Using spatial information we obtained estimates of the averages of the households equivalised income at LAU 2 level in Tuscany: looking at the estimates emerges some dissimilarities between the provinces. This results show the importance to "go deeper", i.e. obtain estimates at the lowest domain level

Concluding remarks

- In this presentation we have shown only part of the results obtained using small area estimation methods
- The focus has been done on the M-quantile method that is an alternative to the Mixed Effect model based methods and the World Bank (ELL) method
- An alternative method based on the M-quantile is the nonparametric M-quantile: the relationship between auxiliary variables and response variable is handle via p-spline
- Methods based on mixed effect model were not presented here, however for some of these methods estimates are available upon request

Essential bibliography

- Breckling, J. and Chambers, R. (1988). M -quantiles. *Biometrika*, **75**, 761–771.
- Brunsdon, C., Fotheringham, A.S. and Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**, 281–298.
- 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.L., Chandra, H., Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, **37**, 153–170.
- Cheli B. and Lemmi, A. (1995). A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. *Economic Notes*, **24**, 115–134.
- Elbers, C., Lanjouw, J. O., Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.
- Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). Geographically Weighted Regression. John Wiley and Sons, West Sussex.
- Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.
- Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68**, 2, 221–238.
- Lombardia M.J., Gonzalez-Manteiga W. and Prada-Sanchez J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of finite population distribution function. *Journal of Statistical Planning and Inference*, **116**, 367–388.
- Marchetti, S., Tzavidis, N. and Pratesi, P. (2012). Nonparametric Bootstrap Mean Squared Error Estimation for M-quantile Estimators for Small Area Averages, Quantiles and Poverty Indicators. *Computational Statistics and Data Analysis*, **56**, 2889–2902.
- Royall, R. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, **73**, 351–358.
- Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2010). Small Area Estimation Via M-quantile Geographically Weighted Regression. [Paper submitted for publication in TEST]
- Salvati, N., Chandra, H., Ranalli, M.G. and Chambers, R. (2010). Small Area Estimation Using a Nonparametric Model Based Direct Estimator. *Journal of Computational Statistics and Data Analysis*, **54**, 2159-2171.
- Tzavidis N., Marchetti S. and Chambers R. (2010). Robust estimation of small area means and quantiles. *Australian and New Zealand Journal of Statistics*, **52**, 2, 167–186.
- Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2007). M-quantile models for poverty mapping. *Statistical Methods & Applications*, **17**, 393–411.

SAMPLE Project and Small Area Estimation Software

Prof. Monica Pratesi

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

Valmiera, 24-28 August 2012

Structure of the Presentation

Part I

The SAMPLE Project

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
- Ending date: 28^{th} February 2011
- Partners:
  - University of Pisa (Coordinator)
  - University of Siena
  - University of Manchester / University of Southampton
  - Universidad Carlos III de Madrid
  - Universidad Miguel Hernandez de Elche
  - Warsaw School of Economics
  - Province of Pisa
  - Simurg Ricerche
  - Glowny Urzad Statystyczny
- Web-site: www.sample-project.eu

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);
- to develop instruments (software, questionnaires, etc.) to aid the implementation of the proposed indicators and procedures.

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
- WP 2 Small area estimation of poverty and inequality indicators
- WP 3 Integration of EU-SILC data with administrative data
- WP 4 Standardisation and application development - Software for living conditions estimates
- WP 5 Management
- WP 6 Information, dissemination of results

CONTRIBUTO EUROPEO: 774.972 EURO
DURATA PROGETTO: 36 mesi: 1 marzo 2008 - 1 marzo 2011

COORDINAMENTO

Prof. Monica Pratesi, e-mail: coordinator@sample-project.eu

UNIVERSITA' DI PISA
DIPARTIMENTO DI STATISTICA
E MATEMATICA APPLICATA ALL'ECONOMIA
Via Cosimo Ridolfi, 10 56124 PISA - ITALY
TEL +39 (0) 50 2216391
FAX +39 (0) 50 2216375
http://www.dipstat.ec.unipi.it

PARTERNARIATO

CRIDIRE - Università di Siena - Centro Interdipartimentale di Ricerca sulla Distribuzione del Reddito (Italia)

CCSR - Cathie Marsh Centre for Census and Survey Research, University of Manchester (UK)

UC3M - Departamento de Estadística, Universidad Carlos III de Madrid (Spagna)

UMH - Centro de Investigación Operativa, Universidad Miguel Hernandez de Elche (Spagna)

WSE - Warsaw School of Economics (Polonia)

PP - Provincia di Pisa - Osservatorio per le Politiche Sociali - Ufficio Politiche Comunitarie (Italia)

SR - Simurg Ricerche (Italia)

GUS-CES - Centre of Statistical Training - Central Statistics Office (Polonia)

Che cos'è Sample?

Small Area Methods for Poverty and Living Condition Estimates

SAMPLE
Small Area Methods for Poverty and Living Condition Estimates
Proj. n° 217565 - FP7 - SSH-2007 - 1

COMMISSIONE EUROPEA

SEVENTH FRAMEWORK PROGRAMME

www.sample-project.eu

Main results of the sample project

- State of the art report on small area estimation and poverty indicators
- Methodological development on small area estimation
  - M-quantile approach development
  - Mixed model approach development
- Data integration
- Software application for stakeholders
- Small area estimation software
- Organization of national and international events and conferences to disseminate project results

Part II

Small Area Estimation Software

Small Area Estimation Software

- One of the goal of the SAMPLE project was to produce functions for small area estimation
- The SAMPLE project partners involved in the small area estimation software development were:
  - Mixed model approach: University of Miguel Hernandez de Elche, University Carlos III de Madrid and marginally the University of Pisa
  - M-quantile approach: University of Pisa and University of Manchester/Southampton
- Basic small area estimation functions, such as the EBLUP and EBP estimation functions, has been included in the SAMPLE Software Application
- Functions are developed for the R statistical software

Mixed model approach functions

Area-level small area estimation functions:

- Fay-Herriot model
- Area-level spatial model
- Area-level time model
- Area-level partitioned time models
- Area-level spatio-temporal models

Unit-level small area estimation functions:

- Unit-level time models
- EB prediction of poverty measures with unit level models
- Fast EB methods for estimation of fuzzy poverty measures

M-quantile approach functions

- M-quantile small area estimation of the mean
- Nonparametric M-quantile estimation of the mean
- M-quantile geographically weighted regression
- M-quantile CD estimators of the quantiles
- Nonparametric M-quantile CD estimators of the quantiles
- M-quantile poverty indicators estimators

Remark 1: The M-quantile approach can be used only as unit level model
Remark 2: The functions are written in R language so they are easy to modify but they are not fast!

Focus on R functions of M-quantile linear model estimators

Function to compute the small area averages:
`mq.sae(y,x,x.outs,regioncode.s,regioncode.r,m,p,tol.value, maxit.value,k.value)`

- y: the (numeric) response vector for sampled units
- x :a $n \times p$ matrix of auxiliary variables which also has include a vector of ones for the intercept term
- $x.outs$: covariate information for out of sample units
- $regioncode.s$: area code for sampled units
- $regioncode.r$: area code for out of sample units
- m: the number of small areas
- p: size of $x + 1$ (including the intercept)
- $tol.value$: convergence tolerance limit for the M-quantile model. Default to 0.0001
- $maxit.value$: maximum number of iterations for the iterative weighted least squares. Default to 100
- $k.value$: tuning constant used with the Huber proposal 2 scale estimation. Default to 1.345

Focus on R functions of M-quantile linear model estimators

`mq.sae` function returns the following arguments:

- *mq.cd*: estimates of small area means using the M-quantile Chambers-Dunstan estimator (Tzavidis et al. 2010)
- *mq.naive*: estimates of small area means using the M-quantile naive estimator (Chambers and Tzavidis 2006)
- *mse.cd*: MSE estimates for the M-quantile CD small area means
- *mse.naive*: MSE estimates for the M-quantile naive small area means
- *code.area*: the unique codes of the small areas

Example of the use of MQ.SAE.mean

Generating population data and drawing a sample (R commands)

```
> # MQ-EBLUP
> source("c:\\MQ_sae.R"); library(pps)
> sigmasq.u=3; sigmasq=6
> m=40
> ni=rep(5,m); Ni=rep(100,m); N=sum(Ni); n=sum(ni)
> set.seed(1973)
> u=rnorm(m,0,sqrt(sigmasq.u)); u=rep(u,each=100)
> e=rnorm(N, 0, sqrt(sigmasq))
> gr=rep(1:40,each=100)
> ar=unique(gr)
> uno=matrix(c(rlnorm(N,log(4.5)-0.5,0.5)),nrow=N,ncol=1)
> y=100+5*uno+u+e
> pop.matrix<-cbind(y,uno,gr); pop<-as.data.frame(pop.matrix)
> names(pop)<-c("y","x","area")
> # Drawing a sample
> s=stratsrs(pop$area,ni)
> x.lme=pop[s,]$x
> y.lme=pop[s,]$y
> regioncode.lme=pop[s,]$area
> pop.r<-pop[-s,]
```

Example of the use of MQ.SAE.mean

Example of R code for running function MQ.SAE.mean

```
tmp<-MQ.SAE.mean(y=y.lme,x=x.lme,regioncode.s=regioncode.lme,m=40,
p=2,x.outs=pop.r[,2], regioncode.r=pop.r[,3],tol.value=0.0001,
maxit.value=100,k.value=1.345)
```

Output of function MQ.SAE.mean

```
> tmp

mq.cd
[1] 115.7275 117.9384 115.3374 115.5339 116.3331 ...

mq.naive
 [1] 115.9003 117.6583 115.0192 115.6947 116.0871 ...

mse.cd
 [1] 0.55237498 0.80473242 1.54140859 0.75538562 2.13604316 ...

mse.naive
[1] 0.09710564 0.02790977 0.16425263 0.05226719 0.12559878...

code.area

[1]  1  2  3  4  5...
```

The MQPovertyLib package

- This package provides functions to estimate mean, quantiles and poverty indicators using the M-quantile approach
- It is think to complain with the EU-SILC survey data and with population data such as population censuses
- M-quantile model is estimated at household level but estimates are returned either at household level or at person level
- Poverty indicators computed are the Head Count Ratio (it measures the incidence of poverty) and the Poverty Gap (it measures the intensity of poverty)
- Authors: Stefano Marchetti, Nicola Salvati, Nikos Tzavidis and Caterina Giusti

Remark: the package is still under development and have no warranty and is released upon request to stefano.marchetti@ec.unipi.it

The MQPovertyLib package

The package provides the following functions:

MQ-Poverty-Lib-package	Functions to estimate means, quantiles, HCRs and PGs for Small Areas using the MQ approach
income.example.sae	Simulated income data for 30 domains
mq.coef	It estimates the beta coefficient of each small area
MQ.SAE.mean	It estimates the small area mean
MQ.SAE.mean.pers	It estimates the small area mean at person (unit) level
MQ.SAE.poverty	It estimates the small area Head Count Ratio and the Poverty Gap indicators at household (cluster) level
MQ.SAE.poverty.persons	It estimates the small area Head Count Ratio and the Poverty Gap indicators at person level
MQ.SAE.poverty.smearing	It estimates the small area Head Count Ratio and the Poverty Gap indicators at household (cluster) level
MQ.SAE.quant	It Estimates the small area quantiles
MQ.SAE.quant.pers	It Estimates the small area quantiles at person level
QRLM	M-quantile linear regression model

There is an help for each function, accessible by the standard help R command (i.e. `?MQ.SAE.mean`)

Estimates of Percentiles

Function used: `MQ.SAE.quant`

- Usage: `MQ.SAE.quant(qgrid, y, x, X, regioncode, regioncodepop, adjseed = max(0.15, mean(y)/500), MSE = FALSE, B = 50, R = 200, method = "su", maxit = 100)`
- Details: This function uses a linear M-quantile model to estimate small area quantiles. It is a smearing type estimator, Chambers and Dunstan (1986). The root mean squared error of this small area estimator is estimated via a specific bootstrap technique, see Marchetti et al. (2012).
- Arguments: `qgrid`, quantiles to be estimated (values from 0 to 1); `y`, the study variable; `x`, design matrix of sample auxiliary variables; `X`, design matrix of population auxiliary variables; `regioncode`, vector of areas IDs for sampled observations; `regioncodepop` vector of areas IDs for population observations; `adjseed`, default to max(0.15, $mean(y)/500$), to be changed if the quantile estimation algorithm do not converge; `MSE`, if set to TRUE estimates the RMSE of the HCR and PG estimators; `B`, set the number of bootstrap population to be generated in the bootstrap procedure; `R`, set the number of bootstrap samples to be drown for each bootstrap population in the bootstrap procedure; `method`, set the method to generate the bootstrap population: "eu" empirical unconditional method, "ec" empirical conditional method, "su" smooth unconditional method, "sc" smooth conditional method; `maxit`, number of maximum iteration in the iterated weighted least squares betas estimation procedure

Estimates of Percentiles

The function returns:

- `quantiles`: The quantiles estimates for each area
- `rmse`: The estimated root mean squared error of the quantiles estimates
- `Area.Code`: The IDs of the small areas

Data used:

- `y`: Household equivalised income, EU-SILC 2008
- `x`: Household size, ownership of dwelling, age, working position and gender of the head of the household, EU-SILC 2008
- `X`: Household size, ownership of dwelling and age, working position and gender of the head of the household, Census 2001

Estimates of Poverty Indexes

Function used: `MQ.SAE.poverty`

- Usage: `MQ.SAE.poverty(y, x, X, regioncode, regioncodepop, L = 50, MSE = TRUE, B = 50, R = 200, method = "eu", pov.l = NULL)`
- Details: This function uses a linear M-quantile model to estimate small area HCR and PG indicators. The estimator uses a Monte Carlo approach to account for the prediction error. The root mean squared error of this small area estimators is estimated via a specific bootstrap techniques, see Marchetti et al. (2012)
- Arguments: `y`, the study variable; `x`, design matrix of sample auxiliary variables; `X`, design matrix of population auxiliary variables; `regioncode`, vector of areas IDs for sampled observations; `regioncodepop` vector of areas IDs for population observations; `L`, number of Monte Carlo runs in the estimation procedure; `MSE`, if set to TRUE estimates the RMSE of the HCR and PG estimators; `B`, set the number of bootstrap population to be generated in the bootstrap procedure; `R`, set the number of bootstrap samples to be drown for each bootstrap population in the bootstrap procedure; `method`, set the method to generate the bootstrap population: "eu" empirical unconditional method, "ec" empirical conditional method, "su" smooth unconditional method, "sc" smooth conditional method; `pov.l`, the poverty line value, if it is set to NULL the poverty line is computed as $0.6 \cdot y$

Estimates of Poverty Indexes

The function returns:

- `HCR.MQ`: Estimates of the Head Count Ratio (or At Risk of Poverty Rate)
- `RMSE.HCR.MQ`: Estimates of the root mean squared error of the HCR estimator
- `PG.MQ`: Estimates of the Poverty Gap
- `RMSE.HCR.MQ`: Estimates of the root mean squared error of the PG estimator
- `Area.Code`: The IDs of the small areas

Data used:

- `y`: household equivalised income, EU-SILC 2008
- `x`: household size, ownership of dwelling, age, working position and gender of the head of the household, EU-SILC 2008
- `X`: household size, ownership of dwelling and age, working position and gender of the head of the household, Census 2001

Estimates of Averages

Function used: `MQ.SAE.mean` (it has been already explained)
Data used:

- `y`: household equivalised income, EU-SILC 2008
- `x`: household size, ownership of dwelling, age, working position and gender of the head of the household, EU-SILC 2008
- `x.outs`: household size, ownership of dwelling and age, working position and gender of the head of the household, Census 2001

Part III

Concluding remarks

Concluding remarks

- SAMPLE project: European project focused on small area poverty estimation and data integration
  - Help stakeholders and decision makers providing information at a local level
  - Development of softwares to make accessible small area estimation methods to pratictioners
- Methods and application for data integration aimed to help decision makers
- Small area estimation software for the R statistical software
  - Functions to estimate small area averages, quantiles and poverty indicators
  - Functions that use either M-quantile and mixed model approach to small area estimation
- Bootstrap technique is time consuming. In our experience it can be used for population of no more than four millions of elementary units

Essential bibliography

- Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, 761–771.
- Brunsdon, C., Fotheringham, A.S. and Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**, 281–298.
- 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.L., Chandra, H., Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, **37**, 153–170.
- Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). Geographically Weighted Regression. John Wiley and Sons, West Sussex.
- Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.
- Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68**, 2, 221–238.
- Lombardia M.J., Gonzalez-Manteiga W. and Prada-Sanchez J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of finite population distribution function. *Journal of Statistical Planning and Inference*, **116**, 367–388.
- Marchetti, S., Tzavidis, N. and Pratesi, P. (2012). Nonparametric Bootstrap Mean Squared Error Estimation for M-quantile Estimators for Small Area Averages, Quantiles and Poverty Indicators. *Computational Statistics and Data Analysis*, **56**, 2889–2902.
- R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`
- Royall, R. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, **73**, 351–358.
- Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2010). Small Area Estimation Via M-quantile Geographically Weighted Regression. [Paper submitted for publication in TEST]
- Salvati, N., Chandra, H., Ranalli, M.G. and Chambers, R. (2010). Small Area Estimation Using a Nonparametric Model Based Direct Estimator. *Journal of Computational Statistics and Data Analysis*, **54**, 2159-2171.
- Tzavidis N., Marchetti S. and Chambers R. (2010). Robust estimation of small area means and quantiles. *Australian and New Zealand Journal of Statistics*, **52**, 2, 167–186.
- Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2007). M-quantile models for poverty mapping. *Statistical Methods & Applications*, **17**, 393–411.