Estimation of income, consumptions and living conditions indicators at local area level by SAE methods: ideas for a possible project to do it for prefectures and counties in China

Monica Pratesi[1,2]

[1]Univeristy of Pisa

[2]Centre ASESD Camilo Dagum

13 December 2018

Overview

Local level analysis

Having an estimation of income, consumptions and living conditions indicators at local level it is important:

i) because, usually, the living conditions in different areas within a country or within a region or provinces are heterogeneous;

ii) in order to plan sub-regional policies or evaluating the results of policy.

Reliable local level statistics

- Usually, the income and expenditure data are obtained by sample surveys whose sample size is not enough to obtain reliable estimates at local level.
- Indeed, survey designs usually focuses on achieving a particular degree of precision for estimates at a level of aggregation higher than that of local areas.
- The **sample size is small or equal to zero** in the target areas and direct estimators do not reach a minimum level of precision.

◦ Two choices

- oversampling (is often not possible!)

$\Rightarrow$ statistical techniques that allow for reliable estimates in that local ('small') domains: **Small Area Estimation (SAE) models**.

Small Area Estimation

- The term "small area" is used to describe domains whose **sample sizes are not large enough** to enable sufficiently precise direct estimates.
- Small Area Estimation methods constitute a set of advanced statistical inference techniques that can be used for the measurement of income and living condition indicators.

Direct vs Indirect Estimates

- The **direct estimates** are produced under the design based approach using only data coming from one survey.
- The **indirect estimates** use auxiliary information (such as census or administrative data available at local area level and geo-coded data about the spatial distribution of domains and units) to improve the quality and accuracy of survey estimates or to break down the known values referred to larger areas by using regression-type models. The relationship between the target parameters and the auxiliary variables is described by a suitable model.
  - model-assisted approach
  - model based approach
  - area level models
  - unit level models

How a SAE model works

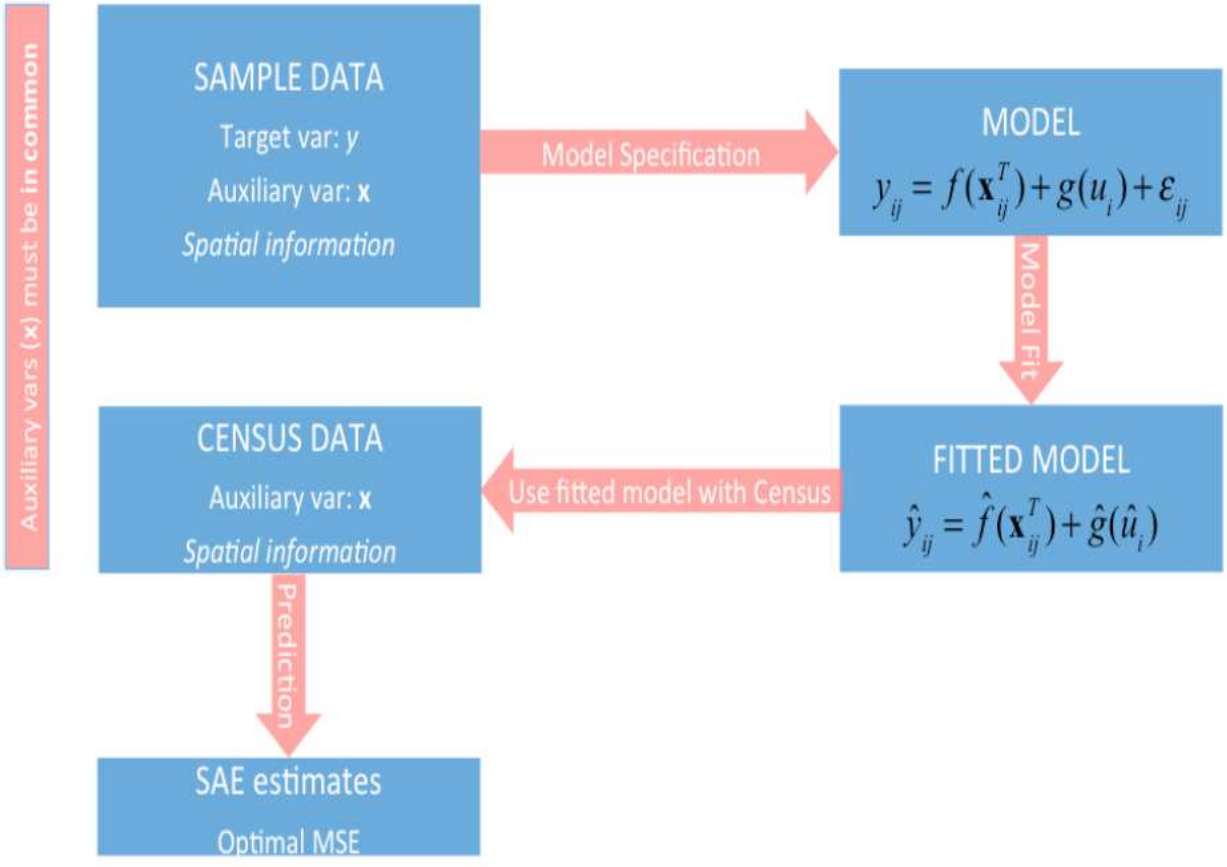

Figure: How a SAE model works. Source: Fao (2015)

Classification of SAE models

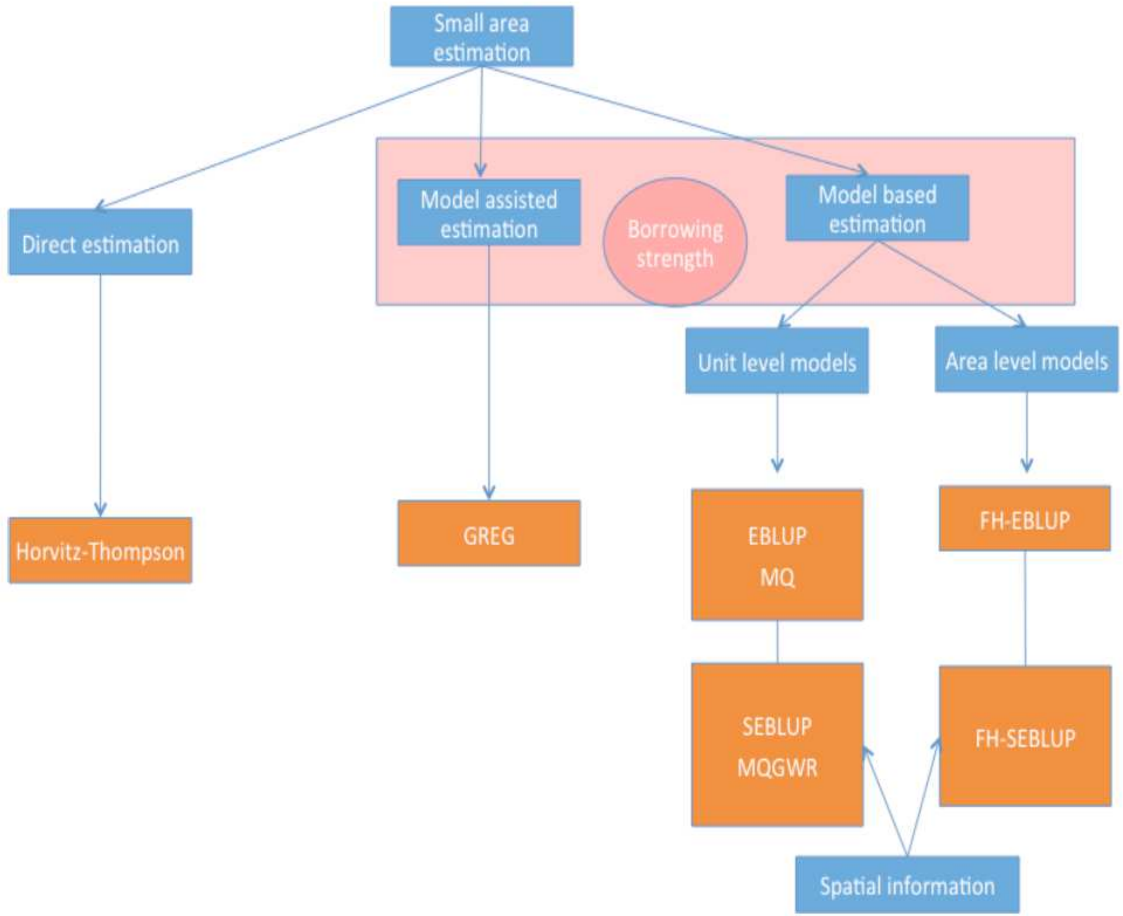

Figure: A classification of the SAE Methods. Source: Fao (2015)

SAE models

- SAE methods **combine survey estimate with auxiliary information to reduce their variability** (hence reduce the CVs).
- Small area model-based estimators explicit small area models that make specific allowance for between area variation.
- In this framework models involve random area-specific effects that account for between area variation beyond that explained by auxiliary variables included in the model.
- Area-level models that relate the small area means to area-specific auxiliary variables are essential if unit level data are not available.

Framework for the production of SAE in Italy using EU-SILC

Target variables: Income or consumption data in order to obtain poverty indicators (such as Head Count Ratio - HCR).
Data availability. EU-SILC provides data at NUTS-2 level (Italian Regions). SAE is needed to obtain estimates at NUTS-3 (110 provinces in Italy).
Example: SAE in Tuscany. Tuscany is a region of Italy (about 60 millions of inhabitants) with 10 provinces and about 4 millions of inhabitants.

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
Italy	22 millions	20,928

Table: Households: number of households in Tuscany provinces in 2001 Census, Sampled: number of households sampled in the provinces of Tuscany in EU-SILC 2008 survey.

Framework for the production of small area official statistics

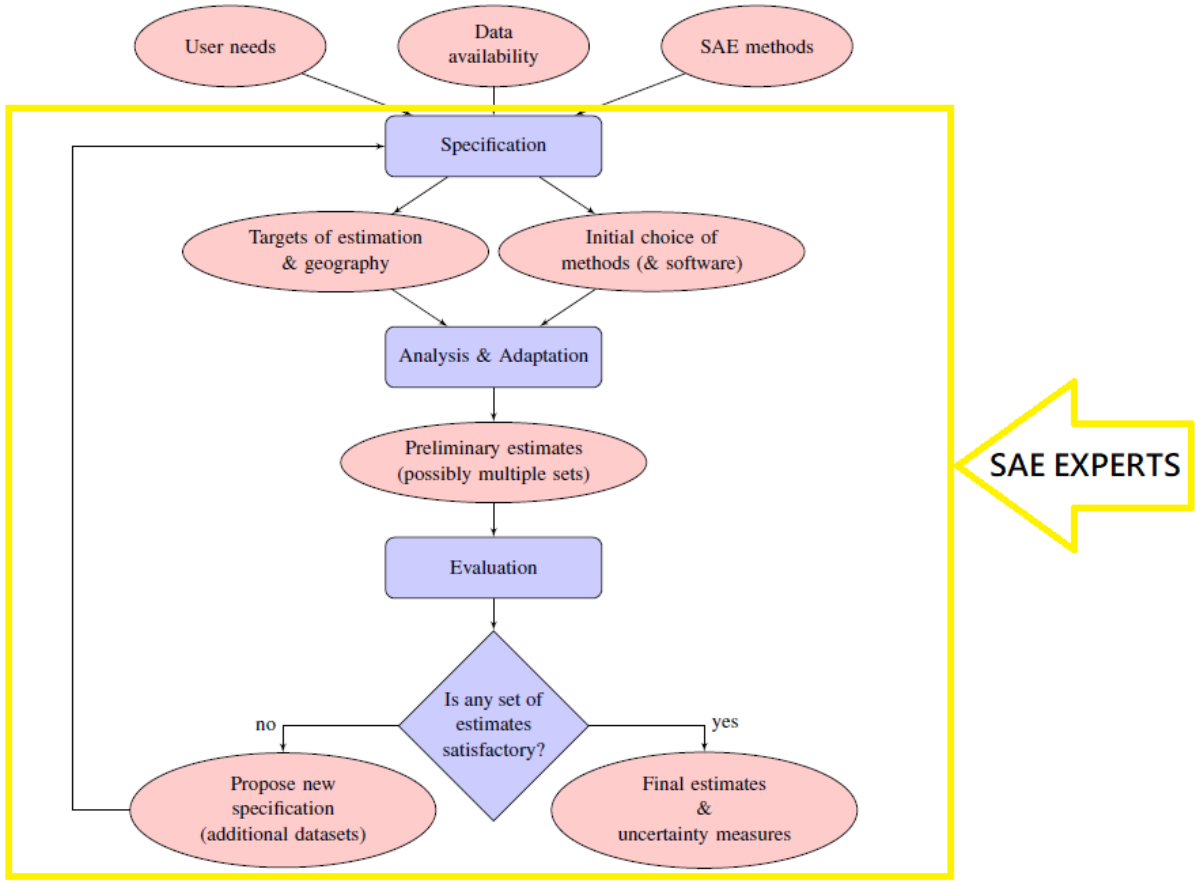

Figure: Framework for the production of small area official statistics.

Geographical inequality of consumption in Italy

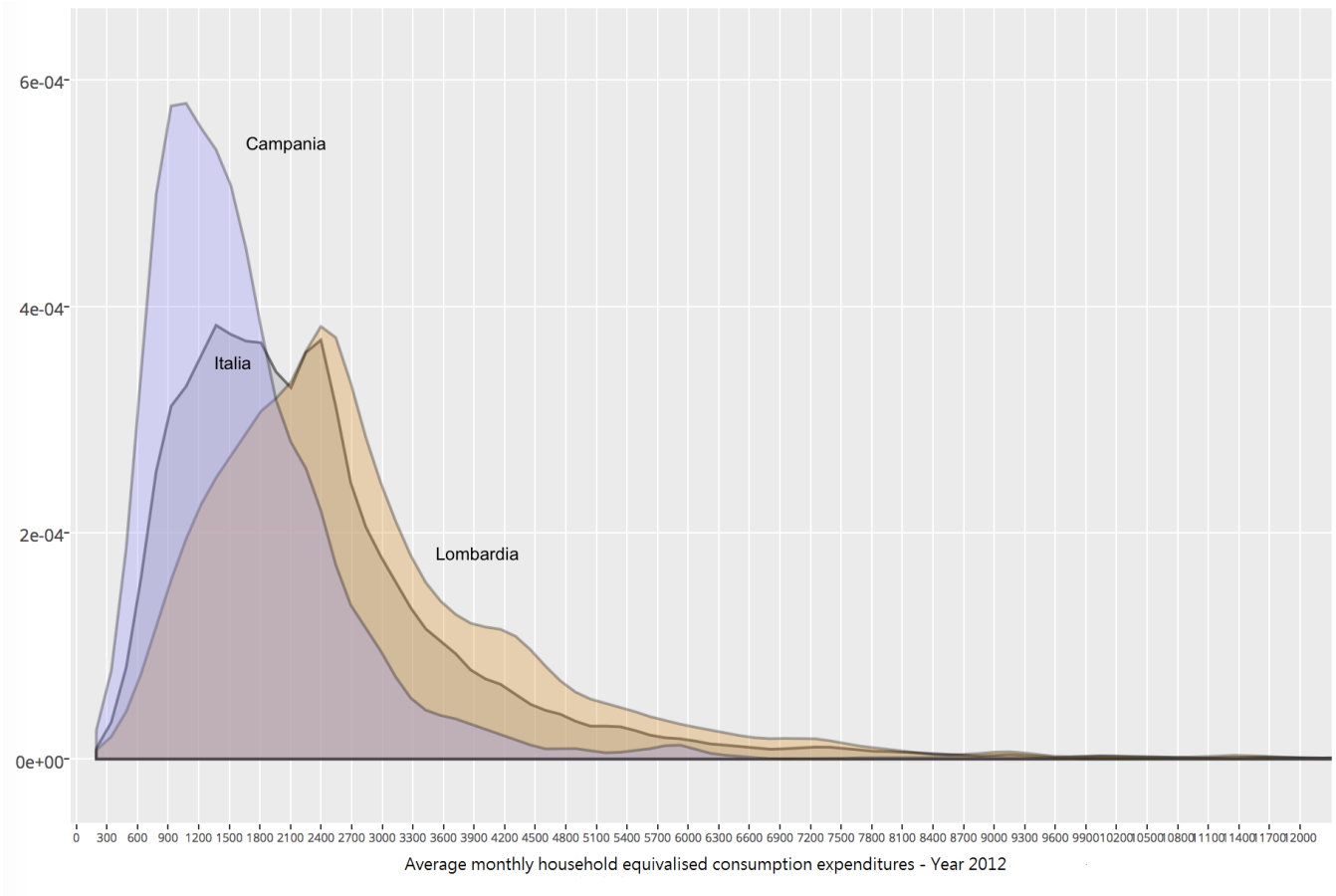

Figure: Distribution of the average monthly household equivalised consumption expenditures in Italy - Year 2012. Source: ISTAT

SAE of Household Consumption Expenditure in Italy

- **User need:** Reliable disaggregated estimates of Household consumption expenditure in order to plan and evaluate the political interventions at provincial level.
- **Target parameters**: HCR (based on the Household Consumption Expenditures) for the 110 provinces in Italy.
- The data on income come from the 2012 wave of Household Budget Surveys (HBS). The estimates are available for large geographic domains (NUTS-1 (national) and NUTS-2 (regional) levels in Italy)
- Auxiliary variables come from the Italian Population Census 2011.

- average taxable income at provincial level
- share of ownership of the house at province level

Area-level data

The Model

- Area-level **Fay Herriot Model (FH)**: model proposed by Fay and Herriot (1979).
- FH Model (linear mixed model):

1 $\hat{\theta}_d^{dir} = \theta_d + e_d$

2 $\theta_d = \mathbf{x}_d^T \boldsymbol{\beta} + u_d$

$$\hat{\theta}_d^{dir} = \mathbf{x}_d^T \boldsymbol{\beta} + u_d + e_d, \tag{1}$$

- θ_d^{dir} parameter of interest;
- $\mathbf{x}_d$ is the vector of p auxiliary variables;
- $\boldsymbol{\beta}$ is the vector of regression coefficient;
- u_d's are area-specific random effects: $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$;
- e_d's are the sampling errors in the d^{th} area: $e_d \overset{iid}{\sim} N(0, \psi_d^2)$;

The Fay-Harriot Model

The **Best Linear Unbiased Predictor (BLUP)** is obtained minimizing $MSE_m(\hat{\theta}_d^{dir})$

- $\min_{\hat{\theta}_d^{dir}} MSE_m(\hat{\theta}_d^{dir}) \rightarrow$

$$\tilde{\theta}_d^{EBLUP} = \mathbf{x}_d^T\tilde{\boldsymbol{\beta}} + \frac{\sigma_u^2}{\sigma_u^2 + \psi_d^2}(\hat{\theta}_d - \mathbf{x}_d^T\tilde{\boldsymbol{\beta}}) = \mathbf{x}_d^T\tilde{\boldsymbol{\beta}} + u_d$$

The BLUP can be rewritten as follows:

$$\tilde{\theta}_d^{BLUP} = \gamma_d\hat{\theta}_d^{dir} + (1-\gamma_d)\mathbf{x}_d^T\tilde{\boldsymbol{\beta}}$$

- $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2+\psi_d^2}$
- ψ_d^2 is assumed known (actually is the estimated MSE of direct estimate)

From the big picture....

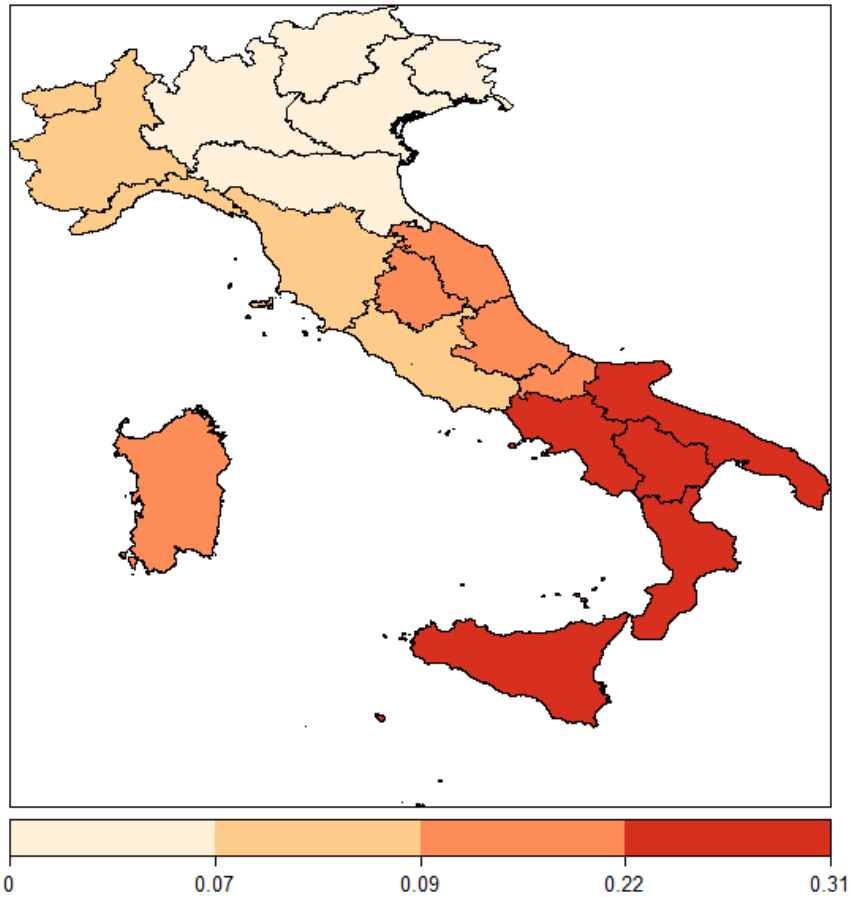

Figure: HCR (consumptions) for the 20 Italian Regions - Household Budget Survey 2012

...to the local level

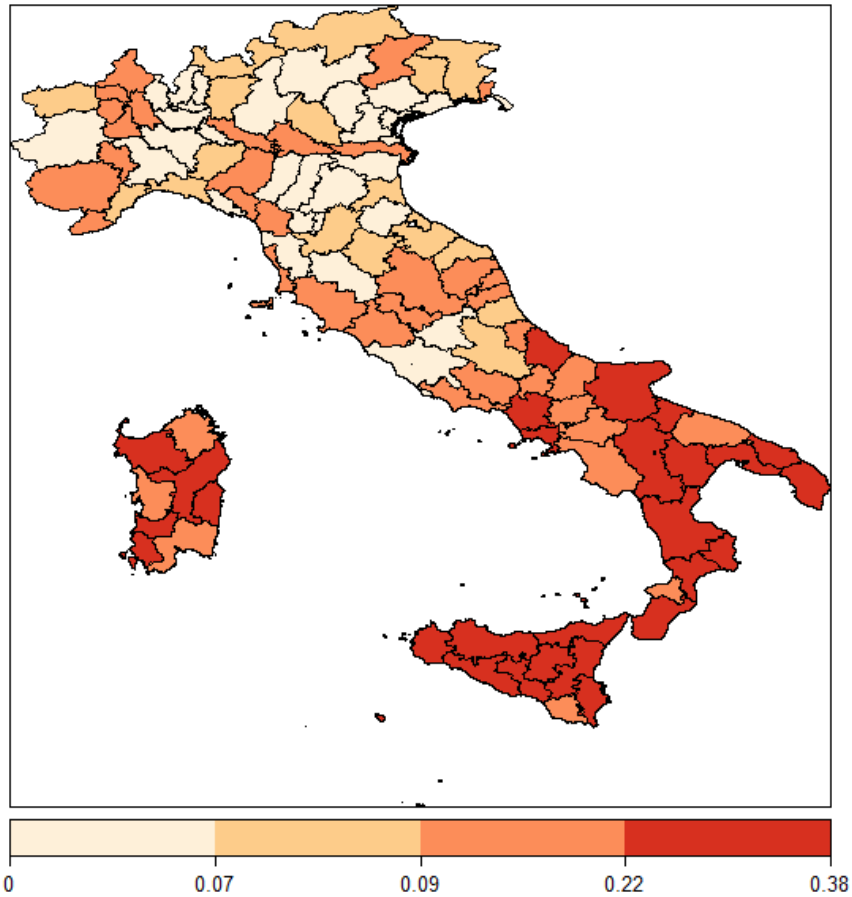

Figure: HCR (consumptions) for the 110 Italian Provinces - Household Budget Survey 2012

Geographical inequality of income in Italy

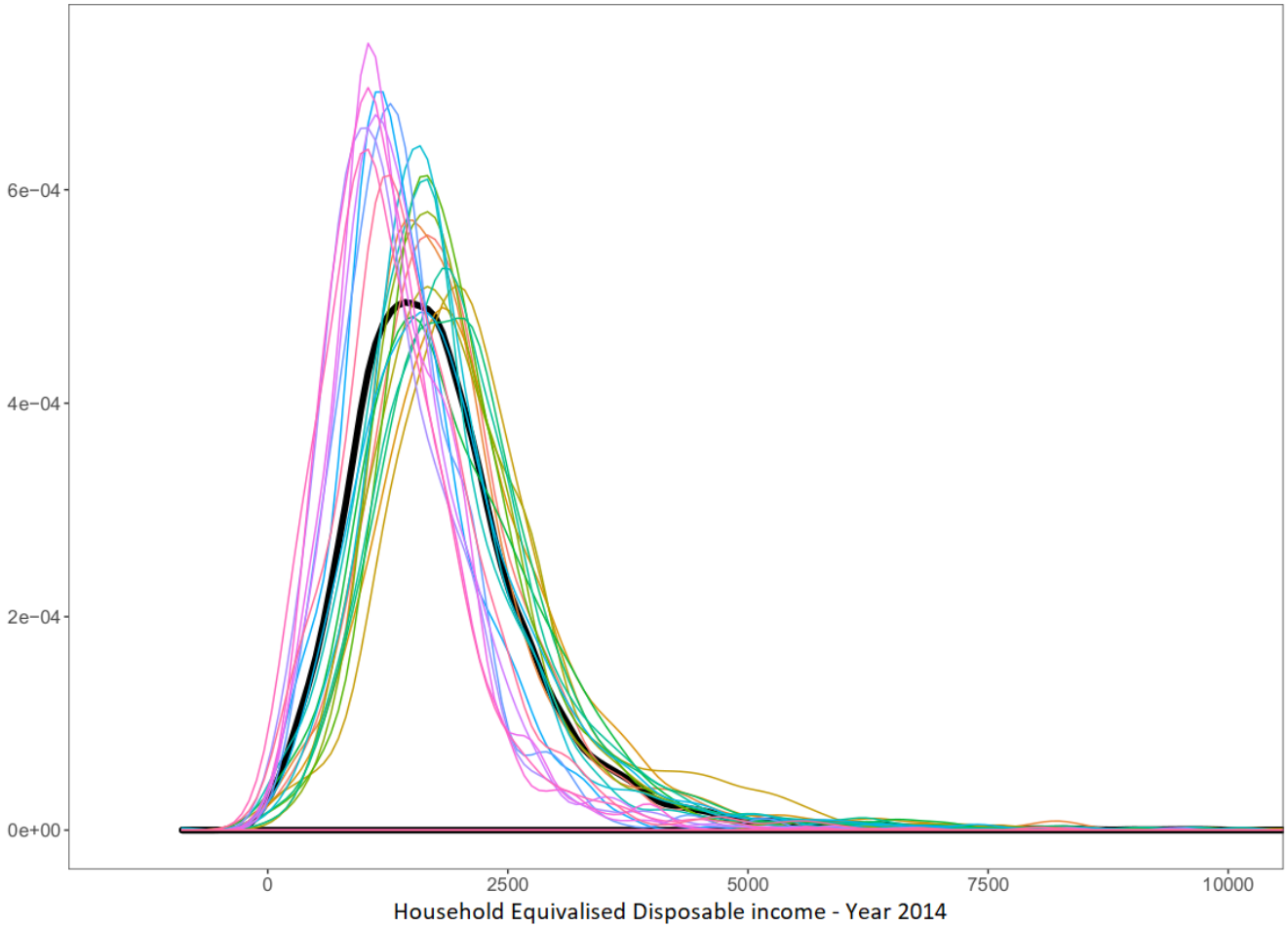

Figure: Distribution of the household equivalised income in Italian Regions - Year 2014. Source: EU-SILC

Small Area Estimates of HCR in Tuscany

- **Target parameters**: HCR (income) for the 57 Local Labour Systems (LLSs) of the Tuscany region, Italy.
- The data on income come from the 2011 wave of Italian EU-SILC survey.
- Auxiliary come form the Italian Population Census 2011.

- proportion of males aged 15–24 with low educational level
- proportion of males aged 25–34 with low educational level
- proportion of non-Italian males aged 25–34
- proportion of unemployed males aged 34–65

Area-level data

The Model

- The spatial dependence among small areas is introduced in the FH model by specifying a linear mixed model with spatially correlated random area effects:

$$\hat{Y}_d = \mathbf{x}_d^T \boldsymbol{\beta} + \mathbf{D}(I_d - \rho \mathbf{W})^{-1}\mathbf{u}, \tag{2}$$

  - $\mathbf{D}$ is a $d \times d$ matrix of known positive constants;
  - $\mathbf{W}$ is a matrix of spatial interaction structure of the small areas;
  - ρ autoregressive coefficient defines the strength of the spatial relationship among the random effects associated with neighbouring areas.

- **R package** `sae` implements small area estimation methods. Other R functions (i.e. M-quantile models) are available at `sample-project.eu` (SAMPLE has been an Europea research programme coordinated by University of Pisa)

Poverty Mapping

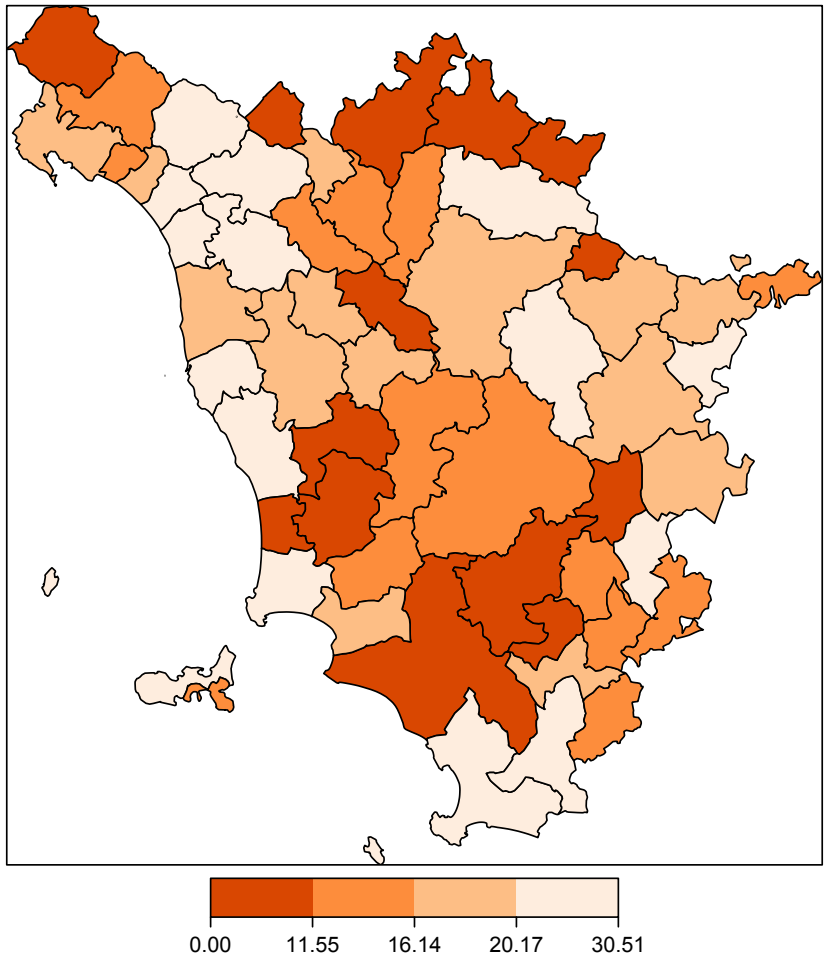

Figure: Estimates of the HCR for LLSs of Tuscany. Source: Giusti et al. (2017)

Confidence Intervals for the HCR

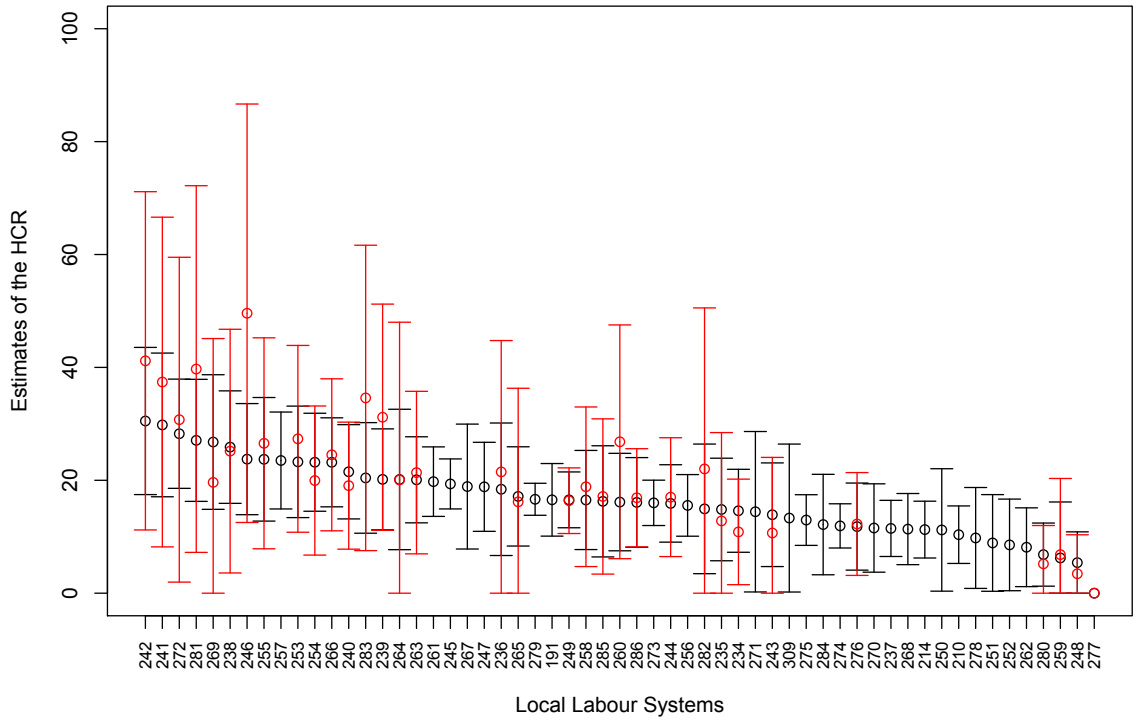

Figure: Confidence Intervals for the HCR of LLSs of the Tuscany region: model-based CIs are represented in black, CIs based on direct estimates are represented in red. Source: Giusti et al. (2017)

China provinces

- Poverty estimates (the target variables may be the income or the consumption expenditures at household level).
- The province(s) where to apply the experiment of SAE could be selected on the basis of the geographical location, the population size, the level of development.
- Example: three provinces different by geography and economy:
  - Sichuan (southwest):
    6th of all 31 provincial regions by GDP in 2017
    2017 GDP per capita (based on mid-year population): 72,851 CNY
    Population: 80,418,200 (2010 Census)
  - Anhui Province (east):
    13th of all 31 provincial regions by GDP in 2017
    2017 GDP per capita (based on mid-year population): 44,206 CNY
    Population: 59,500,510 (2010 Census)
  - Qinghai Province (northwest):
    30th by GDP in 2017
    2017 GDP per capita (based on mid-year population): 44,348 CNY
    Population: 5,626,722 (2010 Census)

SAE in China

- Income and consumption estimates for the Prefectures and/or Counties of the selected provinces
  - 16 prefectures (105 counties) in Anhui Province (east):
  - 21 prefectures (181 Counties) in Sichuan (southwest):
  - 8 prefectures (43 Counties) in Qinghai Province (northwest).
- It is needed to carefully check the availability of surveys (detailed information on the sample design: size of the sample at provincial, prefecture and/or county level; methods of the estimation of the main variables...) and census/administrative data.
- Target indicators in the Households surveys (Urban and Household Survey in China): sampling survey implemented by each provincial survey organizations.
- Auxiliary information in the Population Census (2010): composition of the population; economic activities, family composition, housing information...
- Synergy between Centre Dagum and NBS in order to develop R code to apply SAE methods in China.

Organization and Mission

Tuscan Universities Research Centre Camilo Dagum on Advanced Statistics for the Equitable and Sustainable Development – ASESD

Director: Monica Pratesi, University of Pisa

3 Honorary Fellows:

Achille Lemmi, University of Siena

Luigi Biggeri, University of Florence

Jacques Silber, Economic Department of Bar-Ilan University

54 members from several Italian Universities and National Statistical Institute (ISTAT)

- Promoting applied research on issues that engage the testing of multi – interdisciplinary contributions
- Advanced statistical methods to the study of sustainable and equitable development

SAE and Centre Dagum

- Bertarelli G., Ranalli M.G., Bartoloucci F., D'Alo M., Solari F. (2018). Small Area Estimation for unemployment using latent Markov models. *Surv Methodol*, 44(2)
- Schirripa Spagnolo F., D'Agostino A., Salvati N., Measuring differences in economic standard of living between immigrant communities in Italy, *Qual & Quant*,52(4), 1643–1667
- Giusti C., Masserini L., Pratesi M. (2017). Local Comparisons of Small Area Estimates of Poverty: An Application Within the Tuscany Region in Italy. *Social Indic Res*, 131(1), 235-254.
- Pratesi M. (2016) (ed). *Analysis of Poverty Data by Small Area Estimation*. New York, Wiley.
- Pratesi M. (2015), *Spatial Disaggregation and Small Area Estimation Methods for Agricultural Surveys: Solutions and Perspectives*. FAO Technical Report in the Global Strategy Publications.
- Tzavidis, N., Marchetti, S., Chambers, R. (2010). Robust estimation of small-area means and quantiles. *Aust NZ J Stat*, 52(2), 167-186.
- Pratesi, M., Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat Method Appl*, 17(1), 113-141.

THANK YOU!

Monica Pratesi

monica.pratesi@unipi.it