Poverty Estimation using Small Area Methods

Nikos Tzavidis [1]
Joint work with Timo Schmid, Natalia Rojas & Bea Weidenhammer (Freie Universität Berlin) & Nicola Salvati (University of Pisa)

Session 9
CESS Conference
Budapest, October 21, 2016

[1] Southampton Statistical Sciences Research Institute, University of Southampton (n.tzavidis@soton.ac.uk)

What is Poverty Mapping?

Definition

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.

Examples

- Estimate income distribution at domain level
- Estimate poverty and inequality indicators

Estimation of Complex Indicators

- Growing needs of statistics agencies for estimates at very fine spatial scales
- Model-based methods have dominated recent literature
- Until recently less attention to robustness issues

Examples of Complex Indicators

Income-based indicators

- FGT measures (Foster et al.,1984)

$$FGT(\alpha, t) = \sum_{i=1}^{N} \left(\frac{t - y_i}{t} \right)^{\alpha} \mathbb{1}(y_i \leq t)$$

$\alpha = 0$ - Head Count Ratio; $\alpha = 1$ - Poverty Gap

- The Gini coefficient

$$Gini = \frac{N+1}{N} - \frac{2 \sum_{i=1}^{N} (N+1-i) y_{(i)}}{N \sum_{i=1}^{N} y_{(i)}}$$

- Quintile Share Ratio

$$QSR_{80/20} = \frac{\sum_{i=1}^{N} [y_i \mathbb{1}(y_i > q_{0.8})]}{\sum_{i=1}^{N} [y_i \mathbb{1}(y_i \leq q_{0.2})]}$$

SAE - Data Sources / Requirements

- **Survey Data:** Available for y and for x related to y
- **Census/Administrative Data:** Available for x but not for y

- Access to good auxiliary information is crucial
- Methods require auxiliary information available for every unit in the population - Census/admin micro-data
- Data Hungry Methods: Implementation of currently used methods require access to sensitive data

Model-based Methods - Nested Error Regression Model

Battese, Harter & Fuller, 1988, JASA

Include random area-specific effects to account for between area variation

Notation: (k =domain, i =individual)

$$y_{ik} = \boldsymbol{x}_{ik}^T \boldsymbol{\beta} + \boldsymbol{z}_{ik}^T u_k + \epsilon_{ik}, i = 1, ..., n_k, k = 1, ..., D,$$

$$u_k \sim N(0, \sigma_u), \epsilon_{ik} \sim N(0, \sigma_\epsilon)$$

Some Recent Methodologies

- The World Bank method
  (Elbers et al., 2003, Econometrica)
- The Empirical Best Predictor (EBP) method
  (Molina & Rao, 2010, CJS)
- EBP based on normal mixtures (Elbers & Van der Weidel, 2014; Lahiri and Gershunskaya, 2011)
- Methods based on M-Quantiles
  (Marchetti et al., 2012, CSDA)
- Semi-parametric estimation of the empirical distribution function
  (Tzavidis et al., 2016)

The EBP Method (under normality)

$$\hat{z}_k = N_k^{-1}\Big[\sum_{i\in s_k} z_i + \sum_{i\in r_k} \hat{z}_i^{EBP}\Big]$$

- Estimation uses a unit-level mixed effects model

Summary of the Method

- $\hat{z}_k^{EBP}$ estimated by using the predictive density $f(y_r|y_s)$
- Use sample data to estimate β, σ_u^2, σ_ϵ^2, γ_k
- Generate $u_k^* \sim\ N(0, \hat{\sigma}_u^2(1-\gamma_k))$ and $\epsilon_{ik}^* \sim\ N(0, \hat{\sigma}_\epsilon^2)$

$$y_{ik}^* = \mathbf{x}_{ik}^T\hat{\boldsymbol{\beta}} + \hat{u}_k + u_k^* + \epsilon_{ik}^*$$

- Micro-simulation of a synthetic population of y_{ik}^*.
- Calculate the indicator of interest using the y_{ik}^*.
- Repeat the process L times and average the estimates.
- MSE estimation: Parametric bootstrap

Motivating Alternative Methods

- EBP relies on assumptions about the distribution of the data
- What if these fail?

- **Alternative I:** Explore the use of transformations. Deciding on appropriate transformations is not straightforward, but offers a possible avenue for improving the model
- **Alternative II:** Use robust methods as an alternative to transformations
- **Alternative III:** Modify the parametric assumptions of EBP. Possible only for some distributions

A Robust Alternative - Microsimulation via Quantiles (MvQ) method (Tzavidis et al., 2016)

- Estimate the empirical distribution function (edf)
- Use the edf to generate synthetic populations as in the EBP
- Use each generated population for small area estimation

- $Q_{y|\boldsymbol{x},k}(q|\boldsymbol{x},k)$ denote the quantile function of an unknown $F(y|\boldsymbol{x},k)$
- Interested in estimating this quantile function
- **Simplest case:** Assume a linear model for the quantiles

$$Q_{y|\boldsymbol{x},k}(q|\boldsymbol{x},k) = \boldsymbol{x}_{ik}^T\boldsymbol{\beta}_q + v_k$$

- v_k domain random effect capturing unobserved heterogeneity

Mixed Effects Quantile Regression

- $p(y, v|\boldsymbol{\theta}) = p(y|v, \theta_1)p(v|\theta_2)$
- Use the link between quantile regression and MLE under the Asymmetric Laplace distribution (Yu & Moyeed, 2001, Stat. & Probab. Lett.)
- $p(y|v, \theta_1) \sim ALD(\mu, \sigma, q)$
- with $\mu = \boldsymbol{x}^T \boldsymbol{\beta}_q + v$

- $p(v|\theta_2)$
- Normal (Geraci & Bottai, Stats & Comp, 2013)
- Discrete mixture (Marino, Tzavidis & Schmid, 2016)

Design-based simulation - Setup

Data

- Census data from one state in Mexico
- Outcome is the earned per capita income from work
- Target parameters include the Gini coefficient & median income
- Target areas: Municipalities in the state

Setup

- Design-based simulation with 500 MC-replications from fixed population
- 6 covariates leading to a R^2 of around $40 - 50\%$
- Unbalanced design leading to a sample size of $n = 2195$ ($min = 8$, $mean = 17.6$, $max = 50$)

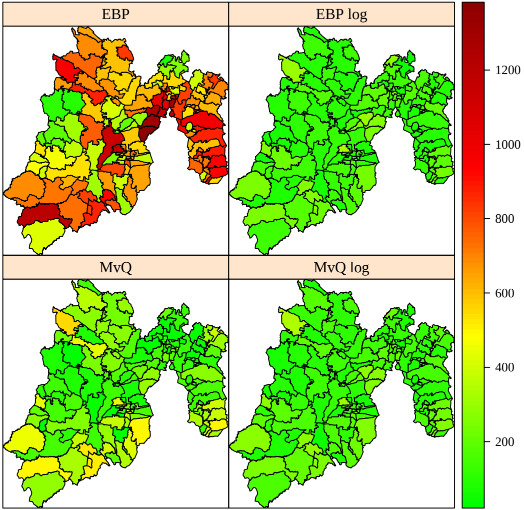
Design-based simulation - Methods

1. EBP - Model: 2-level nested error regression model (households nested within municipalities) with and without log transform for income
2. MvQ - Model: 2-level nested error regression model for the quantiles of income with and without log transform for income

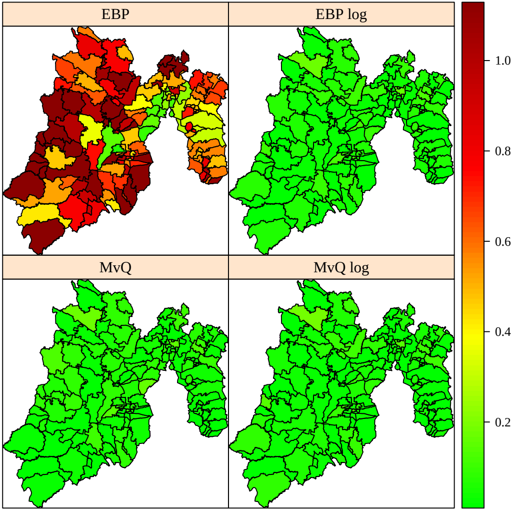
Aims:

- Assess robustness of MvQ when log transform is not used
- Compare the MvQ and EBP methodology

RMSE - Median



RMSE - Gini



Unresolved Challenges I

- Transformations and robust methods can help. However,
- Small departures from the assumed model assumptions will impact upon estimation
- Impact depends on the target of estimation
- E.g. Gini coefficient possibly more difficult to estimate than median income
- MSE estimation that relies on parametric bootstrap can be a risky strategy
- External validation of model-based estimates becomes very important

Unresolved Challenges II

- Currently the biggest challenge with poverty mapping methodologies is access to Census micro-data
- **Possible solution**: Replace Census by a bigger survey that covers all areas/domains
- Adapt methodologies to include measurement error in the covariates coming from the bigger survey
- However, are the estimates of acceptable precision?