The effects of social transfers on the *At-Risk-Of-Poverty* rate

A local comparison with two applications of the Fay-Herriot model

Saverio Pertosa, Linda Porciani, Monica Pratesi

Pisa, 09/05/2018

Summary

1 Background and motivation

2 Fay-Harriot model | EBLUP

3 Data

4 Results - Direct

5 Results - FH

6 Future research

7 Appendix : Power-laws

8 Supporting slides

Motivation

- The interest is in the area income distribution and applying small area estimation techniques to real world data sets and problems that may then help inform policy decisions.
- Traditional surveys allow for adequate precision only at regional level (NUTS2).

- Research question : What is the effect of social transfers on provincial level ?
- How : Increase available knowledge on Italian households' living conditions by *Small area estimations* based on province-level data :
  - test the use of several and innovative datasets as sources of information
  - re-define poverty in terms of regional income distribution

What is small area estimation ?

- Small area estimation is a statistical technique used for estimating parameters for small sub-populations, when the sub-population of interest is included in a larger survey
- An area is regarded as “small” if the sample from the area is not sufficient to produce direct estimates of adequate precision.
- Small area estimation “borrows strength” auxiliary information related to the variable of interest. The modeling approach is quite powerful however the results it yields are highly dependent on the validity of the model (Longford, 2005).

Small Area Estimation : FH

Fay-Herriot's approach employs the parameter's direct estimates and the totals/averages of covariates' values on small areas.

$$y_i = x_i^T \beta + v_i + e_i, \quad i = 1, \ldots, m$$

- x_i is a vector of known *covariates*
- β is a vector of unknown regressors' coefficients
- v_i is an area-specific random effect (uncorrelated with independent variable)
- e_i is a sampling error

Small Area Estimation : FH (2)

It emerges by combining the two model assumptions :

- The direct estimate is available and design-consistent (EU-SILC provides survey weights)

$$\mathbf{y}^{DIR} = \mathbf{y} + \mathbf{e}$$

  with $\mathbf{e} \sim N_m(0, \mathbf{D})$. $\mathbf{D}$ is a diagonal matrix of heteroskedastic elements D_i, assumed known by hypothesis

- A matrix of auxiliary variables is linearly related to $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{v}$$

  with homoskedastic $\mathbf{v} \sim N_m(0, \psi\mathbf{I})$

Under the assumption of normality of errors, given $\hat{\beta}$ and $\hat{\psi}$, the EBLUP is given by

$$\hat{y}_i^{FH} = \gamma_i y_i^{DIR} + (1 - \gamma_i) x_i^T \beta$$

where $\gamma_i = \frac{\psi}{\psi + D_i}$ is called *shrinkage factor*.
EBLUP in this setting is a convex combination between the direct and synthetic estimates.

Parameters' estimation and MSE

- $\hat{\psi}$ can be estimated in four ways. The most common method is **REML**
- $\hat{\beta}$ is a function of $\hat{\psi}$, so there are four ways to *fit* the model

$$\hat{\beta}(\psi) = \left(\sum_{i=1}^{m} \frac{x_i x_i^T}{\hat{\psi} + D_i} \right)^{-1} \left(\sum_{i=1}^{m} \frac{x_i y_i}{\hat{\psi} + D_i} \right)$$

Regardless of the method to fit the model, we have the following MSE, function of ψ

$$MSE(\hat{y}_i^{FH}) = g_{1,i} + g_{2,i} + g_{3,i}$$

approximated correctly (through $\hat{\psi}$) to

$$\hat{MSE}(\hat{y}_i^{FH}) = g_{1,i} + g_{2,i} + 2g_{3,i}$$

- $g_{1,i} = \frac{\psi D_i}{\psi + D_i} = O(1)$, due to random errors (leading term)
- $g_{2,i} = O(m^{-1})$, due to β's estimation
- $g_{3,i}$, due to the estimate of ψ

Estimates are improved (in terms of MSE or CV) thanks to auxiliary information.

Data : Target indicator

- EU-SILC (European Union Statistics on Income and Living Conditions) 2013
  - At-risk-of-poverty rate (HCR) before social transfers (national poverty line)
  - At-risk-of-poverty rate (HCR) after social transfers (national poverty line)
  - At-risk-of-poverty rate (HCR) before social transfers (regional poverty lines)
  - At-risk-of-poverty rate (HCR) after social transfers (regional poverty lines)

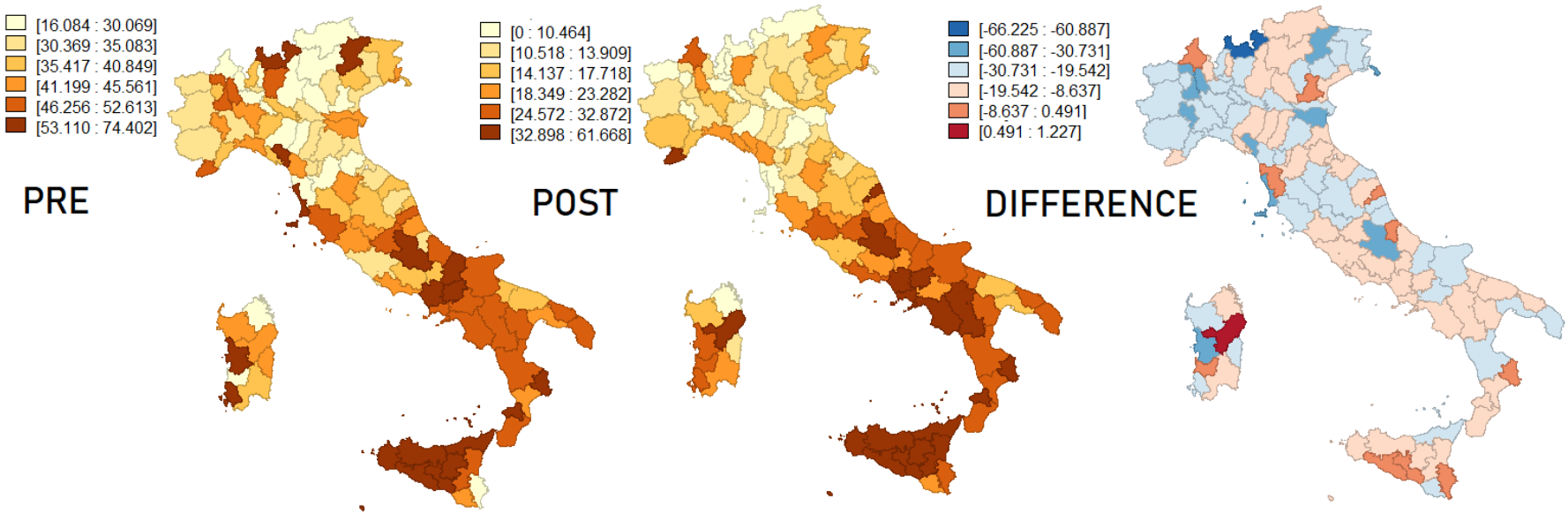
The poverty line is defined as 60% of the median of household *equivalised income*

Auxiliary information

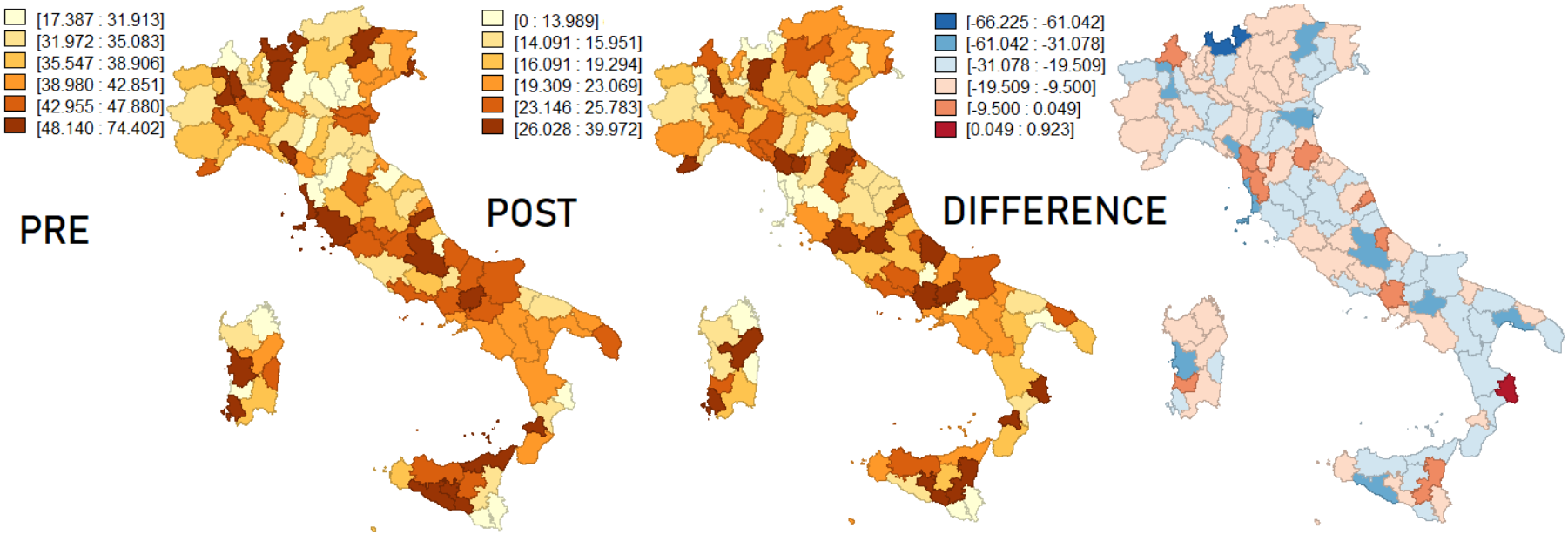
Covariates

- Registry offices (Liste anagrafiche)
  - *Number of families*
  - *Average dimension of families*
  - *Number of domestic partnerships*
- 8000Census
  - *Theft rate*
  - *Net rate of new enrollments in the business registers*
  - *Ability to export (sectors with dynamic global demand)*
  - *Incidence of adults with higher education qualifications*
  - *Unemployment rate*
  - *Incidence of NEETs*
- Tax declaration (Dichiarazione dei redditi)
  - *Mean per-capita income at provincial level*
- Social capital
  - *Social capital* as measured by Guiso, Zingales, Sapienza

Case study : AROP - Direct estimates - National line

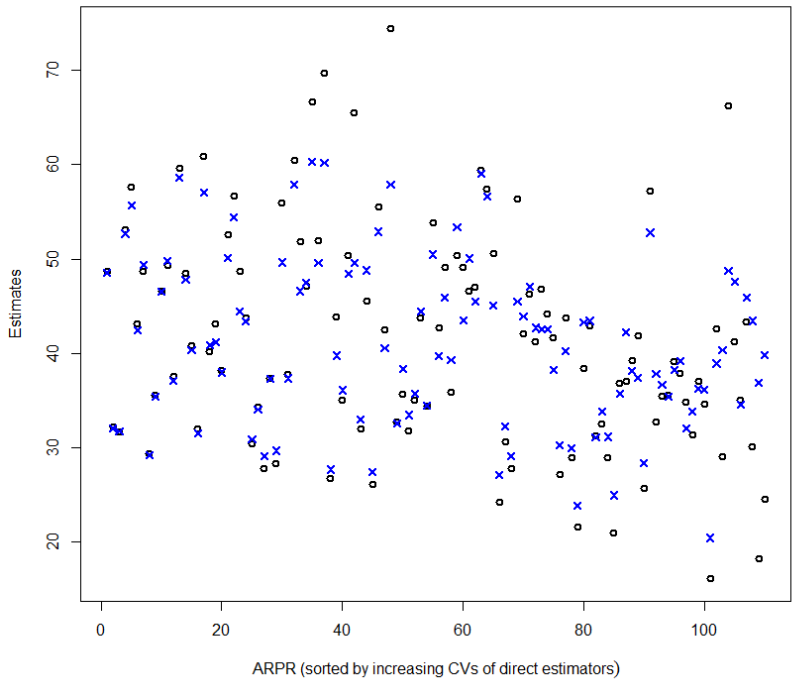


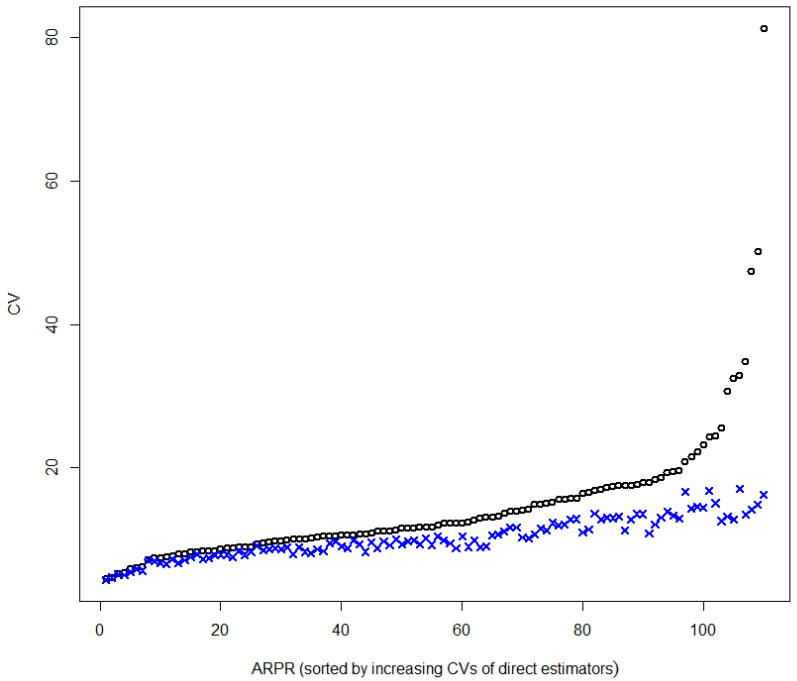
Case study : AROP - Direct estimates - Regional lines



AROP - FH - National line - PRE

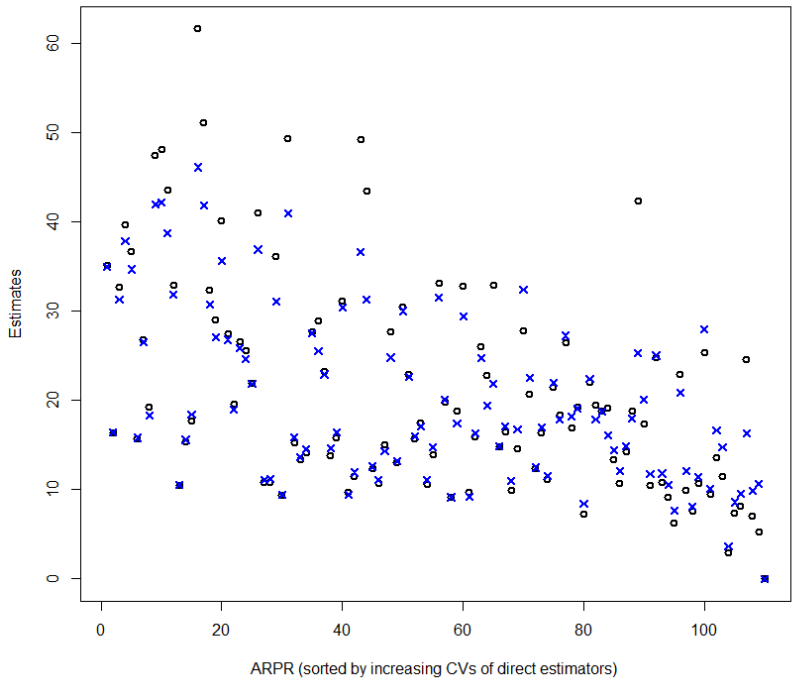
Covariates : • Social capital • Net Income (Tax declaration) • Number of families • Number of domestic partnerships • Theft rate

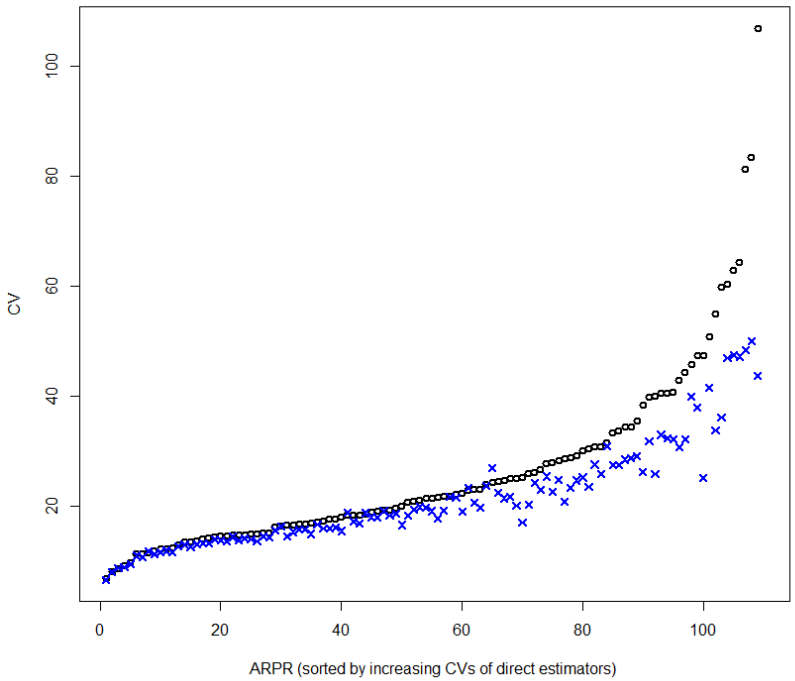




AROP - FH - National line - POST

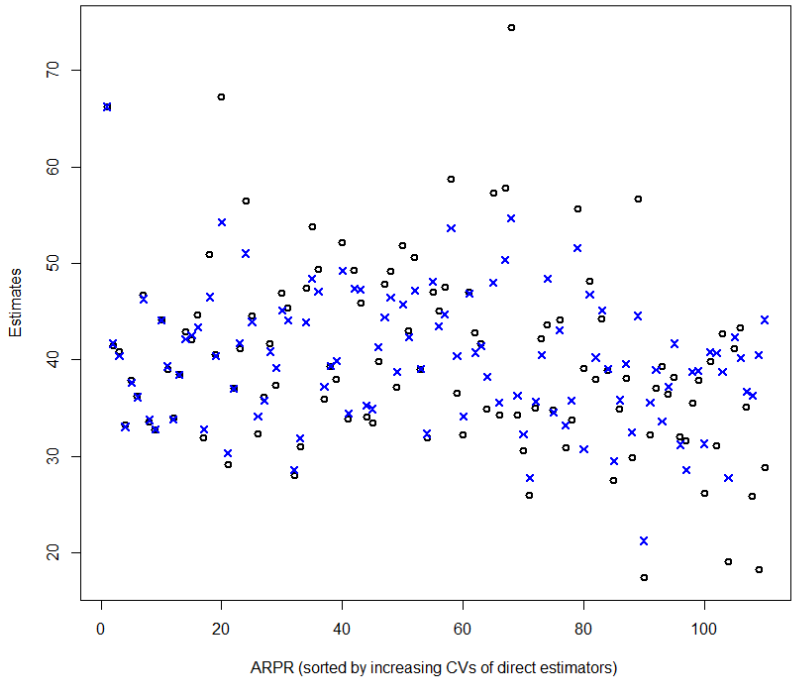
Covariates : • Net income (Tax declaration) • Number of families • Number of domestic partnerships • Average dimension of families • Incidence of adults with a higher educational qualification

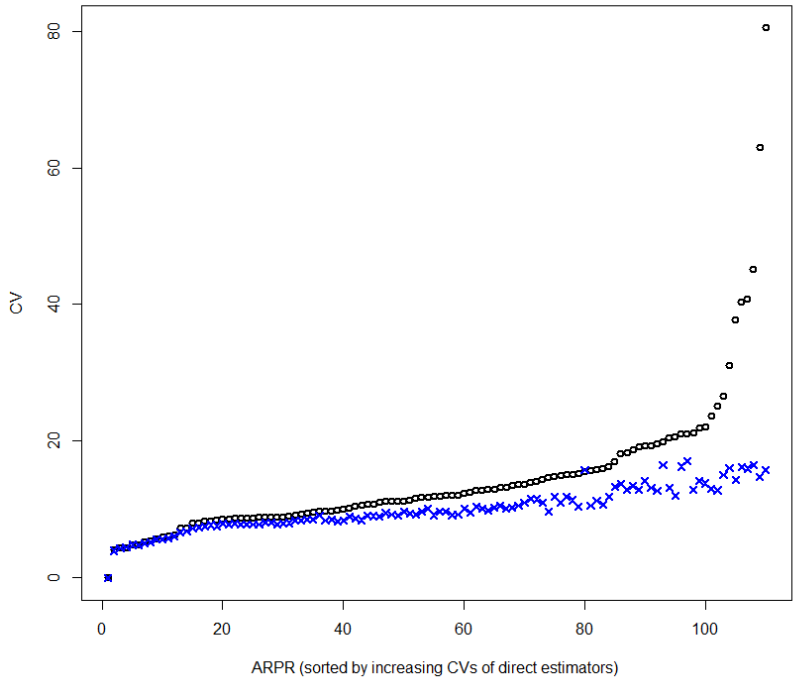




AROP - FH - Regional lines - PRE

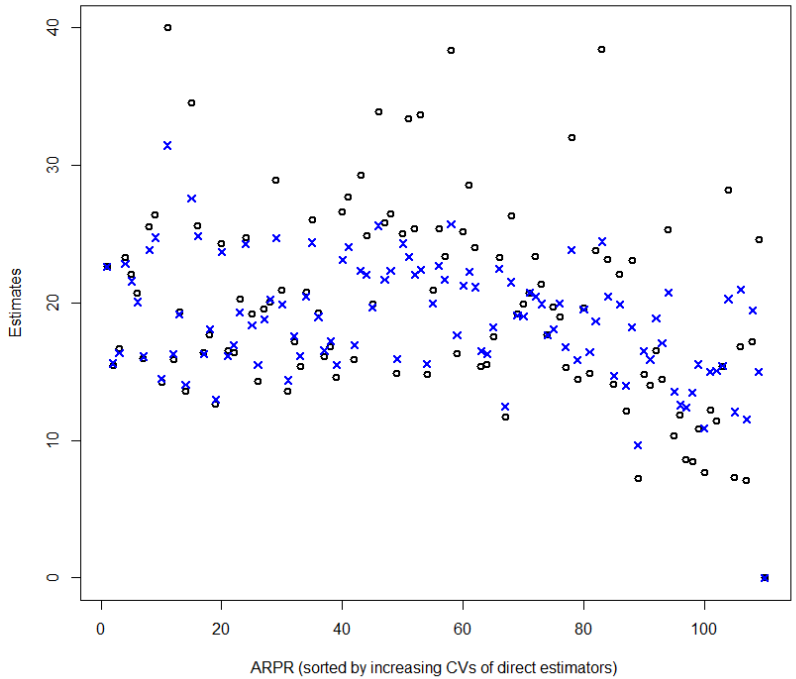
Covariates : • Social capital • Net income • Number of families • Average dimension of families • Theft rate • Net rate of new enrollments in the business registers • Ability to export (sectors with dynamic global demand) • Incidence of adults with higher education qualifications • Unemployment rate • Incidence of NEETs

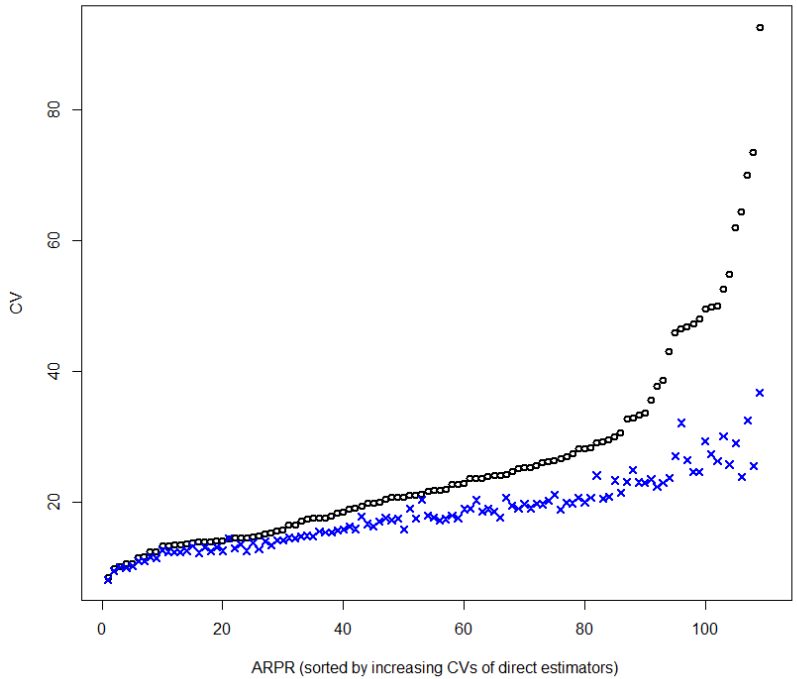




AROP - FH - Regional lines - POST

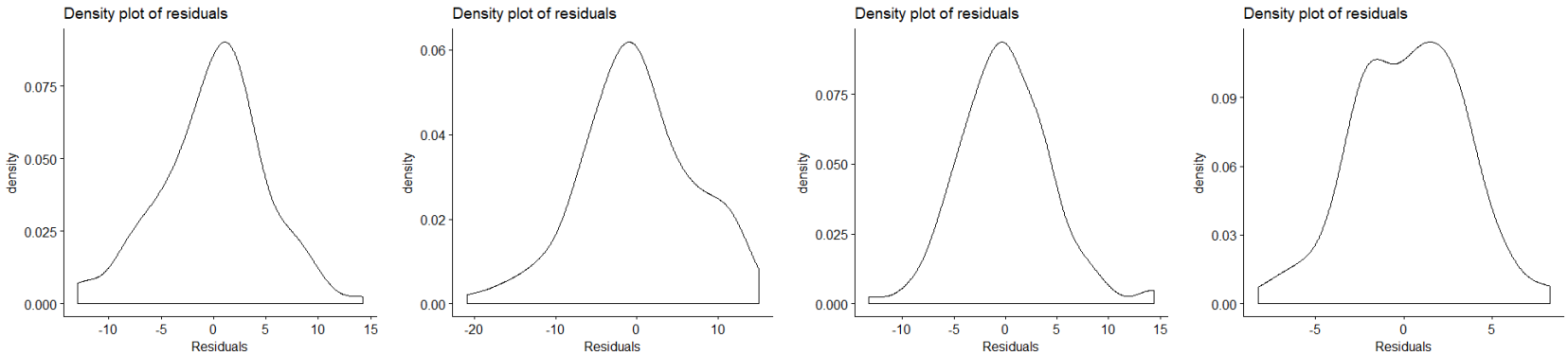
Covariates : • Social capital • Net income (Tax declaration) • Number of families • Average dimension of families • Theft rate • Ability to export (sectors with dynamic global demand) • Incidence of adults with higher education qualifications • Unemployment rate





Hypothesis testing

While the assumption of normality for **e** cannot be tested for lack of data-point, it is crucial to assess it on **v** to evaluate the quality of model's assumptions. The optimality properties of the predictor hugely depend on the extent on which the underlying hypotheses hold.



When $\hat{\psi}$ is low, comparatively the effect of D_i on γ_i is higher, which in turns give more weight to the synthetic estimator rather than the direct one. γ_i also comes into play in reducing g_1 that is - as we have seen - the leading term in defining the magnitude of the MSE. All the following $\hat{\psi}$ have a 0.001-confidence level : as apparent, the existence of these effects cannot be rejected with an extremely high level of confidence

(a) $\hat{\psi}$ = 43.84 (b) $\hat{\psi}$ = 67.7 (c) $\hat{\psi}$ = 37.46 (d) $\hat{\psi}$ = 26.59

Concluding remarks

- Smoother distribution (*overshrinkage effect*)
- Good auxiliary information related to the variables of interest plays a vital role. Coordination and cooperation among agencies and research departments could drastically improve how we measure the quality of policies.
- Area-level models can be more easily implemented because area-level data are more readily available. However, the assumption of known sampling variance is restrictive.

- The FH model is appropriate to get the estimates of interest.
  Through SAE one can study the effect of social transfers on areas on which adequate precision of estimates can not be obtained by traditional survey methods.
- A more in-depth knowledge of local distribution of social protection is a crucial measure for metric-based policymaking.

Future research

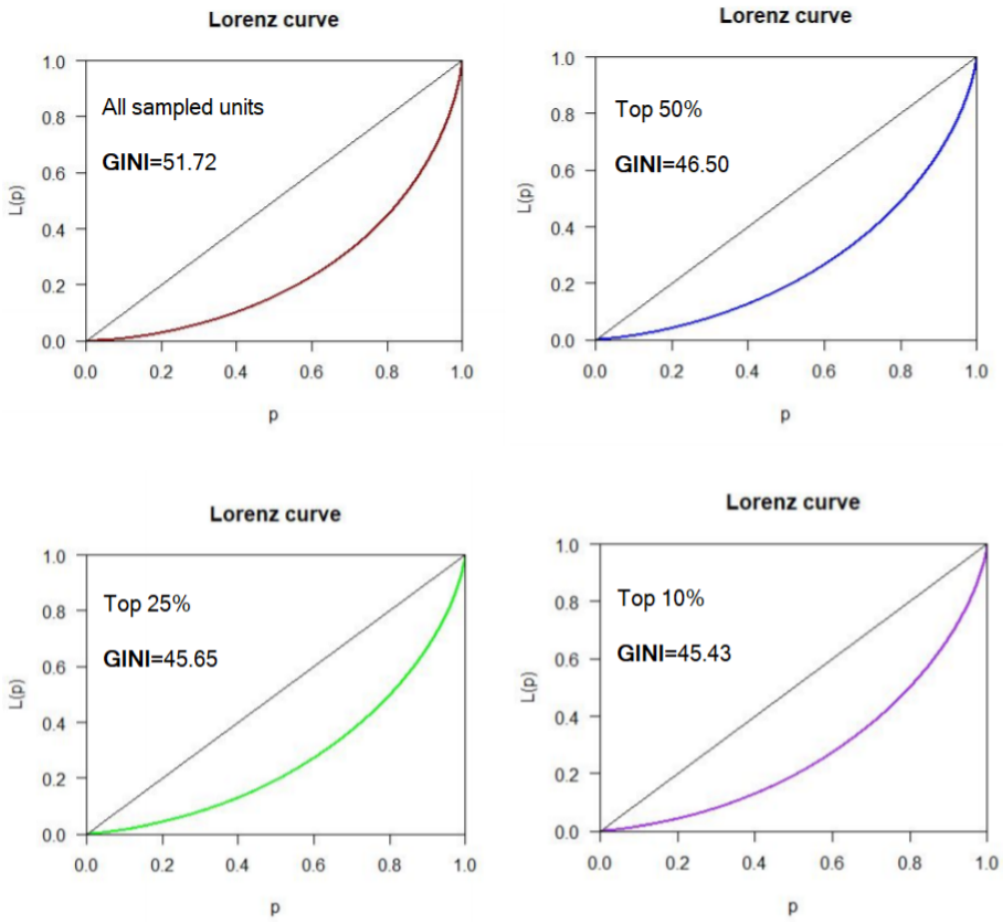
- Transform the proportion of relatively poor households (HCR) through a logit transformation $\log \frac{p}{1-p}$ to have an unbounded variable and thus avoid non-sensical predictions
- Although the proposed approach is generalizable and produced good results, an extension to include spatial and non-linear effects could give new insights
- Covariates based on estimates insert a downward bias in the computation of MSE. Some works in literature, due to lack of information, have assumed that these biases have a negligible effect. A more accurate analysis would require taking into account the sampling design of the covariates (Ybarra and Lohr, 2008)

Thanks

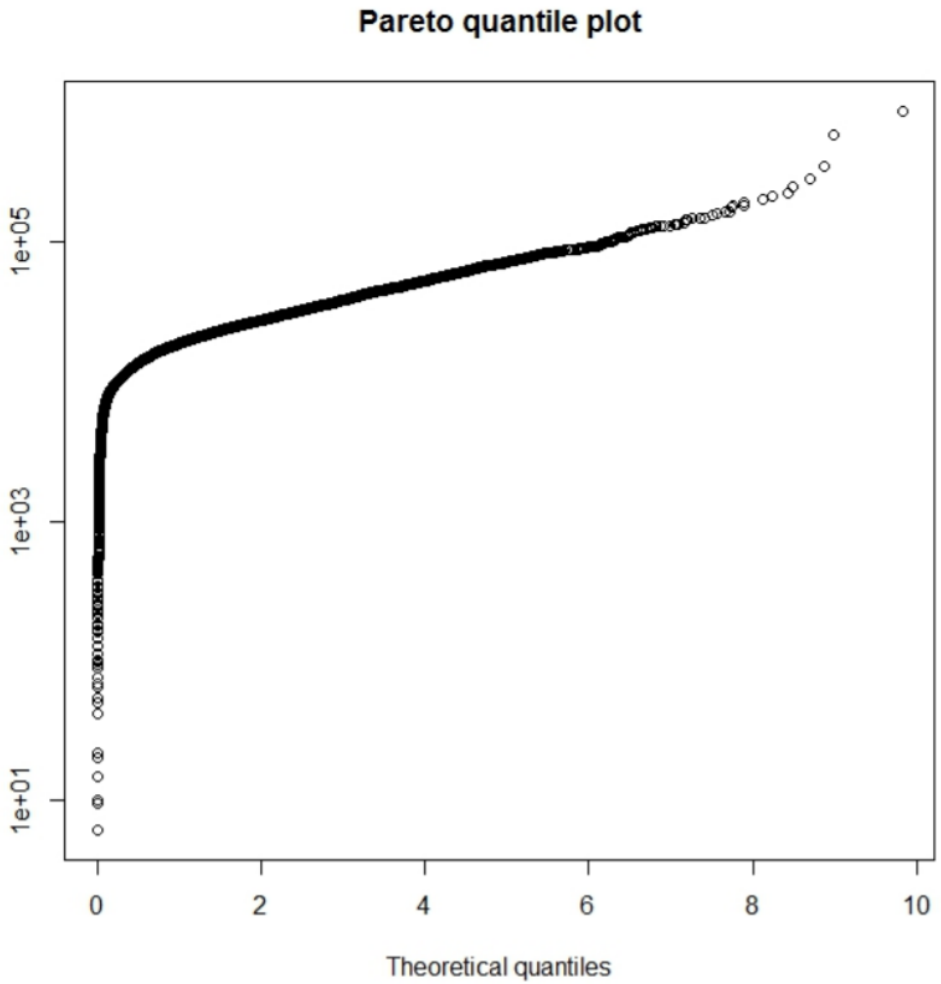
Thanks for your attention

Power-laws and fractality property

Studies on wealth inequality represent a growing field of economics and statistics, that is exposed to fat-tailed data generating processes.



Pareto quantile plot



The estimated tail parameter is 3.31 (upward bias)

Pre-asymptotic measure of Paretianity

- Conventional measures :
  - Tail-exponent
  - Kurtosis
- This new method is based on the rate of convergence of the LLN for finite sums.

Let $X_1, \ldots, X_n$ be i.i.d. RVs with $\mathbb{E}(X) < +\infty$. Let $S_n = S_1 + \cdots + S_n$ be the partial sum of n observations. Define $\mathbb{M}(n) = \mathbb{E}(|S_n - \mathbb{E}(S_n)|)$ as the mean absolute deviation from the mean of n summands. The metric k is defined as the rate of convergence for n additional summands, starting from n_0

$$k_{n_0,n} = min\{\frac{\mathbb{M}(n)}{\mathbb{M}(n_0)}; n = 1, \ldots, n_0\}$$

How many more observations we need to get the same drop in variance from averaging (CLT) as a Gaussian with n_g available data-points ?

$$\hat{n} \approx n_g^{-\frac{1}{k_1 - 1}}$$

With $k_1 = 0.445$ and $n_g = 406$, $\hat{n} = 50124$

Appendix's conclusion

Given the relatively small sample size considered, to "robustify" the small-area approaches when dealing with economic variables it is crucial to have a good survey design stage, able not only to capture outliers - that are essential to have a realistic picture of the wealth distribution - but also to appropriately "weight" every observation. As we have seen, this is particularly hard in the presence of strong skewness and fat-tails. Those are measures of how fast (or rather, slowly) the central limit theorem applies data under consideration

Total Household income after social transfers

Household income components to be added :

1. HY040N : Income from rental of a property or land,
2. HY050N : Family/children related allowances,
3. HY070N : Housing allowances,
4. HY080N : Regular inter-household cash transfer received,
5. HY090N : Interest dividends profit from capital investments in unincorporated business,
6. HY110N : Income received by people aged under 16.

Household income components to be subtracted :

1. HY130N : Regular inter-household cash transfer paid,
2. HY145N : Repayments/receipts for tax adjustment.

Personal income components to be added :

1. PY010N : Net employee cash or near cash income,
2. PY020N : Non-cash employee income,
3. PY050N : Net cash benefits or losses from self-employment,
4. PY080N : Pensions received from individual private plans,
5. PY090N : Unemployment benefits,
6. PY100N : Old-age benefits,
7. PY110N : Survivor's benefits,
8. PY130N : Sickness benefits,
9. PY140N : Education-related allowances.

Equivalized household income

The equivalized disposable income is defined as the household income parametrized by the equivalized household size. Such size is defined by an OECD's scale assigning the following weights :

$$\begin{cases} 1.0 & \text{for the first adult} \\ 0.5 & \text{to other household members aged} \geq 14 \\ 0.3 & \text{to other household members aged} < 14 \end{cases}$$