A Bayesian approach to regional income inequality in Europe using EU-SILC data

Mathias Silva & Michel Lubrano

Aix-Marseille University, CNRS, EHESS, Centrale Marseille, AMSE, France.
8th European User Conference for EU-Microdata

16/03/2023

Hot topic: 'missing rich'

Recent focus on data issues around high incomes (e.g., Bourguignon (2018), Lustig (2020), Flachaire et al. (2022), Blanchet et al. (2022)).

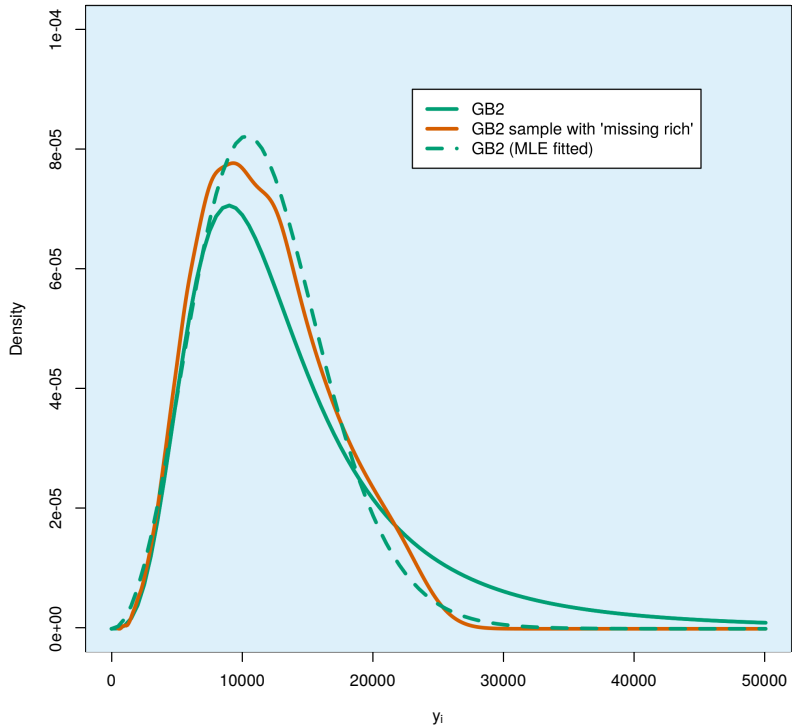
- **Sampling issues:** Non-coverage and sparseness $\Rightarrow$ bias and high s.e.
- **Non-sampling issues:** unit/item non-response, under-reporting, top-coding $\Rightarrow$ bias if correlated with incomes.

Phenomena also affecting EU-SILC data:

- Evidence of 'missing rich' from comparing National Accounts and/or administrative tax data against SILC (e.g., Angel et al. (2019), Ederer et al. (2022), Carranza et al. (2021)).
- Inequality under-estimated when computed on (weighted) sample EU-SILC (e.g., average bias of 2.6-3.38 points for Gini (Carranza et al. (2021), Hlasny and Verme (2018))) .
- Very heterogeneous 'missing rich' patterns across country-years and across register- vs survey-based sources in EU-SILC.

Motivation

Take a typical income distribution model to the typical data disregarding sampling/measurement issues and you've got a problem..



- With issues on the upper tail, we underestimate top income shares, inequality measures (e.g., Gini), mean income (i.e., growth), and overestimate lower income shares.

Coming up...

Some issues with what's been previously done in the literature:

- Several 'corrections' explored, each assuming different forms of 'missing rich' to impute or reweight observations in the data. No unified framework to contrast them, no way of computing s.e.'s in particular.
- Using external data not always feasible and often manipulates income and population analysed.

$\Rightarrow$ **Contribution:** A unified parametric framework integrating replacing and/or non-response mechanisms.

A general parametric framework for 'missing rich' data

Consider the following parametric framework:

$$y_i = \begin{cases} m^{-1}(y_i^{Obs}, \mathbf{X}_i; \eta) \ , \text{ with probability } \varphi(y_i, \mathbf{X}_i; \nu) \\ y_i^{NObs} \ , \text{ with probability } 1 - \varphi(y_i, \mathbf{X}_i; \nu) \end{cases}$$

- $m(\mathbf{y}, \mathbf{X}; \eta) \equiv m(y_i, \mathbf{X}_i; \eta)$ is an **income reporting function**, parametrized by the vector η. States what the observed income of unit i is given her true income and potentially other characteristics $\mathbf{X}_i$. Its inverse $m^{-1}(y_i^{Obs}, \mathbf{X}_i; \eta)$ is the real interest: a **replacing function**.
- $\varphi(\mathbf{y}, \mathbf{X}; \eta) \equiv \varphi(y_i, \mathbf{X}_i; \nu)$ is a **response probability function**, parametrized by the vector ν. Given i's true income and relevant characteristics $\mathbf{X}$, with what probability will she report an income in the data conditional on being sampled?

A general parametric framework for 'missing rich' data

We can expand the population income distribution model $f_{\mathbf{y}}(.;\theta)$ to that of observed incomes $f_{\mathbf{y}^{Obs}}(.;\theta,\eta,\nu)$ under assumed forms for $m(.;\eta)$ and $\varphi(.;\nu)$ as[1]

$$f_{\mathbf{y}^{Obs}}(y_i^{Obs};\theta,\eta,\nu) = \frac{\overbrace{f_{\mathbf{y}}(m^{-1}(y_i^{Obs};\eta);\theta) \times \left(\frac{\partial m^{-1}(y_i^{Obs};\eta)}{\partial y_i^{Obs}}\right)}^{\text{Reporting function: Replacing transformation of } \mathbf{y}} \overbrace{\times\varphi(m^{-1}(y_i^{Obs};\eta);\nu)}^{\text{Non-response: Reweighting of } f_{\mathbf{y}}}}{\underbrace{\int f_{\mathbf{y}}(m^{-1}(y_i^{Obs};\eta);\theta) \times \varphi(m^{-1}(y_i^{Obs};\eta);\nu) \times \left(\frac{\partial m^{-1}(y_i^{Obs};\eta)}{\partial y_i^{Obs}}\right) dy^{Obs}}_{\text{Normalizing constant}}}$$

- $f_{\mathbf{y}^{Obs}}(y_i^{Obs};\theta,\eta,\nu)$ is a model for the actual data. Separates features from the income distribution $f_{\mathbf{y}}(.;\theta)$ through θ to those from the 'missing rich' through η and ν.
- External data can be informative in specifying $m(.;\eta)$ and $\varphi(.;\nu)$ but also in having some guesses for η and ν.

[1] For simplification and without loss of generality we'll consider monotonic forms $m(y_i,\mathbf{X}_i;\eta) \equiv m(y_i;\eta)$ and forms $\varphi(y_i,\mathbf{X}_i;\nu) \equiv \varphi(y_i;\nu)$

Bayesian inference under 'missing rich'

The goal within this framework is to make inference on $\phi = (\theta, \eta, \nu)$, but $f_{\mathbf{y}^{Obs}}(y_i^{Obs}; \theta, \eta, \nu)$ can be complex

$\Rightarrow$ **Approximate Bayesian Computation (ABC)** approach:

i) (Informative) initial guesses: prior probabilities $p(\phi)$.

ii) Weight these guesses against the observed data: No analytical likelihood :(Simulate data from the model $f_{\mathbf{y}^{Obs}}(.; \theta, \eta, \nu)$ and compute a distance to the observed data (e.g., compare CDFs).

iii) Output: A weighted sample of parameter values, with probabilities approximately proportional to the Bayesian posterior probability $p(\phi|y^{Obs}) \propto \frac{p(y^{Obs}|\phi) \times p(\phi)}{p(y^{Obs})}$. Basically, what we learnt from comparing initial guesses to the data.

More...

Applications to EU-SILC

- Household disposable incomes OECD-equivalized (HX090) distribution. 2007, 2011, and 2018 EU-SILC cross-sectional samples.
- GB2 assumption for $f_{\mathbf{y}} \Rightarrow \theta = (\alpha, \beta, p, q)$
- **Linear Progressive Underreporting (LPU (Bourguignon (2018))):**

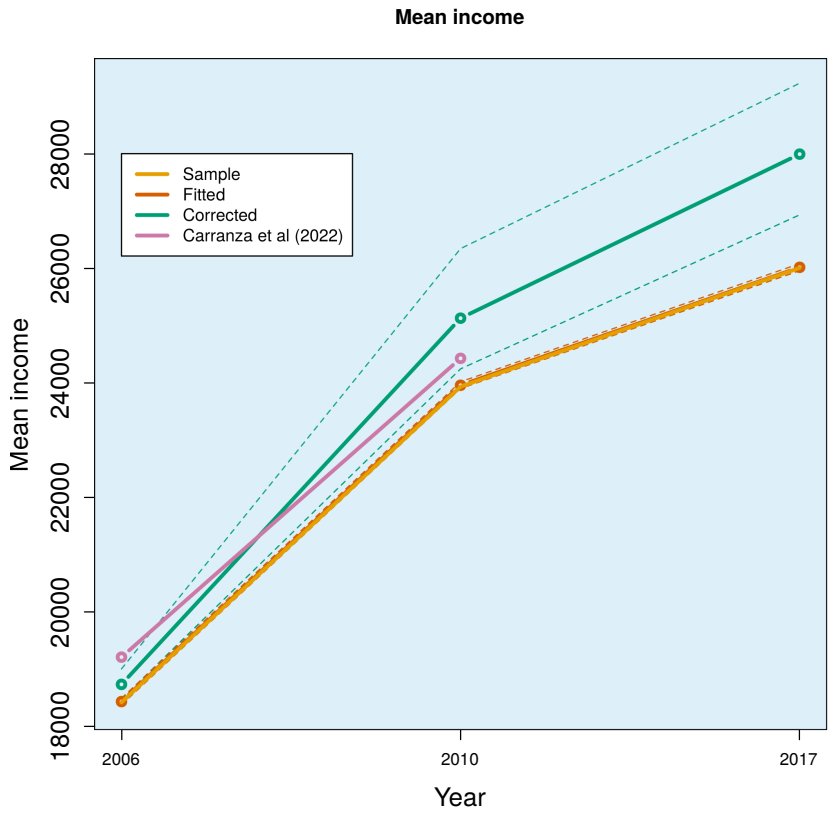
$$m(y_i; \bar{p}, \delta) \equiv y_i - \underbrace{\mathbf{1}(y_i > F_{\mathbf{y}}^{-1}(\bar{p}; \theta))}_{\text{Indiv. with observed incomes above } \bar{p}\text{-th percentile under-report}} \times \underbrace{\delta(y_i - F_{\mathbf{y}}^{-1}(\bar{p}; \theta))}_{\text{Under-reported amount linearly increases with true incomes with slope } \delta}$$

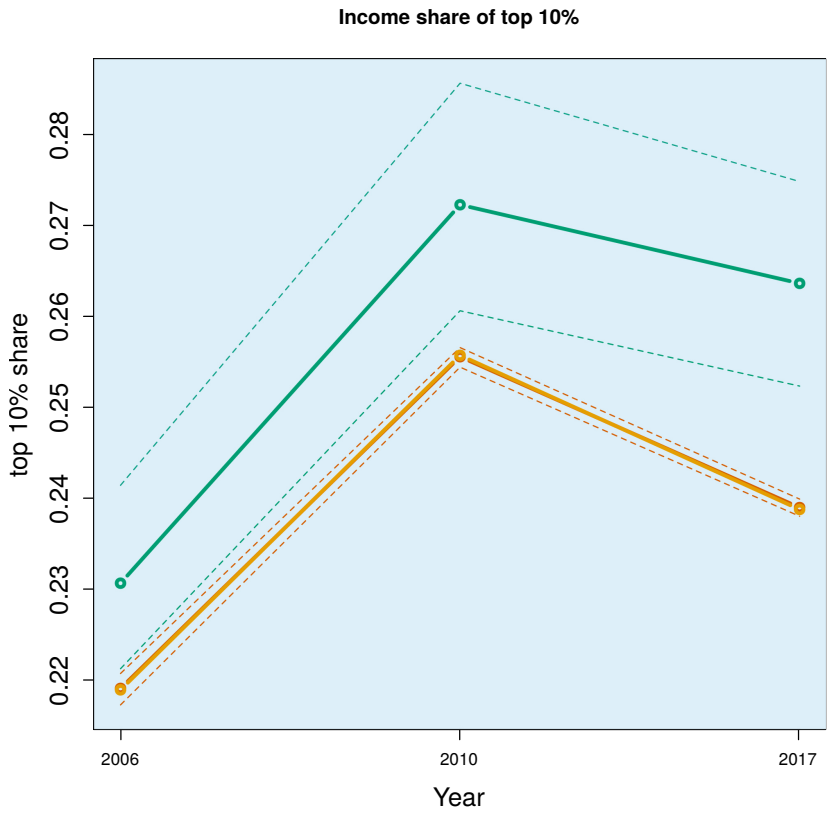
- **Right-truncation (e.g., Alvaredo (2011)):**

$$\varphi(y_i; t) \equiv \begin{cases} 1 \ , & \text{if } y_i \leq F_{\mathbf{y}}^{-1}(t) \\ 0 \ , & \text{if } y_i > F_{\mathbf{y}}^{-1}(t) \end{cases}$$

- $\phi = (\alpha, \beta, p, q; \bar{p}, \delta, t)$
- Comparison benchmark: Distributional National Accounts (DINA) corrections from Carranza et al. (2021).

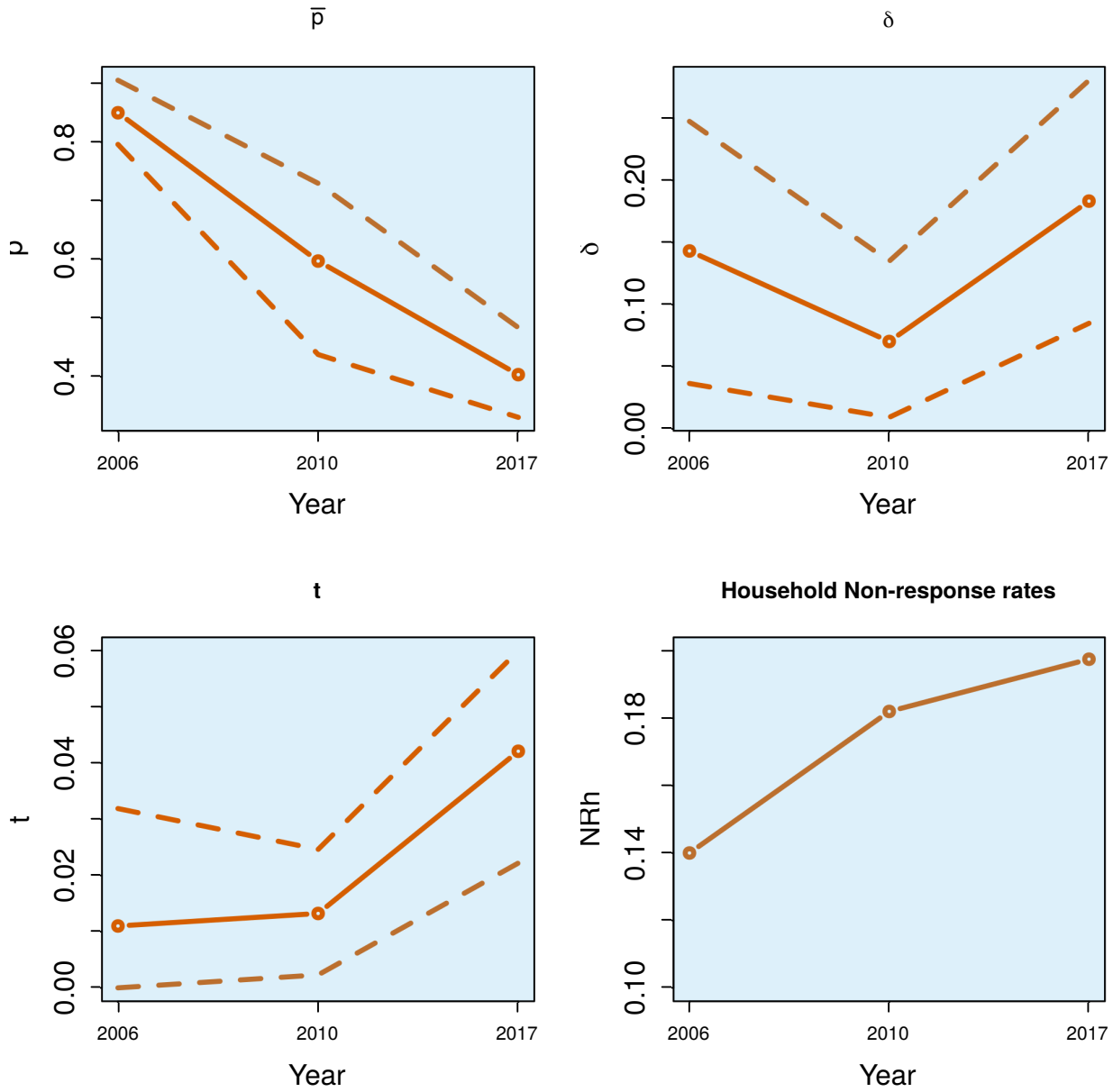
French income distribution





More...

French income distribution



EU-15 income distributions

Table 1: Inequality estimates for EU-15 countries - 2007 vs 2018

Country	Gini 2007	Bias	Top 10% 2007	Bias	Gini 2018	Bias	Top 10% 2018	Bias
Austria	[29.6;32.4]	3.15	[24.4;26.6]	2.77	[29.1;32.6]	2.37	[23.2;25.7]	1.96
Belgium	[30.2;32.3]	4.35	[24.7;26.7]	4.1	[30.7;32.1]	5.04	[25.1;26.4]	4.71
Finland	[32.4;34.3]	6.01	[26.8;28.5]	5.61	[30.8;32.2]	4.49	[26;27.2]	4.28
France	[26.7;28.7]	0.93	[22.1;24.1]	1.17	[29.6;33]	2.34	[25.2;27.5]	2.48
Germany	[33.3;35]	3.18	[26.9;28.6]	3.37	[33.4;36.7]	3.69	[26.6;29.6]	3.82
Ireland	[37.8;39.8]	6.75	[30.2;32.3]	7.04	[33.2;34.4]	2.76	[27.1;28.2]	4.03
Italy	[33.5;36.3]	2.87	[26.1;28.7]	2.75	[34.1;38.1]	2.62	[26.1;29.5]	2.58
Luxembourg	[29.6;31.1]	2.56	[24.2;25.7]	2.57	[31.1;31.8]	-0.22	[24;24.8]	0.86
The Netherlands	[29.8;36.4]	6.63	[25.5;32]	6.46	[30.7;32.1]	3.26	[25.9;27.1]	3.37
Portugal	[40;43.1]	3.34	[31.8;35.4]	4.01	[33.5;36]	1.49	[26.4;28.6]	1.59
Spain	[32.4;34.7]	1.63	[24.5;26.6]	1.63	[33.1;34.6]	1.03	[24.7;26.2]	1.21
Sweden	[24.1;27.3]	1.19	[19.9;22.9]	1.47	[30.1;33.2]	3.87	[23.5;26.6]	4.11
United Kingdom	[34;35.4]	1.53	[26.8;28]	1.95	[34.8;37.6]	3.58	[27.4;30.3]	3.71
EU-15 Average	-	3.39	-	3.45	-	2.79	-	2.98

Note: Posterior distribution Highest Density Intervals (HDI) and mean posterior distribution correction. Greece and Denmark excluded from sample (WIP).

Concluding remarks

Integrating simple 'missing rich' assumptions to standard models of income dist. to exploit data better:

- External information can be integrated through informative priors.
- No need to manipulate income and population concepts to correct for 'missing rich'.
- Uncertainty in posterior distributions integrates uncertainty also on the magnitude of missing/misreported incomes.
- Results in line with previous studies requiring much richer external data.

Thank you for your time and feedback!

ABC: Approximating the likelihood

- In Bayesian inference: $\pi(\theta|\mathbf{y}^{Obs}) \propto L(\theta|\mathbf{y}^{Obs}) \times p(\theta) \Rightarrow$ the likelihood L identifies which values θ considered in $p(\theta)$ are relatively more likely to have generated $\mathbf{y}^{Obs}$ through the model.
- Relatively more likely can be stated in terms of a weighting kernel $K_\epsilon(d(\tilde{\mathbf{S}}, \mathbf{S}^{Obs}))$ with bandwidth ϵ, giving larger weights to values better reproducing the observed data in terms of summaries $\mathbf{S}$ and a distance d (i.e., absolutely lower $d(\tilde{\mathbf{S}}, \mathbf{S}^{Obs})$).
- Approximating the likelihood in this way yields an approximated posterior density:

$$\pi_\epsilon(\theta|\mathbf{S}^{Obs}) \propto \underbrace{K_\epsilon(d(\mathbf{S}, \mathbf{S}^{Obs}))}_{\text{ABC kernel}} \times \underbrace{f(\mathbf{S}|\theta)}_{\text{Prob. of } \mathbf{S} \text{ given } \theta} \times \underbrace{p(\theta)}_{\text{Prior density of } \theta}$$

- $\epsilon \to 0$ increases strictness of approximation and computational cost. $\epsilon \to \infty$ relaxes approximation and forces posterior distribution towards the prior.

Return

ABC: Target posterior distribution

$$\pi_\epsilon(\theta|\mathbf{S}^{Obs}) \propto \underbrace{K_\epsilon(d(\mathbf{S},\mathbf{S}^{Obs}))}_{\text{ABC kernel}} \times \underbrace{f(\mathbf{S}|\theta)}_{\text{Prob. of } \mathbf{S} \text{ given } \theta} \times \underbrace{p(\theta)}_{\text{Prior density of } \theta}$$

Essentially, we've replaced the likelihood for something that's proportional to it:

- Each $\tilde{\theta}$ is translated to its simulated data $\tilde{\mathbf{S}}$. $f(\tilde{\mathbf{S}}|\theta)$ can be 1 if we have parametric expressions for all S statistics in our model. Otherwise, the probability of generating $\tilde{\mathbf{S}}$ from $\tilde{\theta}$ comes into play.
- Each $\tilde{\mathbf{S}}$ is compared to $\mathbf{S}^{Obs}$ and values of $\tilde{\theta}$ giving an $\tilde{\mathbf{S}}$ resembling the data closer than others are given a higher density by the ABC kernel.

  $\Rightarrow$ ABC kernel weights values $\tilde{\theta}$ in a manner proportional to the likelihood.

Return

ABC: MCMC sampling

Initialization:
Set $\Sigma^{(0)}$
Until $K_\tau(\tilde{\epsilon}^{(0)}) > 0$:
Sample $\tilde{\phi}^{(0)}$ from $p(\phi)$
Generate $G\tilde{L}C^{(0)}$ from $f_{\mathbf{y}Obs}(y_i^{Obs}; \tilde{\phi}^{(0)})$
Generate $\tilde{\epsilon}^{(0)} = d(G\tilde{L}C^{(0)}, GLC^{Obs})$
Sampling:
for $t = 1, \ldots, N$ **do**
Sample $\tilde{\phi}^{(t)} \sim g_{\Sigma(t-1)}(\tilde{\phi}, \tilde{\phi}^{(t-1)})$ from the candidate $g_{\Sigma(t-1)}$
Generate $G\tilde{L}C^{(t)}$ from $f_{\mathbf{y}Obs}(y_i^{Obs}; \tilde{\phi}^{(t)})$
Generate $\tilde{\epsilon}^{(t)} = d(G\tilde{L}C^{(t)}, GLC^{Obs})$
Accept and store $(\tilde{\phi}^{(t)}, \tilde{\epsilon}^{(t)})$ with probability:

$$\rho_t = \min\left\{1, \frac{K_\tau(\tilde{\epsilon}^{(t)}) \times p(\tilde{\phi}^{(t)}) \times g_{\Sigma(t-1)}(\tilde{\phi}^{(t-1)}, \tilde{\phi}^{(t)})}{K_\tau(\tilde{\epsilon}^{(t-1)}) \times p(\tilde{\phi}^{(t-1)}) \times g_{\Sigma(t-1)}(\tilde{\phi}^{(t)}, \tilde{\phi}^{(t-1)})}\right\}$$

otherwise store $(\tilde{\phi}^{(t)}, \tilde{\epsilon}^{(t)}) = (\tilde{\phi}^{(t-1)}, \tilde{\epsilon}^{(t-1)})$; Optional: update $\Sigma^{(t)} = Cov(\phi^{(0)}, \ldots, \phi^{(t)})$
end for
Return

French income distribution estimates: Setup

- Data $\mathbf{y}^{Obs}$ summarized through the Generalized Lorenz curve (GLC) at each percentile:

$$GLC^{Obs}(u_k) = \underbrace{s^{Obs}_{(k)}}_{\substack{\text{Share of total income} \\ \text{accumulated at} \\ k\text{ -th percentile}}} \times \underbrace{\mu^{Obs}}_{\text{Sample mean}}, \; u_k = .01, ..., 1$$

- Simulated data $G\tilde{L}C(.;\phi)$ compared to observed through an approximated Wasserstein-1 distance (i.e., compare quantiles one-to-one):

$$d(G\tilde{L}C(.;\phi); GLC^{Obs}) = \sum_{k=2}^{99} \left| (GLC(u_k;\phi) - GLC(u_{k-1};\phi)) - (GLC^{Obs}(u_k) - GLC^{Obs}(u_{k-1})) \right|$$

- Gaussian ABC kernel: $N(0, 25^2)$.

Return

French income distribution estimates: Prior-data conflict

Simple (informal) prior check for big prior-data conflicts:

i) Draw **many** points $\tilde{\phi} \sim p(\phi)$, $p(\tilde{\phi}) > 0$.

ii) Simulate data $\tilde{GLC}(\tilde{\phi})$ from $f_{\mathbf{y}^{Obs}}(y_i^{Obs}; \tilde{\phi}^{(t)})$

iii) Is GLC^{Obs} 'too extreme' on the distribution of simulated $\tilde{GLC}$'s?

Return

French income distribution estimates: Prior-data conflict

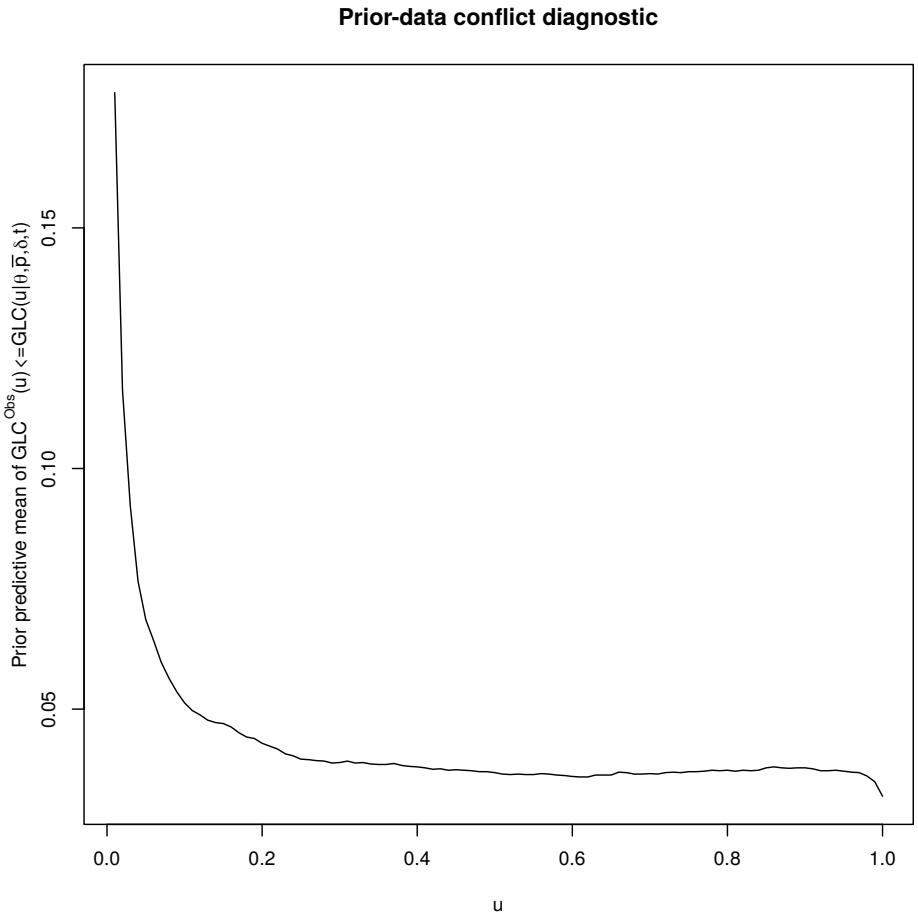


Figure 1: Prior-data conflict diagnostic - 2011

Return

Alvaredo, F. (2011). A note on the relationship between top income shares and the gini coefficient. *Economics Letters*, 110(3):274–277.

Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2019). What did you really earn last year?: explaining measurement error in survey income data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1411–1437.

Blanchet, T., Flores, I., and Morgan, M. (2022). The weight of the rich: Improving surveys using tax data. *The Journal of Economic Inequality*, pages 1–32.

Bourguignon, F. (2018). Simple adjustments of observed distributions for missing income and missing people. *The Journal of Economic Inequality*, 16(2):171–188.

Carranza, R., Morgan, M., and Nolan, B. (2021). Top income adjustments and inequality: An investigation of the eu-silc. *Review of Income and Wealth*.

Ederer, S., Ćetković, P., Humer, S., Jestl, S., and List, E. (2022). Distributional national accounts (dina) with household survey

data: Methodology and results for european countries. *Review of Income and Wealth*, 68(3):667–688.

Flachaire, E., Lustig, N., and Vigorito, A. (2022). Underreporting of top incomes and inequality: A comparison of correction methods using simulations and linked survey and tax data. *Review of Income and Wealth*.

Hlasny, V. and Verme, P. (2018). Top incomes and inequality measurement: a comparative analysis of correction methods using the eu silc data. *Econometrics*, 6(2):30.

Lustig, N. (2020). The missing rich in household surveys: Causes and correction approaches. *ECINEQ Working Paper No. 2020-520*.