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

Michel Lubrano* Mathias Silva‡

October 2021

Abstract

Measurement and/or sampling issues concerning high incomes in survey data can introduce significant biases for measuring levels of inequality. This paper introduces a new approach to account for these 'missing rich' issues in studying the income distributions of EU-member states and of the European Union as a whole using EU-SILC data. A general parametric framework for introducing high-income under-reporting and/or undersampling corrections to parametric distribution models is proposed with detailed derivations for cases where analytical probability density functions, cumulative distribution functions, and Generalized Lorenz curves may be obtained. Specific derivations are presented for two common corrections from the literature. Namely, a right-truncation is exploited as a modelling assumption to allow for corrections for undersampling under the belief that a proportion of the top-incomes population is not represented in the data as in Jorda and Nino-Zarazua (2019). Additionally, a linear progressive under-reporting (LPU) assumption of Bourguignon (2018) is simultaneously introduced as a specific case of this framework to allow for scenarios where only individuals above a certain income threshold under-report their income in the data and the magnitude of under-reporting is linearly increasing with incomes. Under these specific corrections, the framework is taken as a model for each state's data on equivalized disposable household incomes as available on the European Union's Statistics on Income and Living Conditions (EU-SILC) surveys. In doing so, two approaches are explored. Firstly, exact inference is allowed for through the analytically derived likelihood function in the specific case of right-truncation and LPU under a 4-parameter Generalized Beta distribution of the second kind (GB2) model for the population's income distribution. Secondly, because under alternative assumptions for the 'missing rich' corrections or under some data formats such as grouped data with non-reported or random bounds an analytically tractable likelihood function might be unavailable or prohibitively costly

*School of Economics, Jiangxi University of Finance and Economics, China and Aix Marseille University, CNRS, AMSE, Marseille, France.

†Aix Marseille University, CNRS, AMSE, Marseille, France and Ecole Normale Supérieure de Lyon, Lyon, France.

‡Corresponding author: mathias.silva-vazquez@univ-amu.fr. AMSE - Aix-Marseille Université 5-9 Boulevard Bourdet 13001 Marseille, France. This work was supported by French National Research Agency Grant ANR-17-EURE-0020 and by the Excellence Initiative of Aix-Marseille University - A*MIDEX. Declaration of interest: None. `R` replication code for the applications shown in this paper are available upon request and will soon be made available on the authors' Github site. No replication data will be accessible with the replication code. Usual disclaimers apply.

to compute, an approximate inference approach is explored through a new class of simulation-based inference methods known as Approximate Bayesian Computation (ABC) after Beaumont et al. (2002). Both alternatives illustrate the feasibility of inferring all parameters of the income distribution model all while updating prior beliefs on the 'missing rich' margins and magnitudes presumed to affect data at the state-year level. The common framework and compatibility of income data across EU-SILC surveys then allows for aggregating these results across states to study the dynamics of the EU region's interpersonal income distribution without disregarding the heterogeneities concerning the 'missing rich' margins across states and time periods. 'Missing rich' corrected income distribution estimates under this framework are presented for EU15 countries, i.e., Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, and the United Kingdom, and for EU28 countries, which include Bulgaria, Croatia, the Republic of Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Romania, Slovakia, and Slovenia, in addition to those of the EU15, for periods covering the 2005-2017 interval. Two types of benchmarks are taken to gain insight on these estimates. Firstly, 'distribution-free' estimates without any consideration for 'missing rich' issues serve as a *naive* point of comparison to understand the impact and the relevance of including corrections for these issues in terms of the estimated population mean income, income Gini coefficient, top 1% and 10% income shares, middle 40% income share, and bottom 50% income share. Secondly, the World Inequality Database (WID.world) project provides estimates of these quantities from an alternative approach to dealing with 'missing rich' issues which we take as benchmark from a more data-demanding approach. Although WID.world estimates are largely sourced from EU-SILC data, the corrections employed in these estimates exploit several external data sources such as administrative tax data or National Accounts aggregates (Blanchet et al., 2022) and so the disposable household income concept underlying them is not necessarily the same as the one in EU-SILC surveys. Preliminary results illustrate that estimates under the framework proposed in this paper fit the EU-SILC data very accurately while recurring to 'missing rich' corrections which effectively adapt common vague prior beliefs on the nature and magnitude of these issues to the specificities of each state-period. When no corrections are allowed for, our results reproduce the *naive* benchmark estimates. Our 'missing rich' corrected estimates are in general in line with those provided by the WID.world, if somewhat lower in magnitude for some state-years, despite the margins for incompatibilities in comparisons with them due to differences in the concept of disposable household income used. All in all, these findings illustrate the feasibility of our framework for regional income distribution analysis in scenarios where we believe that the only source of data available is prone to problems of undersampling and/or under-reporting of high incomes with heterogeneities in the margin and magnitude of these problems across time and states.

Keywords: 'Missing rich', Approximate Bayesian Computation, EU-SILC.

JEL Code: D31, D63, O15

References

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Blanchet, T., Chancel, L., and Gethin, A. (2022). Why is Europe more equal than the United States? *American Economic Journal: Applied Economics*, 14(4):480–518.

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

Jorda, V. and Nino-Zarazua, M. (2019). Global inequality: How large is the effect of top incomes? *World Development*, 123:104593.