LIS
Working Paper Series

No. 792

Bottom Incomes and the Measurement of Poverty and Inequality

Vladimir Hlasny, Lidia Ceriani, Paolo Verme

May 2020

LIS
CROSS-NATIONAL
DATA CENTER
in Luxembourg

Luxembourg Income Study (LIS), asbl

Bottom Incomes and the measurement of poverty and inequality[*]

Vladimir Hlasny
Economics Department,
Ewha Womans University,
Seoul, Korea

Lidia Ceriani
Georgetown University, USA

Paolo Verme
The World Bank

Abstract

Incomes in surveys suffer from various measurement problems, most notably in the tails of their distributions. We study the prevalence of negative and zero incomes, and their implications for inequality and poverty measurement relying on 57 harmonized surveys covering 12 countries over the period 1995-2016. The paper explains the composition and sources of negative and zero incomes and assesses the distributional impacts of alternative correction methods on poverty and inequality measures. It finds that the main source of negative disposable incomes is negative self-employment income, and that high tax, social security withholdings and high self-paid social-security contributions account for negative incomes in some countries. Using detailed information on expenditure, we conclude that households with negative incomes are typically as well off as, or even better, than other households in terms of material wellbeing. By contrast, zero-income households are found to be materially deprived. Adjusting poverty and inequality measures for these findings can alter these measures significantly.

Note: This paper is also published as an ERF Working Paper and was presented at the ERF-LIS conference on "Inequality Trends Around the Mediterranean."

Keywords: bottom incomes, income inequality, poverty, self-employment, Mediterranean, Middle East, Pareto, random forest

JEL Codes: D31, I32, N35.

[*] This study is part of a pilot cooperation program between Luxembourg Income Study (LIS) and Economic Research Forum (ERF). Hlasny conducted a part of this research as a visiting researcher at LIS (Esch-sur-Alzette, Luxembourg) in February 2019.

1 Introduction

Income surveys are known to exhibit a variety of systematic problems that may bias the measurement of income poverty or inequality such as sampling errors, unit and item non-response, under-reporting, and top coding by statistical agencies. These issues are known to affect the top tail of the distribution and bias the measurement of inequality, an issue that generated a significant body of literature covering high and low-income countries (Atkinson et al., 2011, Cowell and Victoria-Feser, 1996, Hlasny and Verme, 2018, Jenkins et al., 2011). These contributions propose parametric and non-parametric methodologies to correct inequality measures thanks to known top incomes properties (such as Pareto's) and information derived from sources external to surveys (such as tax registers or national accounts).

Less is known about bottom incomes and how their mis-measurement can bias poverty and inequality. Statistical agencies and researchers working on poverty or inequality tend to bottom-code or censor incomes at zero. Some scholars have acknowledged this shortcoming and have proposed to use parametric modeling similar to what is used for top incomes or have studied the sensitivity of inequality indices to changes in bottom values (Cowell and Flachaire, 2007, Van Kerm, 2007, Ceriani and Verme, 2019). However, household surveys are generally assumed to be a good source of information on incomes at the bottom. With a few exceptions (Stich, 1996), this has led to relatively little attention being paid to issues such as zero or negative incomes. This paper studies the prevalence and structure of negative and zero incomes and their potential impact on the measurement of poverty and inequality.

The presence of negative incomes is quite common in household surveys and it is not obvious that these incomes represent poor households. For example, in the sample of 354 household surveys in the Luxembourg Income Study (LIS, 2019) database, 229 surveys contain negative disposable household incomes. In 12 surveys, negative incomes account for over one percent of nonzero incomes, and number up to 584 observations in a national survey.[1] These negative incomes are not trivial in size. Mean negative income is as large in absolute value as 754% of mean nationwide positive income, and exceeds 200% of mean nationwide income in 15 surveys.[2] Whether these negative incomes reflect accurately households' current welfare, or whether they are artifacts of accounting practices or data-entry errors, should be investigated.

Zero incomes are also recurrent in household surveys and the inclusion of these incomes in poverty and inequality measurement presents its own challenge. Among the 354 LIS surveys, 270 surveys contain zero incomes. In 22 surveys, zero incomes account for over

[1] In descending order: ch92, fr84, nl87, pe04, pl95, uk86, kr08, pe10, pe13, fr89, rs06, ie87. Observations with zero incomes are omitted from these counts because zero incomes are a separate problem with an unclear source.

[2] In descending order: nl93, be85, no10, de12, ch82, de01, cz92, no13, co10, de98, gt06, de95, se81, de00, de84. In another 28 surveys, mean negative income is between 100–200% of mean nationwide income.

one percent of non-negative incomes, and number up to 1,213 observations.[3] These zero incomes are often caused by post-survey adjustments such as bottom coding, or replacing missings with zeros, where missings may be caused by item nonresponse, data-entry errors or censoring at zero. Zero incomes could thus be associated with a variety of issues, and survey documentation provided to users typically fails to classify their origins. Again, understanding who is who among zero incomes is essential for generating a consistent ordering among households, and measuring poverty and inequality correctly.

Understanding the bottom tail of an income distribution is also important from a policy perspective, arguably more important than understanding the top tail of the distribution. The bottom tail includes the poor, the income group most in need of assistance and the primary target of social protection policies. Miscounting the poor affects the measurement of poverty and inequality but also contributes to bias targeting exercises such as Proxy Means Testing (PMT) resulting in larger inclusion and exclusion errors. This has direct negative consequences on the livelihood of the poor. By contrast, miscounting the rich affects mostly the measurement of inequality and has limited implications for poverty measurement and targeting.

This paper uses 57 harmonized surveys covering 12 countries to study the prevalence of negative and zero incomes, provides the structure and taxonomy of these incomes and assesses the implications for the measurement of poverty and inequality. It finds that the main source of negative disposable incomes is negative self-employment income and that high tax, social security withholdings and high self-paid social-security contributions account for negative incomes in some countries. Overall, households with negative incomes are typically as well off as, or even better, than other households in terms of material wellbeing. On the contrary, zero-income households are found to be materially deprived. The paper also proposes alternative methods to adjust poverty and inequality measures for these findings and concludes that a proper classification of bottom incomes can alter these measures significantly.

The paper is organized as follows. The next section discusses the conceptual framework used to assess the issues of negative and zero incomes, and the measurement problems posed by them. Section three describes the available data. Section four outlines the main methods used to correct for the negative and zero incomes and assesses the distributional impacts of the corrections. Section five concludes by discussing the results.

2 Conceptual framework

When measuring poverty or inequality, negative and zero incomes are typically either bottom-coded or truncated, and may thus be excluded from measurement. As a result,

[3] In descending order: se67, ci02, ca71, ch82, co04, za08, co07, eg12, it14, at95, ru00, ci15, cl94, cl15, co13, be92, cl09, ca75, hu91, gr95, ru10, cl90. Observations with negative incomes are omitted from these counts for clarity.

inequality and poverty can be mis-measured and, most probably, are under-estimated (Ceriani and Verme, 2019). The reported Gini index, for instance, does not cover non-positive incomes and is based on households' income rank, which is affected by bottom-coding and truncation. The resulting biases in inequality measurement are problematic statically for understanding income distribution within as well as across countries, but also dynamically for understanding the evolution of inequality over time.[4] Negative incomes may also be found among non-poor households so that counting negative incomes among the poor can bias the measurement of poverty upwards. Hence, biases generated by bottom-coding and truncation may operate in opposite direction than biases generated by negative incomes making the proper assessment of poverty and inequality very complex. This section provides a taxonomy and a methodology to properly account for negative and zero incomes when measuring poverty and inequality.

2.1 Definitions

As a measurement unit, we use Disposable Household Income (DHI) equivalized per adult equivalent using the square root of household size. DHI is what is normally reported by households in surveys. It is what is used to measure poverty and inequality in richer countries and, unlike individual incomes, it is a measure that can be used to account for family benefits or liabilities like child benefits or mortgages.

Our measures of inequality and poverty are the Gini index and the poverty headcount ratio. The Gini index is what the existing literature on inequality corrections has predominantly used and is the measure used by LIS for cross-country comparisons.[5] While the Gini is less sensitive to issues in the tails of the distribution than other inequality indexes, it may be quite sensitive to the presence of non-positive incomes and the researcher's decision to keep, truncate or correct them. The poverty headcount ratio is the share of households falling below the poverty line. The headcount ratio is less sensitive to corrections for non-positive incomes than, say, the Foster-Greer-Thorbecke indicators of poverty intensity including the income gap ratio (IGR) (James et al., 1984, 2010).[6] Therefore, the poverty and inequality measures considered are the least sensitives to changes in the tails and, as such, our corrections should be considered as lower bound corrections as compared to other measures.

Negative and zero DHI observations come about for a variety of reasons and it is important to stress that they have very different origin and should be treated separately.

[4] Take for instance the French survey: In 2005 there were no zeros and three negatives, while in 2010 there were 117 zeros and 25 negatives (refer to table A1). The approach to dealing with these observations can affect greatly the estimated growth in inequality.

[5] Refer to the LIS key figures at www.lisdatacenter.org/data-access/key-figures.

[6] As alternative inequality and poverty indexes, we also report Theil's entropy index (aka., generalized entropy GE(2), or half the squared coefficient of variation) and the IGR in the appendix. In fact, even the Gini is sensitive to negative incomes, and may itself become negative if mean income is negative, or greater than 1 in the presence of large negative incomes (Scott and Litchfield, 1994).

Among negative incomes, we should strive to distinguish those ‘valid’ from a welfare and capabilities perspectives - implying that households are unable to meet basic needs and lack essential capabilities - and those due to accounting considerations without real effects on households’ material wellbeing. By contrast, zero incomes are typically generated by post-survey adjustments such as bottom coding, or replacing of missing observations with zeros. One should distinguish the unlikely ‘valid’ zeros from those generated by survey administrators.

Having defined the main aggregate and the distinction between negative and zero incomes, we can now define the components of income that are relevant for our analysis. For this purpose, we will use the taxonomy used by LIS. Negative and zero Household Income (HI) can come in the form of labor (HIL), capital (HIC) or transfer income (HIT) with $HI = HIL + HIC + HIT$, or high income tax liability (HXITI) and social security contributions (HXITS) with $HXIT = HXITI + HXITS$. The income components could be further subdivided into paid employment income (HILE) and self-employment income (HILS) with $HIL = HILE + HILS$, interest and dividends (HICID), voluntary individual pensions (HICVIP), rental income (HICREN) and royalties (HICROY) with $HIC = HICID+HICVIP+HICREN+HICROY$, and social security transfers (HITS) and private transfers (HITP) with $HIT = HITS + HITP$. Liability components could be subdivided into income tax withholdings (HXITIW) and adjustments (HXITIA) with $HXITI = HXITIW + HXITIA$, and social security contributions paid by self (HXITSS) and paid on behalf of others (HXITSB) with $HXITS = HXITSS + HXITSB$. In sum (in red the potential negative incomes):

$$\begin{gathered}
\overbrace{\overbrace{(HILE+HILS)}^{\text{Labor Income }(HIL)} + \overbrace{(HICID+HICVIP+HICREN+HICROY)}^{\text{Capital Income }(HIC)} + \overbrace{(HITS+HITP)}^{\text{Transfers }(HIT)}}^{\text{Household Income }(HI)} \\
- \\
\underbrace{\underbrace{(HXITIW+HXITIA)}_{\text{Taxes }(HXITI)} + \underbrace{(HXITSS+HXITSB)}_{\text{Social Security Contributions }(HXITS)}}_{\text{Fiscal Liability }(HXIT)}
\end{gathered}$$

In each survey, among households with negative disposable incomes, we calculate how many households have negative capital income, negative self-employment income, or tax withholdings and social security contributions higher than gross income. We also calculate mean negative capital income, mean negative self-employment income, and mean excess of fiscal liability over gross income, and compare these means to the mean negative DHI. These statistics indicate how important capital income, self-employment income, and undue liabilities are in bringing about negative disposable incomes in each survey.

As a measure of undue liabilities for taxes and social security contributions, we evaluate the part of DHI that is expected to be non-negative, net of total taxes and contributions. To do so, we subtract from DHI (which is already net of taxes and contributions) three potentially negative income components: self-employment income, interest and dividend income, and private transfers ($DHI - HILS - HICID - HITP$). If the result is negative, this could indicate overpayment in taxes and contributions relative to what was due on current income (net of self-employment, financial-assets and private-transfer earnings). Finally, we distinguish the effect of tax withholdings, adjustments, and social security contributions ($HXITIW, HXITIA, HXITS$).

Even when negative or zero incomes are accurate, including them in the distribution of incomes can be problematic for the purpose of distributional analysis, because these values may not reflect the households' short-term or long-term capabilities, consumption or welfare. Moreover, the negative values may mismeasure households' true annual incomes. Self-employment income in particular is prone to mismeasurement (Eurostat, 2006a). First, evidence from comparing the distribution of self-employment income in survey and tax data in Latin America suggests that this income tends to be underreported in surveys across all distribution quantiles. Hence, negative self-employment incomes may come from underreporting. Second, household surveys provide information over a short sampling period when the self-employed may have been mostly expending resources on self-employment related activities, whereas gains from self-employment may have materialized only later without being captured in the survey snapshot. Third, self-employment income might be more difficult to report accurately in surveys, because the respondents need to recall not only how much they gained from their sales or services but also their annualized investment in self-employment activities.[7]

2.2 Assessing the composition and sources of non-positive incomes

We start by assessing the prevalence of negative and zero incomes across survey waves and across countries. We survey their size distributions, and we identify the likely culprits of the observed values. We then draw qualitative conclusions regarding the true capabilities and wellbeing of the respective households, using information from the available income components and alternative measures of households' economic status. Namely, we evaluate the association between negative or zero incomes and households' observed capabilities including education (secondary or higher, $EDUCLEV \geq 32$) and subjective health rating (good, $HEALTH_C \leq 2$), to uncover patterns and irregularities. We also evaluate the links between incomes and households' functionings including total consumption (HC), food consumption ($HCFOOD$), and home ownership ($OWN \leq 120$), as per data availability across surveys. We compute households' 'monetary overconsumption' as the excess of

[7] The authors are grateful to Holguer Xavier Jara Tamayo for a helpful correspondence on these points.

total monetary consumption (HC) over final monetary income.[8] From the analysis of this overconsumption between households with negative, zero, and positive DHI, we assess the quality of the respective observed DHIs as measures of households' capabilities and welfare.

The careful incidence analysis of non-positive incomes by source and by household type is important, because the relationship may be complex and non-monotonic. Households' overconsumption is a case in point. Accruing debt may be a survival strategy for the poor, investment strategy for the middle class, or a tax evasion strategy for the rich. A testable conjecture may be that small negative incomes are prevalent among chronically poor people who are temporarily in trouble, while large negative values are prevalent among chronically rich people under-reporting, or writing off capital losses or tax assessments from past years (Eurostat, 2005). Disentangling between these groups is essential for deriving a relevant measure of household well-being, which is instrumental for targeting social programs. In sum, non-positive incomes are clearly short of the 'wolf point' of income necessary for bare survival (Davis, 1941:405). Finding that households with non-positive incomes do not have a profile of deprived units, we may wish to truncate the reported non-positive values of individual income sources, or replace them with values from households with matching characteristics.

2.3 Adjusting for non-positive incomes

Standard statistical adjustments for negative and zero incomes include data trimming or bottom coding (Eurostat, 2006b). We apply these corrections and compare them with the corrections provided by two more advanced methods: parametric modeling of non-positive incomes, and non-linear random forest imputation of incomes using information on consumption, asset wealth, savings, investment, and other household characteristics.

Among parametric-modeling studies, Van Kerm (2007)(p. 8) fitted an inversed Pareto distribution to negative incomes, using the following cumulative distribution function:

$$F^L(y;\theta;y^u) = \left(\frac{2y^u - y}{y^u}\right)^{-\theta} \quad for \quad y < y^u \tag{1}$$

where y is income, and y^u is the upper cutoff for modeling bottom incomes, such as $y^u = \min(\max(0.3\mu, Q(0.02)), Q(0.03))$, proposed empirically by Van Kerm (2007) (p.9). $\theta > 0$ is an estimable shape parameter that can be made robust to extreme incomes using Maria-Pia Victoria Feser's (Victoria-Feser and Ronchetti 1994; Victoria-Feser 2000; Cowell

[8] Final monetary income is taken to be inclusive of special transfers and benefits, indirect subsidies, and windfall income, less of other taxes, voluntary contributions, inter-household transfers paid, charity donations, and interest paid. $FMI = DHI + add.transf1\&2 + (HWL + HWC + HWT) - (HXOT + HXVC + HXIH + HXCH + HXINT)$, where $addit.transfers1\&2 = HIATOLD + HIATDIS + HIATSUR + HIATSIC + HIATFAM + HIATEDU + HIATUNE + HIATHOU + HIATCSP + HIATWIC + HIATCAR + HIATVET$.

and Victoria-Feser 2007) optimal B-robust estimator, essentially scaling down the weight of observations deviation from the fitted pattern.

Dagum (1990, 1999); Jenkins and Jäntti (2005); Jäntti et al. (2015) also proposed fitting an exponential distribution to negative data using a point-mass for zero incomes. The corresponding cumulative distribution function is

$$F(y;\alpha;\beta;\gamma;\delta;\pi_2;y^u) = \begin{cases} \pi_1 \exp(\delta y) & \text{for } y < 0 \\ \pi_1 + \pi_2 & \text{for } y = 0 \\ \pi_1 + \pi_2 + (1 - \pi_1 - \pi_2) SM(y;\alpha;\beta;\gamma) & \text{for } y > 0 \end{cases} \tag{2}$$

where π_1 and π_2 are the shares of negative and zero incomes, respectively, α, β, γ, $\delta > 0$ are estimable parameters, SM is the cumulative distribution function of the Singh-Maddala distribution.

Among imputation methods, Ceriani and Verme (2019) have proposed matching estimators to assess the accuracy of components of the welfare aggregate by constructing "the correct sample counterpart for the missing information on the treated outcomes had they not been treated, by pairing each participant with members of the nontreated group" (Blundell and Costa Dias, 2009: 593). In our case, zero and negative incomes would be the treated group and all those with positive incomes the non-treated. The matching is performed based on households' demographics including household composition, sector of employment, and housing.

In this paper, rather than using a matching method, we propose to use a random forest algorithm to predict household welfare based on household observable characteristics. Machine learning algorithms can improve the accuracy of imputation and random forest in particular has been shown to be very effective in prediction exercises as compared to standard econometric models (Breiman, 2001; Haziza and Beaumont, 2007; Zabala, 2015; Athey and Imbens, 2019). This is also the case for poverty predictions as shown by a recent experiment conducted by the World Bank.[9]

2.4 Assessing the distributional impact of imputing non-positive incomes

With the alternative estimates for the distribution of bottom incomes, we re-calculate poverty and inequality measures. We propose a decomposition of inequality and poverty changes due to the different hypotheses on the distribution of bottom incomes. This decomposition clarifies the relative importance of negative and zero incomes in explaining changes to the inequality and poverty measures.

When the income distribution includes both negative and non-negative incomes, the Gini coefficient can be obtained using the Lorenz curve from the Ginis among negative incomes (G_N) and among non-negative incomes (G_{1-N}) by knowing the population share

[9] See Fitzpatrick et al. (2018) and details of this competition on GitHub: https://github.com/worldbank/ML-classification-algorithms-poverty.

of households with negative incomes (π_N) and their share of aggregate net income (S_N, a negative share). Refer to Figure 1 for derivation.

$$G = -G_N \pi_N S_N + \pi_N - S_N + G_{1-N}(1 - \pi_N - S_N + \pi_N S_N) \tag{3}$$

Here G_{1-N} is computed non-parametrically from data, π_N is observed, S_N is observed or computed in a corrected income distribution, and G_N is estimated non-parametrically or parametrically using the corrected distribution of negative incomes.

[Figure 1]

3 Data and descriptive statistics

Our study relies on 57 household surveys from 12 countries for the years 1995-2016, harmonized and made available by LIS and ERF. The LIS database contributes income distributions for seven countries, namely Greece, France, Israel, Italy, Serbia, Slovenia and Spain, while the ERF database contributes surveys for Egypt, Iraq, Jordan, Palestine, and Sudan.[10] These countries are particularly interesting for our analysis because they represent high, medium and low-income countries and high levels of tax evasion, low formal employment, and high rates of self-employment relative to their income level. These are properties generally associated with high prevalence of low reported incomes. Among these surveys, there are also subsets with similar income distributions, yet different prevalence and composition of non-positive incomes.

The data are not without problems. The survey documentation does not explain the source of zero and negative incomes, which implies that, to understand these incomes, we need to rely on within-data evidence. Also, among the variables available, some income components are missing between LIS and ERF surveys (or between the different sources of household data in both repositories), and cannot be assessed across the entire sample of surveys.[11] With non-income variables, the problems are analogous. This explains the various gaps in Tables 1 and 2.

Across the 57 surveys considered, 33 have zero values (57.9%) accounting for up to 173 observations and 1.5% of the sample, and 34 have negative values (59.6%) accounting for up to 107 observations, which can be as large on average as 104% of positive incomes (Table 1). Among the northern Mediterranean surveys, zero incomes are more prevalent than negative incomes in France and Italy, as prevalent in Greece, Serbia and Spain, and entirely or nearly

[10] Egypt 2012 is available in both databases, using data from alternative sources: LIS dataset is from the Labor Market Panel Survey (LMPS), while the ERF dataset is from the Household Income, Expenditure and Consumption Survey (HIECS).

[11] Paid employment income is missing for Iraq, Palestine and Sudan. Self-employment income is missing for Palestine. Rental income is missing for Egypt 1999 and Palestine. Interest earnings and individual pensions are missing for Palestine, Sudan, Greece 1995, Spain 1995, Israel 1997, Italy 1995, and Slovenia.

non-existent in Israel and Slovenia (respectively). Among the southern Mediterranean surveys, only the Egyptian 2012 survey in the LIS database, and the Iraqi 2007 survey contain negative incomes, but zero incomes appear in Iraq 2012, and in the Palestinian and Sudanese surveys as well. These cross-country differences endure qualitatively over time, suggesting that they may have to do with survey instrument problems (e.g., source of income data, type of recall on interviews) and administrators' practices (e.g., bottom coding, imputation), rather than with socio-economic countries' conditions. When negative incomes are present in a survey, they vary across households suggesting that the values represent some meaningful differences in the households' income components. The only exception is Greece 1995, where the 17 negative incomes are all -10,000 drachmas (€-29.35), indicating that they are due to bottom coding of self-employment income.

Table 1 also reports the distribution of self-employment income ($HILS$), undue liabilities for taxes and social security contributions ($DHI - HILS - HICID - HITP$), and the burden of social security contributions alone ($HXITS$). A quick review suggests that, empirically and among the various income components, there is one predominant source of negative disposable incomes: negative self-employment income. The remaining cases are due to unduly high self-paid social security contributions, and other burdens, such as high property taxes, loan repayment, or negative inter-household transfers (e.g., alimonies, remittances, family transfers; Eurostat, 2006a). The prevalence of negative incomes, and the contribution of individual factors - self-employment income, social security contributions, and other burdens - differ across countries and across years. It turns out that capital income is non-negative for all households and all surveys, so it does not contribute to explain negative DHIs (not reported in Table 1).

Surveys for Greece, Italy and Serbia in the LIS database show that up to one percent of households report negative disposable incomes, linked to negative self-employment incomes (of approximately 50-150% of the reported negative DHI). In Greece 2013 and recent Italian surveys, tax and social security withholdings also account for a handful of negative incomes (112% of the reported negative DHI in Greece, and 148-290% in Italy 2010-2014). In Israel, the count of negative $HILS$ is lower, but the values are much larger (of 350-2,000% of the size of the negative DHI). In Spain, the negative incomes are predominantly due to self-employment (120-200%), but in 2004 the three negative incomes were due to large income-tax burdens. In Egypt 2012, there are 191 households with zero incomes, and 10 with negative incomes. These 10 are on account of large negative $HILS$.

In sum, the available evidence suggests that negative $HILS$ is the primary source of negative DHI in three-quarters of all surveys, while in other surveys the problem is mainly due to high social-security and other burdens. Interestingly, when surveys are sorted by the frequency of negative DHI, negative HILS shows up as the top source of their prevalence. When surveys are sorted by the relative magnitude of negative incomes, high inter-household transfers and undue social-security and other burdens dominate as sources of the high relative magnitude of negative incomes. We may generalize that the prevalence of negative incomes is primarily due to negative self-employment incomes, while the extreme values of negative incomes are typically due to extremely high social-security contributions, non-income taxes, and paid remittances.

[Table 1]

[Figure 2 and 3]

3.1 The association of incomes with other socioeconomic outcomes

Next, in each survey where it was available, we calculated mean household consumption, consumption of food, homeownership, self-reported health and education among households with non-positive DHI, and we relate these figures to those for households with positive DHI. This helps to identify the true welfare of households with non-positive DHI across different surveys.[12] We also calculated mean outflow from mortgages, loans and repayments, to proxy for the households' level of debt. Refer to Table 2, and Figure 4. Interestingly, in France, Iraq and Italy, households with negative DHI have higher total consumption, food consumption, and home ownership (>100%) than the respective national means among households with positive DHI. In Greece, Israel and Spain, households with negative DHI fare somewhat worse or at least not clearly better than the national mean. Nevertheless, they are not obviously consumption-deprived.

Information on outflows for mortgage and loan repayment is available for fewer households and surveys, and the only observable pattern is that in France households with negative DHI are more burdened by debts than the national mean, while in Greece and Israel this burden is less. For Italy and Spain, no clear patterns emerge. Regarding the completion of secondary education, it is not entirely clear whether households with negative DHI are more educated (as it appears in France and Greece) or less educated (Iraq). Perception of health is available only for selected survey waves in Greece, Italy and Spain, but households with negative DHI systematically outperform their respective national means.

These patterns differ clearly from those for zero-income households. Zero-income households in France, Italy and Slovenia have total consumption of 19.6-48.3 percent of the respective national means, and food consumption of 40.7-68.7 percent. In all countries where they are available, home ownership rate and debt maintenance are also lower among zero-income households than the national means. On the other hand, their health appears to be better. Their education level is not clearly different from the nationwide statistics, except in France and Italy (clearly worse), and Greece (better).

[12] We have also evaluated this on all 354 surveys in the LIS database. Consumption is available in 43 surveys, and food consumption in 73 surveys. Negative-DHI households do not appear to have unduly low consumption. Mean food consumption of negative-DHI households does indicate some cause for concern. Some negative-DHI households appear to be food-poor.

Regarding their residence, households with negative incomes in Egypt, Iraq, Serbia and Spain are less likely to reside in urban areas, while in France, Greece and Israel they are as likely (and in Italy, more likely) to be urban as their peers with positive incomes. Those with zero incomes in Egypt, Iraq and Sudan are also less likely to be urban, while zero-income households in France, Greece and Serbia are more likely to be urban (and as likely as them in Italy, Palestine, and Spain).

These patterns suggest that households with negative DHI are typically as well off as other households in terms of material wellbeing, or even better off. They appear to be healthier and at least as educated. By contrast, zero-DHI households are materially deprived, even though their human capital (it terms of health and educational attainment) is not clearly lower than that of their compatriots.

[Table 2]

[Figure 4]

4 Adjusting welfare measures for non-positive incomes

4.1 Bottom-coding negative incomes

Table 3 and Figure 6 present the Gini coefficients and the poverty headcount ratios estimated on the source data or corrected using traditional correction methods, that is by bottom-coding incomes at zero, truncating negative incomes, or also truncating zero incomes. Applying these incrementally intrusive approaches one by one - first bottom-coding (censoring) at zero, then deleting (truncating) values that were initially negative, and then deleting all remaining zeros (truncating non-positives) - leads to a systematic monotonic fall in the inequality and poverty indexes.

We find that bottom-coding negative incomes at zero leads to a noticeable decline in the Gini, by up to 1.2 percentage points, particularly in RS13 (1.2pc.pt.), GR07 (0.8pc.pt.), and IL14, RS06, RS10 and RS16 (0.6-0.7pc.pt.). Bottom-coding negative incomes has no effect on poverty because negative observations are below the poverty line by definition. Truncating negative incomes - compared to bottom-coding them at zero - further reduces the Gini by as much as 0.7 percentage points, most notably in ES10, RS06, RS13 and RS16. Truncating negative incomes reduces poverty by as much as 0.5 points, most notably in GR07 and IL12. Finally, deleting zero incomes in addition to negative incomes lowers the Gini by another up to 1.1 percentage points, particularly in IT14 (1.1pc.pt.), EG12 (0.7pc.pt.) and GR95 (0.6pc.pt.), and lowers poverty by up to one percentage point, particularly in EG12 (1.0pc.pt.), and GR95, ES10, IT00 and RS13 (0.6-0.7pc.pt.).

In sum, the traditional corrections for non-positive incomes have noticeable effects even on inequality and poverty measures known to be relatively robust to adjustments in the income-distribution tails. Between the uncorrected values and the values corrected by deleting all non-positive incomes, the Gini falls by up to 2.3 percentage points (2.3pt. in RS13; 1.2-1.8pt. in GR07, IT14, RS06, RS10, RS16, ES10; and 0.8-0.9pt. in EG12, GR95, ES95, ES13), while poverty falls by up to 1.5 points (1.1-1.5pt. in EG12, IT14, ES10; 0.8-0.9pt. in GR95, RS06, RS10, RS13, ES95, ES13).

[Table 3 and Figure 6]

For countries with three or more time observations, we can evaluate how the correction affects trend and volatility in inequality and poverty. Figure 5 shows that in Greece and Serbia (across the 6 or 4 years, respectively), the correction somewhat dampens the downward trend in inequality and poverty, while in Italy (across the 7 years) it strengthens it. The correction does not appear to affect volatility, except in the case of the Serbian Gini, which falls with the correction.

[Figure 5]

4.2 Replacing negative values with parametric Pareto distributions

Following Van Kerm (2007), we proceed by replacing negative income values with smooth parametric distributions estimated on those values. Table 4 reports on an exercise replacing negative income values with inversed one-parameter Pareto (type I) distribution, or the two-parameter generalized Pareto (II) distribution. We find that the Pareto (I) distribution does not provide a good fit to the observed negative incomes, because the estimated Pareto coefficients are universally too low, implying excessive dispersion among negative incomes with an undefined mean. Only for eight surveys (out of 33 surveys containing 2+ negative incomes) do we get plausible results. In these surveys, combining the parametric Ginis for negative incomes with nonparametric Ginis for positive incomes yields trivial corrections to the Ginis reported in Table 3, of the order of 0.01 percentage points of the Gini. This is due to the small number of negative incomes in the surveys. The overall Gini appears robust to the method for treating negative incomes.

The two-parameter generalized Pareto distribution provides a somewhat better fit, thanks to the flexibility provided by the additional parameter. For 18 surveys out of 33, we estimate plausible coefficients, income shares and parametric Ginis. Combining the parametric Ginis for negative incomes with nonparametric Ginis for positive incomes yields small corrections to the Ginis, of up to 0.70 percentage points of the Gini (in Serbia 2006-2013; with an outlier of a 25pc.pt. correction in RS16), and typically of 0.01-0.02 points. Once again, the corrections are very small, on account of the small number of negative incomes in the surveys.

[Table 4]

4.3 Imputation of non-positive incomes with random forest

Next we implement a random forest ensemble classification of positive income observations, to replace non-positive incomes with the households' most likely positive values in relation to other households based on their observed characteristics. The intuition is that, while we cannot trust the non-positive incomes, we can rely on households' other characteristics to impute the most likely positive income value given the households' similarity to other households with positive incomes.

Compared to alternative imputation methods - such as regression prediction and propensity score matching - random forest classification has several advantages including a higher likelihood to find the best fit, lower sensitivity to missing values, and flexibility to the reliance on categorical variables (Zhao et al., 2017). One pitfall is possible overfitting, underscoring the importance of imposing restrictions on the depth of the modeled trees.

The method classifies observations into an endogenously selected number of nodes (positive integer values of income here) on a constructed classification tree and estimates the probability that each observation belongs to each node. This is repeated 100 times. The classification is based on households' observed characteristics, namely household size (binaries for 3 quantiles), urban/rural residence, house ownership, and household head's education (binaries: none, primary, secondary, tertiary), self-perceived health (binaries: bad or very bad, fair, good, very good), age and age squared (binaries for 3 quantiles).[13] These variables are selected as they proxy for households' earning capacity or economic status, and they are available across the majority of surveys.

For each household, we identify the node (or positive income value) with the highest probability, and we replace non-positive incomes in the survey with these best-matched positive values. First, we do this for self-employment income only, because negative $HILS$ is the most prevalent cause of negative DHI. (Zero $HILS$ are very common, among households not engaged in self-employment, so these values are not replaced.) We recalculate DHI for households that initialy had non-positive DHI and negative $HILS$ using the best-matched positive $HILS$ (column 1 in Table 5). Next, we repeat the classification exercise for our measure of income less undue liabilities for taxes and social security contributions ($DHI - HILS - HICID - HITP$). We again recalculate DHI for households that initially had non-positive DHI and negative ($DHI - HILS - HICID - HITP$) using the best-matched positive values (column 2 in Table 5). Finally, we repeat the classification exercise for DHI itself, and we replace non-positive DHI with the best-matched positive values according to the node with the highest probability for the household (column 3 in Table 5).

[13] The algorithm is based on chi-square automated interaction detection (CHAID). The algorithm constructs 100 classification trees. This is thought to produce more accurate predictions than a single classifier such as a logistic model, particularly for out-of-sample units (Luchman, 2015). The routine is estimated in Stata 13 software as follows: $chaidforestX, unordered(hlth_fairhlth_goodhlth_verygoodedu_primedu_secedu_terownhouserural)xtile(ageage2nhhmem, nquantiles$

Table 5 reports the corrections to the Gini coefficients and the poverty headcount ratios. Results in column 1 show that the random forest classification for $HILS$ produces modest changes to the distributions of DHI, because only small numbers of non-positive DHI become positive when their $HILS$ is replaced. The corrections to the Gini are as high as 1.73 percentage points in Serbia (1.22pc.pt. in 2006; 0.82pt. in 2010; 1.73pt. in 2013; 0.77pt. in 2016), 1.02 points in Spain (0.45pt. in 1995; 0.56pt. in 2007; 1.02pt. in 2010; 0.41pt. in 2013), and 0.88 points in Greece 2007, but amount to only 0.01-0.33 points of the Gini for other surveys. The mean Gini correction across all surveys is 0.29 points, while the median is only 0.13 points.

Next, using the random forest classification method on income less undue liabilities for taxes and social security contributions ($DHI - HILS - HICID - HITP$), and using the best-matched positive values for them yields typically larger downward corrections to the national Ginis, of up to 1.51 points. Refer to Table 5 column 2. In Egypt 2012, Greece 2007, Italy (2014), and Spain (2010), the Gini falls by 1.04-1.51 points. The mean Gini correction across all surveys is 0.48 points and the median is 0.21 points.

Finally, using the random forest classification method on DHI itself, and converting all non-positive DHI into positive values yields typically larger downward corrections to the national Ginis, of up to 1.61 points (mean 0.53pt, median 0.43pt). Refer to Table 5, column 3. In Egypt 2012, Greece 2007, Italy (esp. 2014), Serbia (esp. 2013) and Spain (esp. 1995, 2010), the Gini falls by as much as 1.04-1.61 points. The correction to the Serbian Gini is surprisingly weaker than the correction in column 1, when negative $HILS$ alone was being replaced and non-positive DHIs were being recomputed using the new $HILS$. Here, non-positive DHIs are imputed directly, and all of them are turned into positive values, but the Gini falls only by up to 1.4 points (1.0pt. in 2006, 0.7pt. in 2010, 1.4pt. in 2013, and 0.6pt. in 2016). The explanation lies in the extent of the correction. In column 1, negative $HILS$ were corrected by a larger extent, so that the few corrected non-negative DHIs rose significantly above zero, while in column 2 the correction to all non-positive DHIs put them just above zero.

Our poverty measure is also sensitive to the random forest corrections. Results in Table 5, column 1 show that the random forest classification of $HILS$ produces a significant change only for EG12. However, corrections in column 2 are significant for EG12, GR07, IT14, ES10 and ES13 and corrections in column 3 are significant for EG12, GR07, IT00, IT14, ES04, ES10 and ES13. In all these cases, corrections lead to lower poverty. This is evidently due negative incomes being reclassified into positive ones by the random forest classifier. For completeness, Figure 7 shows the time trends in the Ginis and the poverty headcount ratios, both uncorrected and corrected, for Greece, Italy and Serbia where three or more time observations are available. The correction using random forest imputation does not appear to dampen volatility, except in the case of the Serbian Gini, exactly as we found in Figure 5 for the traditional correction methods.

[Table 5]

[Figure 7]

5 Discussion

The paper reviewed the prevalence of negative and zero incomes in household surveys, and their implications for the measurement of inequality and poverty. We relied on 57 harmonized surveys for 1995-2016 from 12 states around the Mediterranean region. We find that there is one predominant source of negative disposable incomes across most countries: negative self-employment income (particularly relevant in Egypt, Israel, and Spain). In addition, tax and social security withholdings (France, Greece, Israel and Spain), and unduly high self-paid social security contributions (Israel 2007-10 and Italy 2010) also account for a handful of negative incomes in several countries.

Using several observable measures of households' characteristics, we find that households with negative DHI are typically as well off as other households in terms of material wellbeing, or fare even better. They appear to be healthier and at least as educated. By contrast, zero-DHI households are materially deprived, even though their human capital stocks (it terms of health and educational attainment) are not clearly lower than those of their compatriots.

To correct income distributions for the unreliable non-positive incomes, we implemented two alternative methods beside traditional bottom-coding or truncation: the recently promulgated approaches of replacing extreme income observations with smooth parametric distributions; and imputation using random forest classification of incomes. The results of these estimations were summarized in Tables 3-5. We find that the traditional approaches produce non-trivial corrections of up to 2.3 points of the Gini, and 1.5 points of the poverty headcount ratio. The enduring problem with these approaches is that they do not use all information available in surveys, they do not replace unreliable zero or negative incomes with more realistic values, and they produce income distributions that are truncated at the bottom or have discontinuous point-mass at zero incomes (Ostasiewicz and Vernizzi, 2017).

Corrections via replacement with parametric distributions are less effective, possibly because of the poor fit with the evaluated distribution functions, and because they restrict their focus on incomes under the same presumed distribution function - that is, negative incomes but not zero incomes. Pareto distributions do not fit the observed negative incomes well, with the estimated coefficients being too low, implying unrealistically large dispersion among negative incomes. The two-parameter generalized Pareto distribution fits better, giving rise to realistic parametric means and Ginis for negative incomes, but still yields very small corrections to the overall Ginis, of up to 0.7 points. One reason is that this approach does not address parametrically the point-density at zero incomes, even though these incomes are sometimes more prevalent than negative incomes. Zero incomes are thus left uncorrected. Moreover, the corrected incomes retain their unrealistic negative sign, so the approach can be said, at best, to provide a cosmetic correction for the problem of extremely low incomes. Finally worth noting, because this correction replaces incomes below a poverty threshold with other values below the unchanged threshold (which is based on median income), the poverty headcount ratio is unaffected.

Imputation of negative and zero incomes using random forest classification among positive incomes appears to be a viable approach for dealing with non-positive incomes, as it produces a continuous distribution of overall incomes without a point-density at zero, and converts non-positive incomes into realistic positive values based on households' observed characteristics. This imputation has shown sensible results across multiple countries and across two model specifications that were tried, and it lowers the estimated Gini by up to 1.7 percentage points.

These preliminary estimations, conducted under rather conservative assumptions and modeling specifications, suggest that the poverty-identification and inequality-measurement problems posed by negative and especially zero incomes are not trivial, and deserve attention and careful modeling by academics and practitioners. In relation to the 'static' problem of non-positive incomes, our corrections produce more accurate inequality and poverty indexes for the majority of countries. However, in relation to the 'dynamic' problem of non-positive incomes for measuring the evolution of inequality and poverty, we find only limited evidence that our corrections reduce the volatility of inequality and poverty indexes across survey waves, as would be desired from a correction method.

Where do we go from here? Going beyond Pareto distributions, which do not fit too well, should allow us to model negative as well as zero incomes more sensibly. Efficiency improvements could also be made to the random forest classification method, since we have limited ourselves to evaluating only a simple robust specification. More importantly, extending the analysis to a greater range of bottom incomes - say the extreme 5-10 percent as the top-income literature has been doing, or all incomes falling short of households' consumption - promises to yield more determinate corrections. We should find more clearly that the corrections provide a dynamic benefit in the form of reduced volatility of inequality and poverty indexes. With the corrected bottom incomes, we should be able to re-evaluate their impact on multidimensional deprivation and poverty, and the true incidence of development.

The policy implications of this ongoing research are clear. Our results are relevant for the assessments of poverty depth, fiscal redistribution, aid targeting, and in the MENA region the tackling of evasion and the use of natural resource revenues. Since uprisings in the MENA region have been linked to the problems of poverty and unequal economic opportunities, a better understanding of the scale of these problems can give policymakers the tools to bring social discontent down, and even fix some traps and obstacles to economic growth.

References

Athey, S. and G. Imbens (2019). Machine learning methods economists should know about. *arXiv* (1903.10075).

Atkinson, A. B., T. Piketty, and E. Saez (2011, March). Top incomes in the long run of history. *Journal of Economic Literature 49*(1), 3–71.

Blundell, R. and M. Costa Dias (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources 44*(3), 565–640.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Ceriani, L. and P. Verme (2019, January). The inequality of extreme incomes. *ECINEQ Working Paper* (490).

Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics 141*, 1044–72.

Cowell, F. A. and P. Van Kerm (2015). Wealth inequality: a survey. *Journal of Economic Surveys 29*(4), 671–710.

Cowell, F. A. and M.-P. Victoria-Feser (1996, December). Poverty measurement with contaminated data: A robust approach. *European Economic Review 40*(9), 1761–1771.

Dagum, C. (1990). A model of net wealth distribution specified for negative, null and positive wealth, a case study: Italy. In C. Dagum and M. Zenga (Eds.), *Income and Wealth Distribution, Inequality and Poverty*, pp. 42–56. Berlin: Springer.

Dagum, C. (1999). A study on the distributions of income, wealth and human capital. *Revue Europeenne des Sciences Sociales 37*(133), 231–268.

Davis, H. T. (1941). *The theory of econometrics.* Bloomington, Principia Press.

Eurostat (2005). Current treatment of taxes and their implication on negative income and on comparability between countries. EU-SILC Documents TFMC-1, European Commission.

Eurostat (2006a). Self-employment income. EU-SILC Documents TFMC-02/06, European Commission.

Eurostat (2006b). Some proposals on the treatment of negative incomes. EU-SILC Documents TFMC-15/06, European Commission.

Fitzpatrick, C. A., P. Bull, and O. Dupriez (2018). Machine learning for poverty prediction: A comparative assessment of classification algorithms. *Wired at: www.github.com.*

Haziza, D. and J.-F. Beaumont (2007). On the construction of imputation classes in surveys. *International Statistical Review 75*, 25–43.

Hlasny, V. and P. Verme (2018). Top incomes and the measurement of inequality in Egypt. *World Bank Economic Review 32*(2), 428–455.

James, F., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. *Econometrica 52*(3), 761–766.

James, F., J. Greer, and E. Thorbecke (2010). The Foster–Greer–Thorbecke (FGT) poverty measures: 25 years later. *Journal of Economic Inequality 8*(4), 491–524.

Jäntti, M., E. Sierminska, and P. V. Kerm (2015). Modelling the joint distribution of income and wealth. Discussion Paper 9190, IZA.

Jenkins, S. P., R. V. Burkhauser, S. Feng, and J. Larrimore (2011, January). Measuring inequality using censored data: a multiple-imputation approach to estimation and inference. *Journal of the Royal Statistical Society Series A 174*(1), 63–81.

Jenkins, S. P. and M. Jäntti (2005). Methods for summarizing and comparing wealth distributions. Working Paper 2005-05, ISER.

Kibekbaev, A. and E. Duman (2016). Benchmarking regression algorithms for income prediction modeling. *Information Systems 61*, 40–52.

LIS (2019). LIS Database http://www.lisdatacenter.org (multiple countries; February-November 2019). Luxembourg: LIS, Luxembourg Income Study.

Luchman, J. N. (2015). Random forest ensemble classification based on chi-square automated interaction detection (chaid) as base learner. Statistical Software Components S457932, Boston College Department of Economics.

Nussbaum, M. (2011). *Creating capabilities: the human development approach.* Cambridge, MA: Harvard University Press.

Ostasiewicz, K. and A. Vernizzi (2017). Decomposition and normalization of absolute differences, when positive and negative values are considered: applications to the gini coefficient. *Quantitative Methods in Economics 18*(3), 472–491.

Scott, C. D. and J. A. Litchfield (1994). Inequality, mobility and the determinants of income among the rural poor in chile, 1968-1986. STICERD Discussion Paper 53, London School of Economics.

Sen, A. (2000). *Development as freedom.* New York: Anchor Books.

Stich, A. (1996). Inequality and negative income. *Journal of the Italian Statistical Society 5*(3), 297–305.

Van Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. Working Paper 2007-01, CEPS-Instead IRISS.

Zabala, F. (2015). Let the data speak: Machine learning methods for data editing and imputation. Working Paper 31, Conference of European Statisticians, United Nations Economic Commission for Europe.

Zhao, P., X. Su, T. Ge, and J. Fan (2017). Propensity score and proximity matching using random forest. *Contemporary Clinical Trials 47*, 85–92.

Table 1: Components in negative incomes, household surveys included in this study

Country	HH	Zero DHI		Mean neg. DHI /Mean pos. DHI		Mean neg.HILS/mean neg. DHI among neg. DHIs		Mean neg. (DHI-HILS-HICID-HITP)/mean neg.DHI among neg. DHIs		Mean HXITS/mean (HI-HILS-HICID-HITP) among neg. DHIs		Source
		Num	Share (%)	Num	Share (%)	Num	Share (%)	Num	Share (%)	Num	Share (%)	
EG99	23975	0	0.000	0	.	0	.	0	.	0	.	HIECS
EG04	47095	0	0.000	0	.	0	.	0	.	0	.	HIECS
EG08	23428	0	0.000	0	.	0	.	0	.	0	.	HIECS
EG10	7719	0	0.000	0	.	0	.	0	.	0	.	HIECS
EG12^{a}	7528	0	0.000	0	.	.		.		.		HIECS
EG12^{a}	12039	173	1.437	28	12.60	28	113.990	0	.	0	.	LMPS
EG15	11988	0	0.000	0	.	0	.	0	.	0	.	HIECS
FR00^{L}	10305	4	0.039	14	58.32	0	.	14	141.89	0	.	BDF
FR05^{L}	10240	0	0.000	3	57.86	0	.	3	141.73	0	.	BDF
FR10^{L}	15797	117	0.741	25	29.73	16	194.200	10	155.08	0	.	BDF
GR95^{L}	4842	50	1.033	17	0.28	17	100.000	0	.	0	.	GRECHP
GR00^{L}	3895	18	0.462	4	1.05	3	18.880	4	84.18	0	.	GRECHP
GR04^{L}	5568	21	0.377	18	27.03	16	151.820	7	16.77	0	.	EU-SILC
GR07^{L}	6503	26	0.400	29	100.96	18	71.470	25	59.70	0	.	EU-SILC
GR10^{L}	6024	30	0.498	23	3.82	2	50.190	23	233.00	0	.	EU-SILC
GR13^{L}	8616	6	0.070	8	6.50	0	.	8	100.00	0	.	EU-SILC
IQ07	17822	28	0.157	12	25.53	12	166.430	0	.	0	.	HIES
IQ12	25146	12	0.048	0	.	0	.	0	.	0	.	HIES
IL97^{L}	5230	0	0.000	0	.	0	.	0	.	0	.	HES
IL01^{L}	5787	0	0.000	19	32.89	6	374.420	14	3.93	19	61	HES
IL05^{L}	6272	0	0.000	17	18.92	7	473.440	13	6.38	17	24	HES
IL07^{L}	6172	0	0.000	18	1.95	1	397.230	17	79.99	18	469	HES
IL10^{L}	6168	0	0.000	10	1.27	0	.	10	105.00	10	672	HES
IL12^{L}	8742	0	0.000	45	16.71	3	2284.580	42	6.68	45	17	HES
IL14^{L}	8465	0	0.000	35	103.84	6	39.970	31	129.54	35	39	HES
IL16^{L}	8903	0	0.000	31	25.53	4	681.880	27	65.92	31	14	HES
IT95^{L}	8134	16	0.197	14	48.26	14	124.910	0	.	0	.	SHIW
IT98^{L}	7147	61	0.854	7	70.05	7	165.440	0	.	0	.	SHIW
IT00^{L}	8000	75	0.938	2	13.06	2	113.210	0	.	0	.	SHIW
IT04^{L}	8012	16	0.200	4	71.07	4	134.380	1	0.00	4	22	SHIW
IT08^{L}	7977	39	0.489	0	.	0	.	0	.	0	.	SHIW
IT10^{L}	7941	47	0.592	1	3.77	1	100.320	0	.	1	152.75b	SHIW
IT14^{L}	8151	122	1.497	2	9.76	2	406.160	1	148.14	0	.	SHIW
JO0	2518	0	0.000	0	.	0	.	0	.	0	.	HEIS
JO06	2897	0	0.000	0	.	0	.	0	.	0	.	HEIS
JO08	2746	0	0.000	0	.	0	.	0	.	0	.	HEIS
JO10	2845	0	0.000	0	.	0	.	0	.	0	.	HEIS
JO13	4850	0	0.000	0	.	0	.	0	.	0	.	HEIS
PS10	3757	3	0.080	0	.	0	.	0	.	0	.	PECS
PS11	4317	8	0.185	0	.	0	.	0	.	0	.	PECS
RS06^{L}	4560	33	0.724	46	51.51	46	180.730	0	.	0	.	HBS
RS10^{L}	4585	19	0.414	26	67.15	26	151.170	0	.	0	.	HBS
RS13^{L}	4517	35	0.775	47	86.31	47	150.340	0	.	0	.	HBS
RS16^{L}	6448	48	0.744	38	79.17	38	145.710	0	.	0	.	HBS
SI97^{L}	2577	0	0.000	0	.	0	.	0	.	0	.	HBS
SI99^{L}	3859	0	0.000	0	.	0	.	0	.	0	.	HBS
SI04^{L}	3725	1	0.027	0	.	0	.	0	.	0	.	HBS
SI07^{L}	3697	0	0.000	0	.	0	.	0	.	0	.	HBS
SI10^{L}	3924	1	0.025	0	.	0	.	0	.	0	.	HBS
SI12^{L}	3663	0	0.000	0	.	0	.	0	.	0	.	HBS
ES95^{L}	5928	29	0.489	38	3.08	38	128.580	2	12.00	0	.	ECV
ES00^{L}	4776	4	0.084	11	3.33	9	85.680	0	.	0	.	ECV
ES04^{L}	12950	112	0.865	3	3.49	0	.	0	.	0	.	EU-SILC
ES07^{L}	13014	30	0.231	44	84.74	30	189.310	19	19.49	0	.	EU-SILC
ES10^{L}	13109	93	0.709	107	29.52	104	128.650	66	29.47	0	.	EU-SILC
ES13^{L}	11965	44	0.368	53	21.17	39	148.260	27	87.88	0	.	EU-SILC
SD09	7913	28	0.354	0	.	0	.	0	.	0	.	NBHS

Notes to Table 1: Years refer to income-reference years. Surveys were harmonized by LIS and ERF. Observation counts are those with disposable household income non-missing. No income variables available for Tunisia, Somalia, and 1996–2009 Palestine. BDF - Budget de Famille; ECV - Encuesta de Condiciones de Vida; EU SILC - EU Statistics on Income & Living Conditions; GR ECHP – Greek Household Income & Living Conditions Survey; HBS - Household Budget Survey; HEIS - Household Expenditure & Income Survey; HES - Household Expenditure Survey; HIECS - Household Income, Expenditure & Consumption Survey; HIES - Household Income & Expenditure Surveys; LMPS - Labor Market Panel Survey; NBHS - National Baseline Household Survey; PECS - Palestinian Expenditure & Consumption Survey; SHIW - Indagine sui Bilanci delle Famiglie (Survey of Household Income and Wealth). L: Survey is from the LIS database, else from the ERF database. a: For Egypt 2012, ERF database includes data from the Household Income, Expenditure and Consumption Survey (HIECS), while LIS database includes data from the Labor Market Panel Survey (LMPS). We report figures for HIECS and LMPS. b: In IT10, the single household with $DHI < 0$ has $(HI - HILS - HICID - HITP) = 0$; therefore $HI = 1,890$ is used

Table 2: Characteristics of households with negative or zero incomes

	Attributes of hhds with neg. DHI as % of nationwide mean[b]							Attributes of hhds with zero DHI as % of nationwide mean[b]						
	Consump.	Food expend.	Outflow from mortg., loans & repaymts.	Home-ownership	Good health	Upper secondary education	Urban	Consump.	Food expend.	Outflow from mortg., loans & repaymts.	Home-ownership	Good health	Upper secondary education	Urban
EG99	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EG04	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EG08	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EG10	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EG12[a]	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EG12[a]	.	.	.	136.8	112.2	61.7	18.1	.	.	.	101	104.5	105.6	76.5
EG15	.	.	.	.	.	.	.	.	.	.	.	.	.	.
FR00	177.1	127.4	72.5	138.7	.	119.8	102.8	49.8	43.8	0	49.3	.	48.8	133.9
FR05	159	150.8	171.9	45.7	.	174	97	.	.	.	.	.	.	.
FR10	103.8	122.8	162.4	98.8	.	123.7	106.1	58.8	36.6	32.5	75.4	.	56.4	127.8
GR95	.	.	.	95.1	.	24.7	.	.	.	.	81.1	.	127.9	.
GR00	.	.	0	118.2	.	193.3	72.4	.	.	14.3	60.9	.	161.2	80.8
GR04	.	.	.	96.8	.	97.4	118.7	.	.	.	113.1	.	138.1	131.3
GR07	.	.	.	77.2	96.4	108.3	108.6	.	.	.	43.6	108.8	130.4	144.8
GR10	.	.	84	107.9	110.4	147.1	103.4	.	.	67.5	81.2	118.3	42.4	127.5
GR13	.	.	.	106.7	120.1	114.8	117.1	.	.	.	66.9	101.6	119.2	159.5
IQ07	151.1	124.1	.	104.4	.	27.6	1.8	38.6	41.1	.	0	.	199.4	65.7
IQ12	.	.	.	.	.	.	.	78.4	63.3	.	76.9	.	65.7	67.4
IL97	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IL01	75.4	62.5	43.1	92.3	.	94.2	95.6	.	.	.	.	.	.	.
IL05	64.3	78.3	47.8	127.2	.	101.8	.	.	.	.	.	.	.	.
IL07	61.9	60.2	15.1	83.6	.	63.9	97.3	.	.	.	.	.	.	.
IL10	55.1	57.8	73.8	78.2	.	117.6	101.2	.	.	.	.	.	.	.
IL12	65.6	52.6	.	49.4	.	109.2	96.8	.	.	.	.	.	.	.
IL14	158.2	94.1	.	80.5	.	109.6	96.2	.	.	.	.	.	.	.
IL16	48.2	40.3	.	75.5	.	117.5	98.1	.	.	.	.	.	.	.
IT95	105	115.3	.	139.8	124.1	152.7	120.5	58	59.1	.	53.1	61.7	60.5	79.8
IT98	94.6	107.1	27.4	145.9	.	222.7	99.3	55.5	60.4	47.4	47.3	.	16	117.2
IT00	111.8	88.3	1,505.90	80.9	.	0	123.6	57.8	62.3	12.8	74.6	.	22.7	110.7
IT04	128.9	115.4	0	82.6	.	0	130.1	81.6	97.4	0	128.8	.	85	123
IT08	.	.	.	.	.	.	.	54.7	74.7	5.5	65.9	98.8	39.7	113.1
IT10	.	87.3	0	0	133.4	0	126.6	54.2	61.8	13.3	75.5	110.5	93.6	106.9
IT14	67.6	71.2	.	146.5	.	251.5	125.1	53.2	56.9	.	47.5	.	58.1	106.5
JO02	.	.	.	.	.	.	.	.	.	.	.	.	.	.
JO06	.	.	.	.	.	.	.	.	.	.	.	.	.	.
JO08	.	.	.	.	.	.	.	.	.	.	.	.	.	.
JO10	.	.	.	.	.	.	.	.	.	.	.	.	.	.
JO13	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PS10	.	.	.	.	.	.	.	32.8	26.8	.	118.8	.	0	120.8
PS11	.	.	.	.	.	.	.	55.5	44.6	.	96.4	.	108.3	104.7
RS06	128.1	105.4	.	102.7	.	31	33.3	67.7	65.6	.	88.1	.	124.6	132.2
RS10	131.7	108.4	.	105.6	.	33.2	22.5	63.2	62.6	.	68.2	.	136.2	138.6
RS13	126.6	120.3	.	102.3	.	8.8	11	55.3	60.9	.	91.5	.	119	119.9
RS16	129.1	124.4	.	114	.	51	19.3	61.4	63.5	.	77.4	.	101.6	118.3
SI97	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SI99	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SI04	.	.	.	.	.	.	.	27.6	73.2	.	0	.	137.1	.
SI07	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SI10	.	.	.	.	.	.	.	27.1	62.1	.	130.7	0	128.5	.
SI12	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ES95	.	.	.	81.6	.	107.8	.	.	.	.	80	.	81.1	.
ES00	.	.	0	106.8	.	76.8	.	.	.	0	0	.	0	.
ES04	.	.	.	95.7	.	0	25.6	.	.	.	76.7	.	83.6	102.8
ES07	.	.	.	106.2	100.4	132.1	93.1	.	.	.	70.7	82.2	100.7	110.1
ES10	.	.	.	111.7	100.1	70.7	84.3	.	.	.	68.9	91.4	99.3	109.3
ES13	.	.	.	91.3	115.1	89.4	95.5	.	.	.	49	100.8	68	113
SD09	.	.	.	.	.	.	.	86.1	110.4	.	109.2	.	13.5	51.4

Notes: Years refer to income-reference years. Surveys were harmonized by LIS and ERF. Observation counts are those with disposable household income non-missing. Samples weighted using household weights.

a: For Egypt 2012, ERF database includes data from the Household Income, Expenditure and Consumption Survey (HIECS), while LIS database includes data from the Labor Market Panel Survey (LMPS). We report figures for HIECS (first row) and LMPS (second row).

b: Nationwide mean computed only among non-negative values.

Table 3: Gini coefficients and poverty headcount ratios

			Inequality				Poverty			
	$DHI <$ 0 (#)	$DHI =$ 0 (#)	Gini	Gini ($DHI \geq 0$, bottom-code at 0)	Gini ($DHI \geq 0$, truncate $DHI < 0$)	Gini ($DHI > 0$)	Poverty HCR (%)	HCR ($DHI \geq 0$, bottom-code at 0)	HCR ($DHI \geq 0$, truncate $DHI < 0$)	HCR ($DHI > 0$)
EG99	0	0	35.04 (0.44)	–	–	–	4.48 (0.13)	–	–	–
EG04	0	0	34.55 (0.24)	–	–	–	6 (0.11)	–	–	–
EG08	0	0	33.66 (0.41)	–	–	–	5.35 (0.15)	–	–	–
EG10	0	0	32.43 (0.47)	–	–	–	5.49 (0.26)	–	–	–
EG12[a]	0	0	31.33 (0.48)	–	–	–	4.98 (0.25)	–	–	–
EG12[a]	28	173	53.17 (1.17)	53.12 (1.18)	53 (1.18)	52.32 (1.19)	18.59 (0.35)	18.59 (0.35)	18.4 (0.35)	17.37 (0.35)
EG15	0	0	35.2 (1.48)	–	–	–	5.05 (0.2)	–	–	–
FR00	14	4	33.07 (0.28)	32.97 (0.27)	32.89 (0.27)	32.86 (0.27)	9.01 (0.28)	9.01 (0.28)	8.93 (0.28)	8.9 (0.28)
FR05	3	0	33.04 (0.29)	33.01 (0.29)	32.99 (0.29)		9.68 (0.29)	9.68 (0.29)	9.65 (0.29)	–
FR10	25	117	34.32 (0.38)	34.24 (0.38)	34.1 (0.38)	34.04 (0.38)	9.99 (0.24)	9.99 (0.24)	9.8 (0.24)	9.72 (0.24)
GR95	17	50	40.42 (0.54)	40.42 (0.54)	40.23 (0.54)	39.61 (0.53)	18.56 (0.56)	18.56 (0.56)	18.36 (0.56)	17.7 (0.55)
GR00	4	18	39.13 (0.55)	39.13 (0.55)	39.03 (0.55)	38.71 (0.54)	17.59 (0.61)	17.59 (0.61)	17.61 (0.61)	17.34 (0.61)
GR04	18	21	37.67 (0.48)	37.55 (0.47)	37.34 (0.47)	37.07 (0.47)	14.09 (0.47)	14.09 (0.47)	13.93 (0.46)	13.72 (0.46)
GR07	29	26	37.21 (0.65)	36.44 (0.51)	36.08 (0.5)	35.81 (0.5)	12.83 (0.41)	12.83 (0.41)	12.36 (0.41)	12.09 (0.41)
GR10	23	30	36.76 (0.54)	36.74 (0.54)	36.51 (0.54)	36.2 (0.54)	14.34 (0.45)	14.34 (0.45)	14.07 (0.45)	13.8 (0.45)
GR13	8	6	36.86 (0.54)	36.86 (0.54)	36.81 (0.54)	36.76 (0.54)	14.13 (0.38)	14.13 (0.38)	14.07 (0.37)	14.01 (0.37)
IQ07	28	12	42.25 (0.63)	42.24 (0.63)	42.22 (0.63)	42.12 (0.63)	13.61 (0.26)	13.61 (0.26)	13.58 (0.26)	13.44 (0.26)
IQ12	0	12	41.29 (0.92)	–	–	41.28 (0.92)	16.69 (0.24)	–	–	16.69 (0.24)
IL97	0	0	37.99 (0.51)	–	–	–	16.32 (0.51)	–	–	–
IL01	19	0	38.79 (0.51)	38.61 (0.5)	38.38 (0.49)	–	16.81 (0.49)	16.81 (0.49)	16.51 (0.49)	–
IL05	17	0	39.63 (0.63)	39.56 (0.63)	39.38 (0.63)	–	18.46 (0.49)	18.46 (0.49)	18.24 (0.49)	–
IL07	18	0	39.34 (0.38)	39.33 (0.38)	39.13 (0.37)	–	17.94 (0.49)	17.94 (0.49)	17.75 (0.49)	–
IL10	10	0	41.04 (0.69)	41.03 (0.69)	40.91 (0.69)	–	19.31 (0.5)	19.31 (0.5)	19.14 (0.5)	–
IL12	45	0	39.41 (0.38)	39.28 (0.36)	38.95 (0.36)	–	17.63 (0.41)	17.63 (0.41)	17.28 (0.41)	–
IL14	35	0	39.5 (0.42)	38.89 (0.33)	38.63 (0.33)	–	18.86 (0.43)	18.86 (0.43)	18.65 (0.42)	–
IL16	31	0	37.76 (0.33)	37.58 (0.3)	37.27 (0.29)	–	18.22 (0.41)	18.22 (0.41)	17.9 (0.41)	–
IT95	14	16	37.43 (0.48)	37.3 (0.47)	37.18 (0.47)	37.06 (0.47)	14.86 (0.39)	14.86 (0.39)	14.72 (0.39)	14.62 (0.39)
IT98	7	61	38.71 (0.62)	38.6 (0.62)	38.53 (0.61)	38.12 (0.62)	15.39 (0.43)	15.39 (0.43)	15.3 (0.43)	14.97 (0.42)
IT00	2	75	37.08 (0.47)	37.08 (0.47)	37.07 (0.47)	36.57 (0.47)	13.35 (0.38)	13.35 (0.38)	13.34 (0.38)	12.72 (0.37)
IT04	4	16	36.67 (0.58)	36.64 (0.58)	36.62 (0.58)	36.5 (0.58)	11.35 (0.35)	11.35 (0.35)	11.39 (0.35)	11.24 (0.35)
IT08	0	39	36.14 (0.54)	–	–	35.78 (0.54)	11.7 (0.36)	–	–	11.33 (0.36)
IT10	1	47	35.43 (0.46)	35.43 (0.46)	35.41 (0.46)	35.04 (0.46)	11.37 (0.36)	11.37 (0.36)	11.34 (0.36)	11.05 (0.35)
IT14	2	122	36.44 (0.48)	36.43 (0.48)	36.41 (0.48)	35.27 (0.47)	12.82 (0.37)	12.82 (0.37)	12.79 (0.37)	11.35 (0.35)
JO02	0	0	40.28 (1.28)	–	–	–	14.35 (0.7)	–	–	–
JO06	0	0	40.32 (1.29)	–	–	–	11.64 (0.6)	–	–	–
JO08	0	0	40.33 (1.66)	–	–	–	11.26 (0.6)	–	–	–
JO10	0	0	41.01 (1.88)	–	–	–	11.57 (0.6)	–	–	–
JO13	0	0	37.82 (1.03)	–	–	–	12.33 (0.47)	–	–	–
PS10	0	3	42.52 (0.69)	–	–	42.48 (0.69)	18.85 (0.64)	–	–	18.79 (0.64)
PS11	0	8	41.23 (0.6)	–	–	41.11 (0.6)	19.13 (0.6)	–	–	18.98 (0.6)
RS06	46	33	40.26 (0.51)	39.52 (0.45)	38.9 (0.44)	38.46 (0.44)	18.08 (0.57)	18.08 (0.57)	17.42 (0.56)	17.19 (0.56)
RS10	26	19	38.53 (0.5)	37.97 (0.45)	37.59 (0.45)	37.29 (0.44)	15.6 (0.54)	15.6 (0.54)	15.13 (0.53)	14.78 (0.53)
RS13	47	35	40.71 (0.76)	39.48 (0.61)	38.88 (0.61)	38.37 (0.61)	15.89 (0.54)	15.89 (0.54)	15.59 (0.54)	14.99 (0.54)
RS16	38	48	39.65 (0.46)	39.09 (0.4)	38.78 (0.39)	38.32 (0.39)	16.7 (0.46)	16.7 (0.46)	16.45 (0.46)	16.07 (0.46)
SI97	0	0	30.4 (0.49)	–	–	–	10.07 (0.59)	–	–	–
SI99	0	0	30.86 (0.43)	–	–	–	11.26 (0.51)	–	–	–
SI04	0	1	31.74 (0.43)	–	–	31.7 (0.43)	12.01 (0.53)	–	–	11.98 (0.53)
SI07	0	0	31.87 (0.41)	–	–	–	12.53 (0.54)	–	–	–
SI10	0	1	34.37 (0.45)	–	–	34.31 (0.45)	14.99 (0.57)	–	–	14.92 (0.57)
SI12	0	0	35.65 (0.48)	–	–	–	13.91 (0.57)	–	–	–
ES95	38	29	39.5 (0.46)	39.47 (0.46)	39.06 (0.46)	38.73 (0.46)	14.33 (0.46)	14.33 (0.46)	13.83 (0.45)	13.48 (0.45)
ES00	11	4	38.84 (0.54)	38.83 (0.54)	38.7 (0.54)	38.64 (0.54)	17.66 (0.55)	17.66 (0.55)	17.5 (0.55)	17.41 (0.55)
ES04	3	112	36.42 (0.29)	36.42 (0.29)	36.41 (0.29)	35.91 (0.28)	16.6 (0.33)	16.6 (0.33)	16.58 (0.33)	16.15 (0.32)
ES07	44	30	35.34 (0.32)	34.95 (0.29)	34.72 (0.29)	34.6 (0.28)	15.63 (0.32)	15.63 (0.32)	15.44 (0.32)	15.3 (0.32)
ES10	107	93	37.53 (0.29)	37.18 (0.28)	36.65 (0.27)	36.13 (0.27)	15.72 (0.32)	15.72 (0.32)	15.24 (0.32)	14.64 (0.31)
ES13	53	44	38.03 (0.32)	37.87 (0.32)	37.53 (0.31)	37.21 (0.31)	15.2 (0.33)	15.2 (0.33)	14.84 (0.33)	14.44 (0.32)
SD09	0	28	54.48 (1.06)	–	–	54.37 (1.06)	21.66 (0.46)	–	–	21.6 (0.46)

Notes: Years refer to income-reference years. Surveys were harmonized by LIS and ERF. Standard errors in parentheses. '–' For clarity of presentation: Because of the absence of zero/negative incomes, the statistics are same as in the preceding column, and are thus omitted.

a: For Egypt 2012, ERF database includes data from the Household Income, Expenditure and Consumption Survey (HIECS), while LIS database includes data from the Labor Market Panel Survey (LMPS). We report figures for HIECS (first row) and LMPS (second row).

Table 4: Estimates of Pareto, and generalized Pareto distributions among negative incomes

			Pareto 1						Pareto 2					
	Actual Neg-income Gini	Actual Pos-income Gini	Pareto (I) coeff. α	Pop. Share (%)	Estim. income share	Estim. neg-income Gini	Mean neg. income	Semipar. (Pareto I) Gini	Pareto (II) coeff. $\log(\sigma)$	χ	Estim. income share	Estim. neg-income Gini	Mean neg. income	Semipar. (Pareto I) Gini
EG12a	60.44	53	0.321 (0.048)	0.26	–	–	–	–	7.46 (0.371)	0.516 (0.33)	0.04	34.76	-3,589	53.18
FR00	64.63	32.89	0.165 (0.019)	0.12	–	–	–	–	7.813 (0.738)	1.656 (0.797)	–	–	–	–
FR05b	30.37	32.99	0.93 (0.397)	0.03	–	–	–	–	10.492 (–)	-1.222 (–)	0.02	–	-16,209	33.04
FR10	61.43	34.1	0.345 (0.038)	0.2	–	–	–	–	8.626 (0.373)	0.576 (0.329)	0.08	40.44	-13,151	34.34
GR95	0	40.23	1.443 (–)	0.32	-0.001	0.53	-48	40.42	–	–	–	–	–	–
GR00	61.23	39.03	1.273 (1.06)	0.16	-0.001	0.647	-137	39.13	3.827 (0.889)	1.293 (0.834)	0	–	–	–
GR04	54.87	37.34	0.323 (0.038)	0.32	–	–	–	–	8.357 (0.47)	0.303 (0.408)	0.1	17.87	-6,117	37.68
GR07	64.71	36.08	0.16 (0.008)	0.56	–	–	–	–	9.351 (0.321)	0.598 (0.28)	0.7	42.69	-28,656	37.38
GR10	40.56	36.51	0.405 (0.045)	0.36	–	–	–	–	7.102 (0.301)	-0.362 (0.23)	0.01	–	-893	–
GR13	56.45	36.81	0.924 (0.424)	0.08	–	–	–	–	6.304 (0.646)	0.729 (0.565)	0.01	57.29	-2,016	36.87
IQ07	69.47	42.22	0.373 (0.067)	0.03	–	–	–	–	7.034 (0.431)	0.57 (0.339)	0.01	39.83	-2,639	42.25
IL01	81.56	38.38	0.736 (0.294)	0.38	–	–	–	–	7.567 (0.444)	1.804 (0.544)	–	–	–	–
IL05	80.59	39.38	0.346 (0.096)	0.29	–	–	–	–	7.288 (0.578)	1.836 (0.669)	0	–	–	–
IL07	33.07	39.13	1.536 (0.333)	0.32	-0.008	0.483	-3,163	39.34	7.8 (0.312)	0.097 (0.208)	0.01	5.07	-2,702	39.34
IL10b	20.67	40.91	0.452 (0.046)	0.2	–	–	–	–	8.329 (–)	-1.151 (–)	0	–	-1,927	–
IL12	90.13	38.95	1.5 (0.436)	0.55	-0.013	0.5	-3,926	39.3	7.649 (0.211)	0.836 (0.19)	0.04	71.84	-12,804	39.34
IL14	75.1	38.63	0.361 (0.068)	0.42	–	–	–	–	8.293 (0.378)	2.599 (0.546)	–	–	–	–
IL16	85.72	37.27	0.338 (0.052)	0.5	–	–	–	–	7.767 (0.327)	1.543 (0.38)	–	–	–	–
IT95	53.53	37.18	0.509 (0.069)	0.19	–	–	–	–	8.756 (0.407)	0.396 (0.313)	0.11	24.71	-10,516	37.45
IT98	51.9	38.53	0.769 (0.263)	0.11	–	–	–	–	9.185 (0.588)	0.542 (0.457)	0.11	37.16	-21,280	38.76
IT00b	3.58	37.07	12.22 (10.8)	0.01	-0.001	0.043	-2,886	37.08	8.428 (–)	-1.49 (–)	0	–	-1,837	–
IT04	32.83	36.62	1.134 (0.601)	0.03	-0.063	0.788	-48,048	36.72	10.259 (0.66)	-0.649 (0.541)	0.02	–	-17,301	–
IT14b	12.26	36.41	2.843 (2.17)	0.04	-0.004	0.213	-2,609	36.44	8.342 (–)	-1.405 (–)	0	–	-1,746	–
RS06	63.5	38.9	0.216 (0.012)	1.01	–	–	–	–	11.497 (0.28)	0.576 (0.25)	0.65	40.46	-232,222	40.42
RS10	60.46	37.59	0.194 (0.015)	0.6	–	–	–	–	11.963 (0.504)	0.802 (0.485)	0.92	66.9	-790,927	39.23
RS13	62.73	38.88	0.162 (0.0084)	0.99	–	–	–	–	12.614 (0.266)	0.52 (0.234)	1.01	35.14	-626,768	40.88
RS16	74.03	38.78	0.284 (0.024)	0.5	–	–	–	–	11.892 (0.362)	0.993 (0.357)	18.52	98.68	-22000000	64.84
ES95b	14.25	39.06	0.51 (0.02)	0.66	–	–	–	–	6.838 (–)	-1.413 (–)	0.02	–	-388	–
ES00	46.67	38.7	0.84 (0.185)	0.2	–	–	–	–	6.155 (0.386)	0.429 (0.249)	0.01	27.28	-826	38.84
ES04	36.15	36.41	2.534 (2.58)	0.02	0	0.246	-680	36.42	5.421 (0.963)	1.581 (0.778)	0	–	–	–
ES07	55.08	34.72	0.202 (0.013)	0.34	–	–	–	–	9.811 (0.298)	0.235 (0.258)	0.31	13.32	-23,836	35.36
ES10	48.76	36.65	0.246 (0.007)	0.84	–	–	–	–	8.905 (0.135)	0.045 (0.094)	0.25	2.31	-7,721	37.53
ES13	52.78	37.53	0.202 (0.01)	0.54	–	–	–	–	8.589 (0.233)	0.077 (0.188)	0.12	4.02	-5,823	38.04

a. For Egypt 2012, results for the LMPS survey (LIS database) are provided.

b. For negative incomes in 5 surveys (FR05, IL10, IT00, IT14, ES95), estimation of the generalized Pareto (II) model did not achieve convergence. All estimates should be viewed with caution as they may not be maximum-likelihood estimates; standard errors for them are not reported.

c. For negative incomes in 25 surveys, estimation of the Pareto (I) model yielded α coefficients < 1, implying wide dispersion with an undefined parametric mean. For these surveys, we omit parametric estimates of the income share and the Gini, as they would be outside of reasonable bounds.

d. For negative incomes in 15 surveys, estimation of the generalized Pareto (II) model yielded implausible pairs of coefficients σ, χ giving positive income shares and Ginis outside of the unit interval. For these surveys, we omit parametric estimates of the income share and the Gini, as they would be outside of reasonable bounds.

Table 5: Random forest imputation of negative incomes: Gini and poverty HCR estimates

		(1) Random forest classification & imputation of $HILS < 0$				(2) Random forest classification & imputation of $DHI - HILS - HICID - HITP \leq 0$				(3) Random forest classification & imputation of $DHI \leq 0$		
	$DHI \leq 0$ (#)	$DHI \leq 0$ corrected	corrected to $DHI > 0$	Corrected Gini	Corrected P0	$DHI \leq 0$ corrected	corrected to $DHI > 0$	Corrected Gini	Corrected P0	$DHI \leq 0$ corrected to $DHI > 0$	Corrected Gini	Corrected P0
EG12[a]	201	29	29	52.92	18.45***	201	201	51.87	17.99***	201	51.82	18.01***
FR00	18	–		–	–	18	18	32.84	8.89	18	32.85	8.86
FR05	3	–		–	–	3	3	32.99	9.65	3	32.99	9.65
FR10	142	16	16	34.17	9.86	128	140	34.19	9.84	142	34.06	9.79
GR95	67	17	17	40.34	18.48	–		–	–	67	40.27	17.77
GR00	22	3	4	39.06	17.76	22	22	38.76	17.36	22	38.55	17.36
GR04	39	16	16	37.33	14.08	36	39	37.08	13.81	39	36.9	13.71
GR07	55	16	19	36.34	12.52	49	55	35.87**	12.01**	55	35.61**	12.01**
GR10	53	2	3	36.74	14.31	53	53	36.02	13.84	53	36.02	13.68
GR13	14	0	1	36.86	14.13	14	14	36.73	13.98	14	36.73	13.98
IQ07	40	12	12	42.21	13.58	–		–	–	40	42.1	13.41
IQ12	12	–		–	–	–		–	–	12	41.28	16.68
IL01	19	6	6	38.55	16.76	14	15	38.67	16.6	19	38.46	16.6
IL05	17	7	7	39.51	18.37	12	16	39.56	18.4	17	39.45	18.39
IL07	18	1	1	39.33	17.94	17	17	39.23	17.79	18	39.22	17.79
IL10	10	–		–	–	10	10	40.97	19.31	10	40.97	19.31
IL12	45	3	3	39.28	17.6	42	44	39.32	17.63	45	39.14	17.63
IL14	35	6	14	39.45	18.82	31	34	38.83	18.83	35	38.79*	18.86
IL16	31	4	4	37.56	18.18	27	29	37.63	18.22	31	37.46	18.22
IT95	30	14	14	37.17	14.74	–		–	–	30	36.99	14.51
IT98	68	7	7	38.54	15.3	–		–	–	68	37.95	14.95
IT00	77	2	2	37.07	13.34	–		–	–	77	36.42	12.57**
IT04	20	4	4	36.62	11.4	18	20	36.49	11.25	20	36.47	11.22
IT08	39	–		–	–	–		–	–	39	35.67	11.2
IT10	48	1	1	35.4	11.34	–		–	–	48	34.96	10.91
IT14	124	2	2	36.41	12.78	124	124	34.93***	11.41***	124	34.96***	11.11***
PS10	3	–		–	–	–		–	–	3	42.46	18.77
PS11	8	–		–	–	–		–	–	8	41.08	18.96
RS06	79	46	46	39.03**	17.61	–		–	–	79	39.31*	18.08
RS10	45	26	26	37.71*	15.44	–		–	–	45	37.88	15.6
RS13	82	47	47	38.98**	15.44	–		–	–	82	39.36*	15.89
RS16	86	38	38	38.88*	16.55	–		–	–	86	39.01	16.7
SI04	1	–		–	–	–		–	–	1	31.71	11.97
SI10	1	–		–	–	–		–	–	1	34.3	14.91
ES95	67	39	39	39.05	13.96	60	67	38.62*	13.7	67	38.46**	13.59
ES00	15	9	11	38.69	17.53	–		–	–	15	38.56	17.4
ES04	115	–		–	–	–		–	–	115	35.87*	15.86**
ES07	74	30	36	34.78*	15.49	53	73	35.15	15.36	74	34.58**	15.11
ES10	200	104	106	36.51***	15.26	173	200	36.49***	14.76***	200	36.03***	14.15***
ES13	97	39	45	37.62	14.95	77	97	37.38**	14.5**	97	37.16***	14.28***
SD09	28	–		–	–	–		–	–	28	54.3	21.58

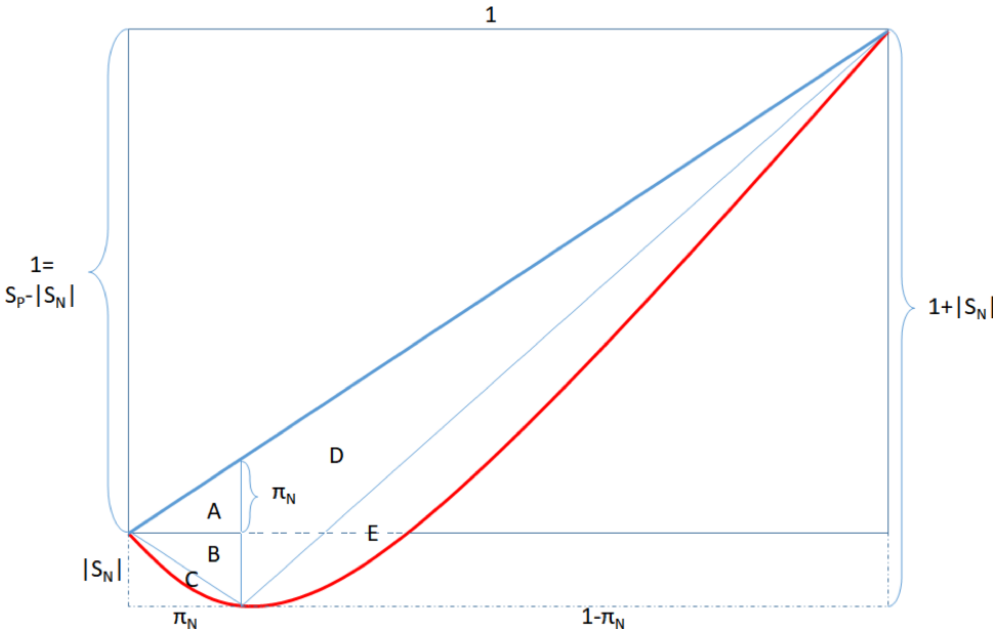
Notes: Years refer to income-reference years. Surveys were harmonized by LIS and ERF. Surveys restricted to those containing non-positive DHIs.

* corrected estimate within 90% (** 95%) confidence interval of uncorrected estimate.

a. For EG12, results for the LMPS survey (LIS database) are provided.

b. For PS10, PS11, self-employment income is unavailable, while for FR00, FR05, IQ12, IT08, SI97, SI99, SI04, SI07, SI10, ES04 and SD09, self-employment income is non-negative. '–' Analysis could not be performed due to the absence of negative or any HILS/HICID/HITP

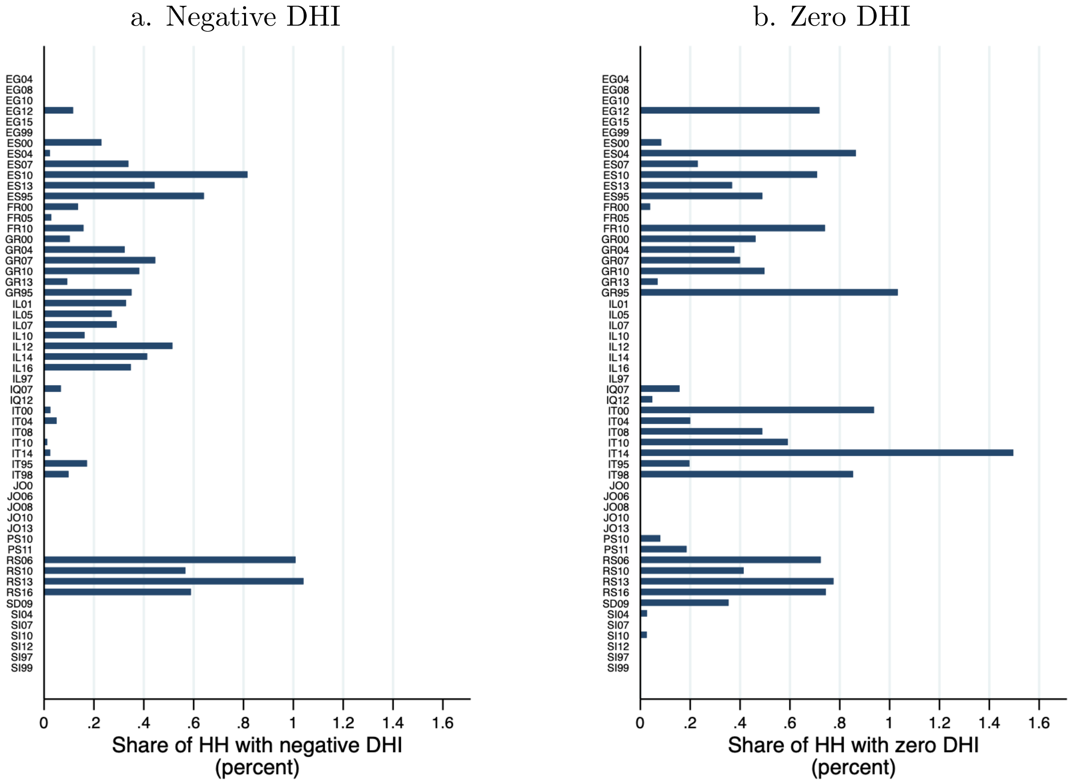
Figure 1: Gini coefficient decomposition using Lorenz curve



Legend: Red line shows the actual Lorenz curve, thick blue line shows the perfect-equality Lorenz curve. π_N is the population share of households with negative incomes; S_N is the (negative) aggregate net-income share of households with negative incomes, and S_P is the aggregate net-income share of households with non-negative incomes ($S_p \geq 100\%$), so that $S_P - |S_N| = 100\%$. The Gini is equal to the areas $(A+B+C+D+E)/0.5$, where $A = (\pi_N^2)/2$, $B = (\pi_N|S_N|)/2$, $C = (\pi_N|S_N|G_N)/2$, $D = ((1-\pi_N)(\pi_N + |S_N|))/2$, and $E = ((1+|S_N|)(1-\pi_N)G_(1-N))/2$. Here G_N is the Gini coefficient estimated among negative incomes, either non-parametrically or parametrically. $G_(1-N)$ is the Gini estimated non-parametrically among non-negative incomes. The overall Gini can thus be computed as:

$$\begin{aligned} G &= ((A+B+C+D+E))/0.5 \\ &= \pi_N^2 + \pi_N|S_N| + G_N\pi_N|S_N| + (1-\pi_N)(\pi_N + |S_N|) + G_(1-N)(1+|S_N|)(1-\pi_N) \\ &= -G_N\pi_N S_N + \pi_N - S_N + G_(1-N)(1-\pi_N - S_N + \pi_N S_N) \end{aligned}$$

Figure 2: Share of households with nonpositive disposable household income

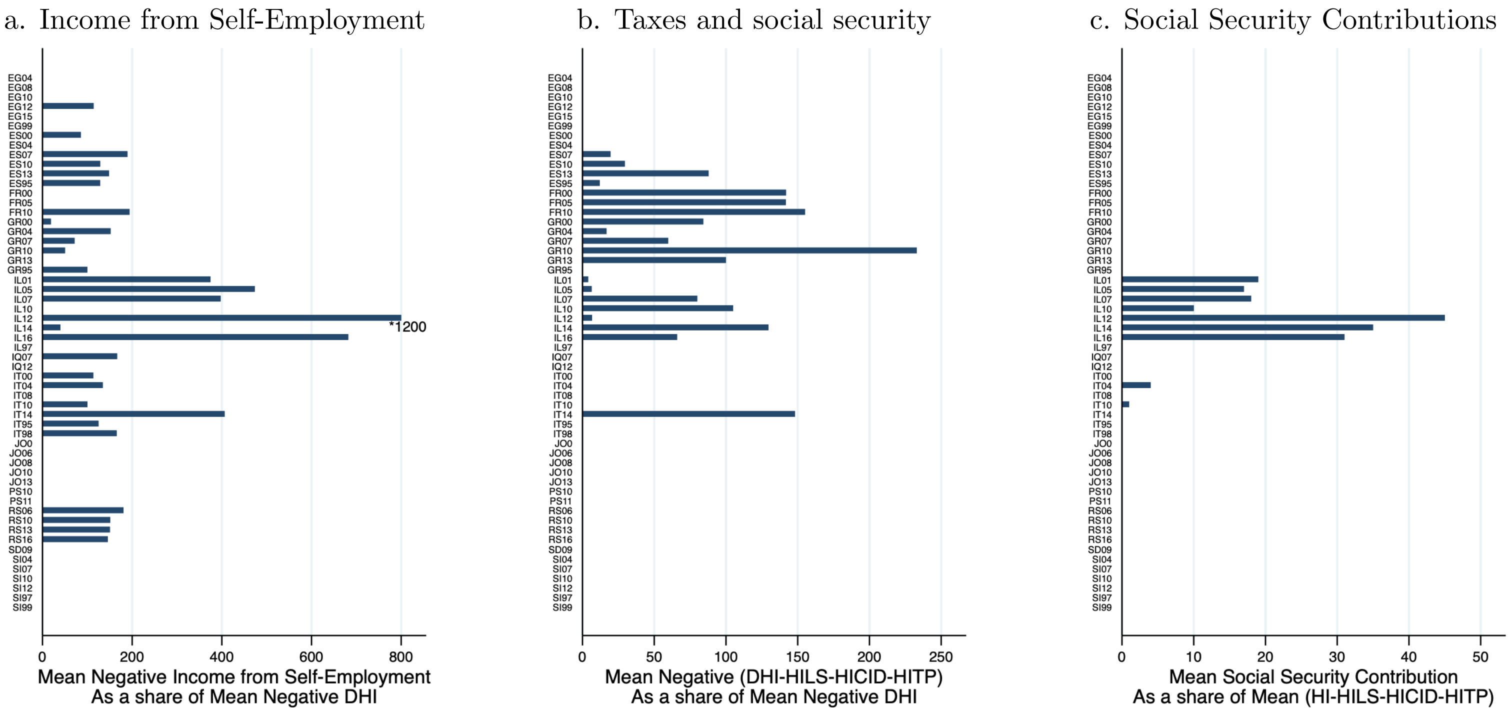


a. Negative DHI

b. Zero DHI

Source: Authors' elaboration from Table 1.

Figure 3: Mean negative income from different sources as a share of mean negative disposable household income



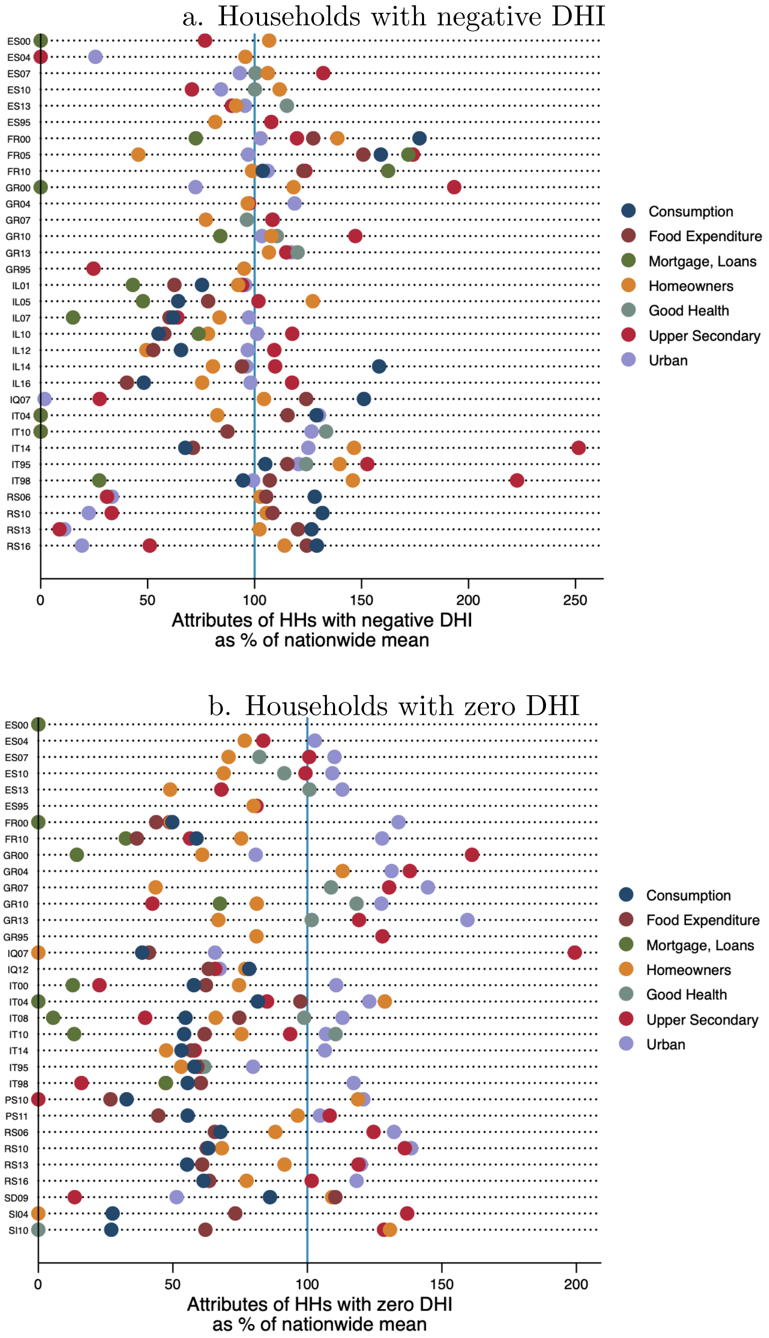
a. Income from Self-Employment

b. Taxes and social security

c. Social Security Contributions

Source: Authors' elaboration from Table 1.

Figure 4: Socioeconomic Characteristics of households with nonpositive disposable household income

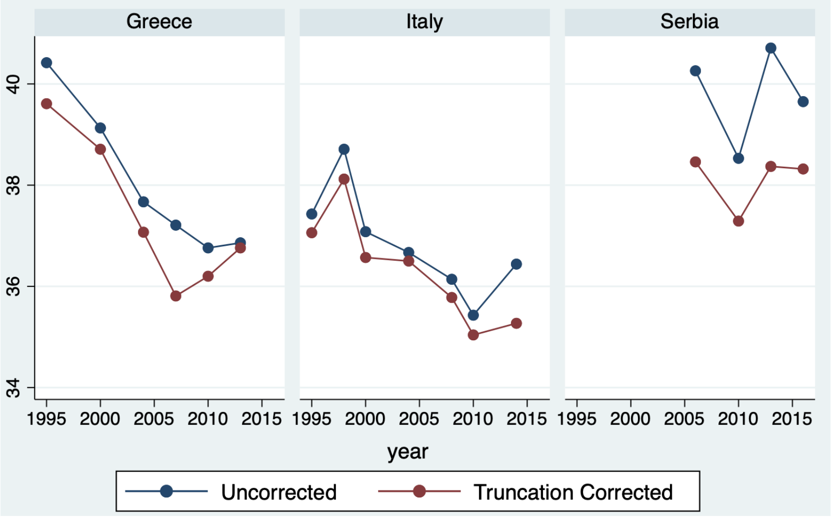

a. Households with negative DHI

b. Households with zero DHI

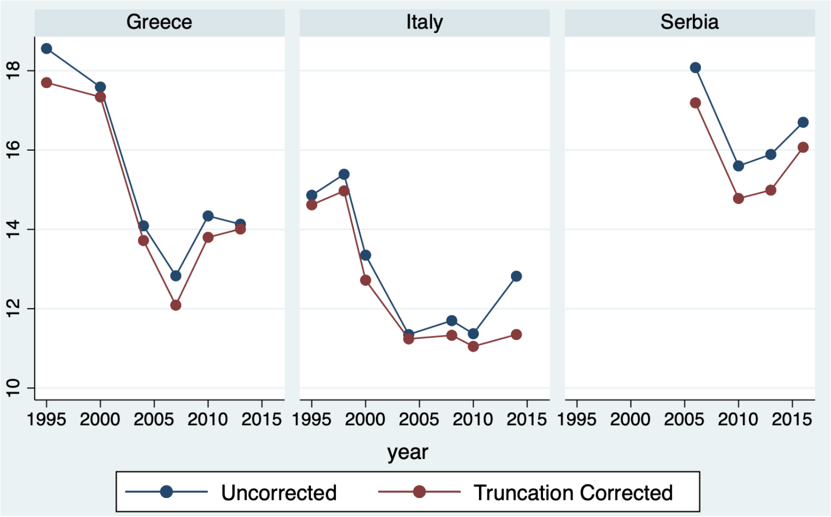
Source: Authors' elaboration from Table 2.

Figure 5: Inequality and poverty on uncorrected vs truncation-corrected DHI distribution: Variation over time

i. Inequality (Gini)

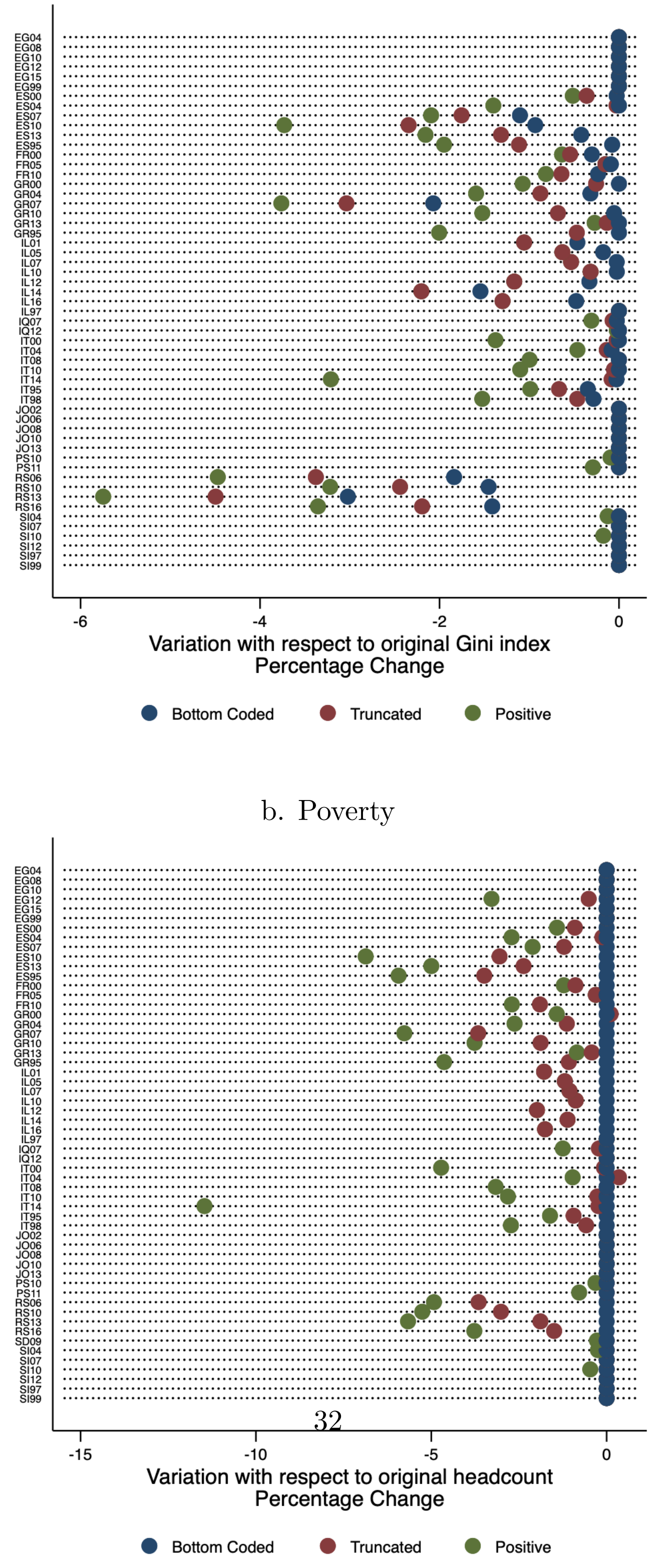


ii. Poverty (H0)



Notes: The horizontal axis shows the time dimension from the first wave to the last wave. Greece, Italy and Serbia are selected as the only countries with 3+ survey waves with corrected estimates. Numbers in this figure are taken from Table 3, using the entire income distribution (uncorrected), or using only positive incomes ($DHI > 0$; corrected distribution)

Figure 6: Distributional changes with different corrections of non-positive incomes

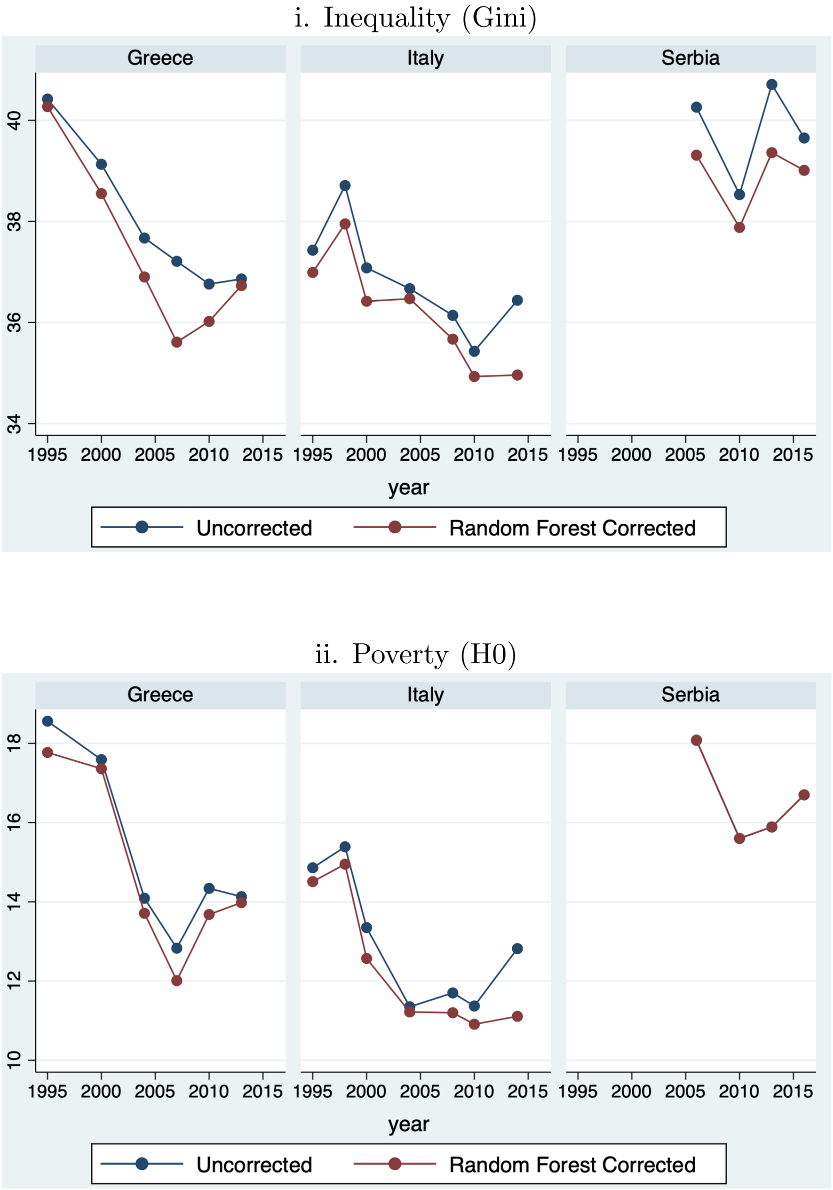


a. Inequality

b. Poverty

Source: Authors' elaboration from Table 3.

Figure 7: Inequality and poverty on uncorrected vs random-forest corrected DHI distribution: Variation over time



i. Inequality (Gini)

ii. Poverty (H0)

Notes: The horizontal axis shows the time dimension from the first wave to the last wave. Greece, Italy and Serbia are selected as the only countries with 3+ survey waves with corrected estimates. The corrected series in this figure are taken from Table 5 column 3