the review of income and wealth

Review of Income and Wealth
Series 68, Number 4, December 2022
DOI: 10.1111/roiw.12535

BOTTOM INCOMES AND THE MEASUREMENT OF POVERTY AND INEQUALITY

BY VLADIMIR HLASNY*

Ewha Womans University

LIDIA CERIANI

Georgetown University

AND

PAOLO VERME

The World Bank

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 data sets from the Luxembourg Income Study and Economic Research Forum databases, covering 12 Mediterranean countries over the period 1995–2016. This 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 withholding, 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 off than, other households in terms of material well-being. On the contrary, zero-income households are found to be materially deprived. Adjusting poverty and inequality measures for these findings can alter these measures significantly.

JEL Codes: D31, I32, N35

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

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 nonresponse, 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 has generated a significant body of literature covering high- and low-income countries (Atkinson *et al.*, 2011; Cowell and Victoria-Feser, 1996; Hlasny and Verme, 2018b; Hlasny, 2020; Jenkins *et al.*, 2011). These contributions propose parametric and nonparametric methods to correct inequality measurement based on known top income properties (Hlasny,

*Correspondence to: Vladimir Hlasny, Economics Department, Ewha Womans University, Seoul, Korea (vhlasny@gmail.com).

2021) or 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. While consumption, which is always nonnegative, could serve as a better welfare aggregate among poorer households, many high- and middle-income countries opt to work with incomes when measuring poverty or inequality. In these cases, a typical approach by statistical agencies and researchers with respect to bottom incomes is to bottom-code or censor incomes at zero. Some scholars have acknowledged this as a potential 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 negative or zero incomes. This paper studies their prevalence and composition, and their potential impact on the measurement of poverty and inequality.

The presence of negative incomes is quite common in household surveys. It is not obvious that these incomes represent poor households. For example, in the sample of 354 data sets in the Luxembourg Income Study (LIS) database made available to the authors in February 2019, 229 data sets contained negative disposable household incomes (DHIs). In 12 data sets, negative incomes accounted for over 1 percent of nonzero incomes and numbered up to 584 observations in a national survey. These negative incomes were not trivial in size. Mean negative income was as large in absolute value as 754 percent of mean nationwide positive income, and exceeded 200 percent of mean nationwide positive income in 15 data sets. Whether these negative incomes reflect accurately households' current welfare, or whether they are artifacts of accounting practices, data-entry errors, or statistical agencies' treatment, 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 data sets evaluated, 270 contained zero incomes. In 22 data sets, zero incomes accounted for over 1 percent of nonnegative incomes and numbered up to 1213 observations. These zero incomes were 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.

Negative and zero incomes can be critical for the measurement of poverty and inequality. The majority of inequality and poverty measures are defined on positive incomes only, and scholars tend to drop these observations by default. This is the case, for example, of measures that include logarithmic or fractional power transformations such as most of Foster, Greer, and Thorbecke (FGT) poverty measures, most Generalized Entropy (GE) measures, or Atkinson and Watts indices.[1] In other cases, including negative observations can alter the properties of

[1] We are grateful to an anonymous referee for suggesting to emphasize this point.

some measures. The Gini index, for example, can lose the upper bound of 1, and the Lorenz curve can be below zero if negative values are included in estimates. These changes in boundaries because of bottom incomes can make these inequality measures less appealing to practitioners.

Understanding the bottom tail of income distributions is also important from a policy perspective, arguably more important than understanding the top tail of the distributions. 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 biasing poverty 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 data sets covering 12 Mediterranean 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, high social security withholding, 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 off than, other households in terms of material wellbeing. By contrast, zero-income households are found to be materially deprived. This paper also proposes alternative methods to adjust poverty and inequality measures for the suspected issues and concludes that a proper classification of bottom incomes can alter these measures nontrivially.

This 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 3 describes the available data. Section 4 outlines the main methods used to correct for the negative and zero incomes and assesses the distributional impacts of the corrections. Section 5 concludes with a discussion of the results.

2. Definitions and Methods

When measuring poverty or inequality, negative and zero incomes are typically either bottom-coded or truncated by statistical agencies or researchers, and may thus be excluded from measurement. As a result, inequality and poverty can be mis-measured and most probably are underestimated (Ceriani and Verme, 2019). 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.[2] Negative incomes may also be found among non-poor households so that counting negative

[2] 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.

incomes among the poor can bias the measurement of poverty upwards. Therefore, biases generated by bottom-coding and truncation may operate in the opposite direction as biases generated by negative incomes, making the proper assessment of poverty and inequality very complex. This section provides taxonomy and methodologies to properly account for negative and zero incomes when measuring poverty and inequality.

2.1. *Data and Definitions*

Our study relies on 57 household data sets from 12 Mediterranean countries for the years 1995–2016, harmonized and made available through a partnership between LIS and ERF. The LIS database consists of microdata collected from six continents spanning five decades, harmonized and arranged for free use among registered users over a remote-execution online platform. The ERF database is available through ERF's Open Access Micro Data Initiative (OAMDI), in collaboration between ERF and national statistical agencies in the ERF region. The ERF data sets have been cleaned, harmonized, and made available to registered users for full download. In 2019, LIS and ERF joined forces to offer registered users access to a combined database harmonized according to a common template. The present paper was among a select group of studies authorized under the pilot program to assess "inequality trends around the Mediterranean," which informed the choice of country data sets evaluated here.[3]

The LIS database contributes income distributions for seven countries, namely Greece, France, Israel, Italy, Serbia, Slovenia, and Spain, while the ERF database contributes the distributions for Egypt, Iraq, Jordan, Palestine, and Sudan.[4] These countries are particularly interesting for our analysis because they encompass high-, medium-, and low-income countries, and exhibit 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 frequency of low reported incomes. Among these surveys, there are also subsets with similar income distributions, yet different prevalence and composition of nonpositive incomes.

The data are not without problems. 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. In addition, among the variables available, some income components are missing between LIS and ERF data sets (or between the alternate sources of Egyptian data in both repositories), and cannot be assessed across the entire sample of data sets.[5] With non-income variables, the problems are analogous. This explains the various gaps in Tables 1 and 2.

[3] For information on the LIS and ERF databases, refer to www.lisdatacenter.org and https://erf.org.eg/erf-data-portal. The coverage of country-years and variables in the joint database is explained in www.lisdatacenter.org/our-data/erf-lis-database.

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

[5] 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.

TABLE 1
Components in Negative Incomes: Data Sets 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[aL]	12039	173	1.437	28	12.60	28	113.99	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	194.20	10	155.08	0	.	BDF
GR95[L]	4842	50	1.033	17	0.28	17	100.00	0	.	0	.	GRECHP
GR00[L]	3895	18	0.462	4	1.05	3	18.88	4	84.18	0	.	GRECHP
GR04[L]	5568	21	0.377	18	27.03	16	151.82	7	16.77	0	.	EU-SILC
GR07[L]	6503	26	0.400	29	100.96	18	71.47	25	59.70	0	.	EU-SILC
GR10[L]	6024	30	0.498	23	3.82	2	50.19	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	166.43	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	374.42	14	3.93	19	60.51	HES
IL05[L]	6272	0	0.000	17	18.92	7	473.44	13	6.38	17	24.12	HES
IL07[L]	6172	0	0.000	18	1.95	1	397.23	17	79.99	18	468.77	HES
IL10[L]	6168	0	0.000	10	1.27	0	.	10	105.00	10	672.33	HES
IL12[L]	8742	0	0.000	45	16.71	3	2284.58	42	6.68	45	17.34	HES
IL14[L]	8465	0	0.000	35	103.84	6	39.97	31	129.54	35	39.31	HES
IL16[L]	8903	0	0.000	31	25.53	4	681.88	27	65.92	31	14.47	HES

TABLE 1
Continued

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 (%)	
IT95^{L}	8134	16	0.197	14	48.26	14	124.91	0	.	0	.	SHIW
IT98^{L}	7147	61	0.854	7	70.05	7	165.44	0	.	0	.	SHIW
IT00^{L}	8000	75	0.938	2	13.06	2	113.21	0	.	0	.	SHIW
IT04^{L}	8012	16	0.200	4	71.07	4	134.38	1	0.00	4	22.32	SHIW
IT08^{L}	7977	39	0.489	0	.	0	.	0	.	0	.	SHIW
IT10^{L}	7941	47	0.592	1	3.77	1	100.32	0	.	1	152.75[b]	SHIW
IT14^{L}	8151	122	1.497	2	9.76	2	406.16	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	180.73	0	.	0	.	HBS
RS10^{L}	4585	19	0.414	26	67.15	26	151.17	0	.	0	.	HBS
RS13^{L}	4517	35	0.775	47	86.31	47	150.34	0	.	0	.	HBS
RS16^{L}	6448	48	0.744	38	79.17	38	145.71	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	128.58	2	12.00	0	.	ECV
ES00^{L}	4776	4	0.084	11	3.33	9	85.68	0	.	0	.	ECV
ES04^{L}	12950	112	0.865	3	3.49	0	.	0	.	0	.	EU-SILC

TABLE 1
Continued

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 (%)	
ES07[L]	13014	30	0.231	44	84.74	30	189.31	19	19.49	0	.	EU-SILC
ES10[L]	13109	93	0.709	107	29.52	104	128.65	66	29.47	0	.	EU-SILC
ES13[L]	11965	44	0.368	53	21.17	39	148.26	27	87.88	0	.	EU-SILC
SD09	7913	28	0.354	0	.	0	.	0	.	0	.	NBHS

Notes: 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, $HI = 1890$ 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, and Repaymts.	Home-ownership	Good Health	Upper Secondary Education	Urban	Consump.	Food Expend.	Outflow from Mortg., Loans, and 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	.	.	.	101.0	104.5	105.6	76.5
EG15	.	.	.	.	.	.	.	.	.	.	.	.	.	.
FR00	177.1	127.4	72.5	138.7	.	119.8	102.8	49.8	43.8	0.0	49.3	.	48.8	133.9
FR05	159.0	150.8	171.9	45.7	.	174.0	97.0	.	.	.	.	.	.	.
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	.	.	0.0	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	.	.	84.0	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	.	0.0	.	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	105.0	115.3	.	139.8	124.1	152.7	120.5	58.0	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	.	16.0	117.2
IT00	111.8	88.3	1505.90	80.9	.	0.0	123.6	57.8	62.3	12.8	74.6	.	22.7	110.7
IT04	128.9	115.4	0.0	82.6	.	0.0	130.1	81.6	97.4	0.0	128.8	.	85.0	123.0
IT08	.	.	.	.	.	.	.	54.7	74.7	5.5	65.9	98.8	39.7	113.1
IT10	.	87.3	0.0	0.0	133.4	0.0	126.6	54.2	61.8	13.3	75.5	110.5	93.6	106.9

TABLE 2
Continued

	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, and Repaymts.	Home-ownership	Good Health	Upper Secondary Education	Urban	Consump.	Food Expend.	Outflow from Mortg., Loans, and Repaymts.	Home-ownership	Good Health	Upper Secondary Education	Urban
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	.	0.0	120.8
PS11	.	.	.	.	.	.	.	55.5	44.6	.	96.4	.	108.3	104.7
RS06	128.1	105.4	.	102.7	.	31.0	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	11.0	55.3	60.9	.	91.5	.	119.0	119.9
RS16	129.1	124.4	.	114.0	.	51.0	19.3	61.4	63.5	.	77.4	.	101.6	118.3
SI97	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SI99	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SI04	.	.	.	.	.	.	.	27.6	73.2	.	0.0	.	137.1	.
SI07	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SI10	.	.	.	.	.	.	.	27.1	62.1	.	130.7	0.0	128.5	.
SI12	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ES95	.	.	.	81.6	.	107.8	.	.	.	.	80.0	.	81.1	.
ES00	.	.	0.0	106.8	.	76.8	.	.	.	0.0	0.0	.	0.0	.
ES04	.	.	.	95.7	.	0.0	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	.	.	.	49.0	100.8	68.0	113.0
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 nonnegative values.

As a measurement unit, we use *DHI* equivalized per adult equivalent using the square root of household size. *DHI* is what is normally reported by households in surveys, which 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.

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.[6] 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 nonpositive 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, which is less sensitive to corrections for nonpositive incomes than, say, the FGT indicators of poverty intensity including the income gap ratio (*IGR*) (Foster *et al*., 1984, 2010).[7] Therefore, the poverty and inequality measures considered here are rather insensitive to changes in the tails and, as such, our corrections should be considered as lower bound corrections as compared to those conducted on 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 origins and should be treated separately. Among negative incomes, we should strive to distinguish those "valid" from the welfare and capabilities perspectives—implying that households are unable to meet basic needs and lack essential capabilities—and those because of accounting considerations without real effects on households' material well-being.[8] 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 HI can come in the form of labor (HIL), capital (HIC) or transfer income (HIT), or high income tax liability (HXITI) and social security contributions (HXITS). The income components could be further subdivided into paid employment income (HILE) and self-employment income (HILS), interest and dividends (HICID), voluntary individual pensions (HICVIP), rental income (HICREN) and royalties (HICROY), and social security transfers (HITS) and private transfers

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

[7] As alternative inequality and poverty indexes, we also report Theil's entropy index (a 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).

[8] There are factors that could explain negative incomes that cannot be checked with available data but are nevertheless possible. For example, if there is a time mismatch between the recording of family composition and income, this may artificially alter household income (HI). Business incomes and outlays may accrue at different points in time across reporting years, and with different frequencies. Bottom coding of income components and top coding of liabilities for negative DHI could also artificially result in negative incomes.

(HITP). Liability components (HXITI) could be subdivided into income tax withholdings (HXITIW) and adjustments (HXITIA), and social security contributions paid by self (HXITSS) and paid on behalf of others (HXITSB). In sum (in bold the potential negative incomes)[9]:

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

In each data set, among households with negative disposable incomes, we calculate the frequencies of negative capital incomes, negative self-employment incomes, or tax withholding 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 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 data set.

As a measure of undue liabilities for taxes and social security contributions, we evaluate the part of *DHI* that is expected to be nonnegative, 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 individual effects of tax withholding, adjustments, and social security contributions.

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 mis-measure even households' actual annual incomes. Self-employment income in particular is prone to mis-measurement (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. Therefore, negative self-employment incomes may come from under-reporting. Second, household surveys provide information over a short

[9] Among these components, we have $HI = HIL + HIC + HIT$; $HXIT = HXITI + HXITS$; $HIL = HILE + HILS$; $HIC = HICID + HICVIP + HICREN + HICROY$; $HIT = HITS + HITP$; $HXITI = HXITIW + HXITIA$; $HXITS = HXITSS + HXITSB$.

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 have gained from their sales or services but also their annualized investment in self-employment activities.[10]

2.2. *Assessing the Composition and Sources of Nonpositive Incomes*

We start by assessing the prevalence of negative and zero incomes across country data sets. We then survey their size distributions, and we identify the likely culprits of the observed values. We draw qualitative conclusions regarding the true capabilities and well-being 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 secondary or higher education, and subjective health rating of good or better, to uncover patterns and irregularities. We also evaluate the links between incomes and households' functionings including total consumption, food consumption, and home ownership, as per data availability across data sets. We compute households' "monetary overconsumption" as the excess of total monetary consumption over final monetary income.[11] From the analysis of this overconsumption between households with negative, zero, and positive *DHI*, we assess the quality of the respective observed *DHI*s as measures of households' capabilities and welfare.

The careful incidence analysis of nonpositive incomes by source and by household type is important, because the relationship may be complex and nonmonotonic. Households' over-consumption 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 real 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, nonpositive incomes are clearly short of the "wolf point" of income necessary for bare survival (Davis, 1941, p. 405). Finding that households with nonpositive incomes do not have a profile of deprived units, we may wish to truncate the reported nonpositive values of individual income sources, or replace them with positive values from households with matching characteristics.

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

[11] 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.

2.3. *Adjusting for Nonpositive Incomes*

Statistical corrections for problems in the tails of income distributions can be characterized as broadly pursuing one of the following two approaches: (1) reweighting, whereby original observations are kept intact while weights are recalibrated, or (2) replacing, whereby weights are kept intact but some observations are removed and replaced by others artificially generated (Hlasny and Verme, 2018a). To address negative and zero incomes, standard statistical adjustments have included data trimming (essentially a reweighting exercise) or bottom coding (replacing) (Eurostat, 2006b). We apply these corrections and compare them with the corrections provided by two more advanced replacing methods, one parametric and one nonparametric: parametric modeling of nonpositive incomes, and random forest imputation of incomes using information on households' composition, sector of employment, housing, and other characteristics.

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

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

where y is income, and y^u is the upper cutoff for modeling bottom incomes, such as $y^u = min\,(max(0.3\mu, Q(0.02)), Q(0.03))$, where μ is the mean income and $Q()$ are the quantiles, proposed empirically by Van Kerm (2007). $\theta > 0$ is an estimable shape parameter that can be made robust to extreme incomes using an optimal B-robust estimator (Victoria-Feser and Ronchetti, 1994), 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

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

where π_1 and π_2 are the shares of negative and zero incomes, respectively, α, β, γ, and $\delta > 0$ are the estimable parameters, and 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, p. 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.

Rather than using a matching method, we propose to use a random forest algorithm to predict household welfare based on households' 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.[12]

2.4. *Assessing the Distributional Impact of Imputing Nonpositive Incomes*

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

When the income distribution includes both negative and nonnegative incomes, the Gini coefficient can be obtained using the Lorenz curve from the Gini coefficients among negative incomes (G_N) and among nonnegative incomes (G_{1-N}) by knowing the population share 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 (also refer to Ostasiewicz and Vernizzi, 2017).

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

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

It is important to note that when one estimates the Gini coefficient with negative values, the upper limit of the Gini may be larger than one. This makes the upper bound of the Gini open. Therefore, when two Gini coefficients derived from two income distributions with different shares of negative incomes are compared, the upper bounds of the Gini are likely to differ (also refer to De Battisti and Vernizzi (2019)).

3. Data and Descriptive Statistics

Across the 57 data sets evaluated, 33 contain zero values for *DHI* (57.9 percent) accounting for up to 173 observations in a data set or 1.5 percent of the sample. Thirty-four data sets contain negative values (59.6 percent) accounting for up to 107 observations, which average (in absolute value) as much as 104 percent of the mean of positive incomes in a data set (Table 1). Among the northern Mediterranean surveys, zero incomes are more prevalent than negative incomes in

[12] Refer to Fitzpatrick and Dupriez (2018) and details of this competition on GitHub: https://github.com/worldbank/ML-classification-algorithms-poverty.

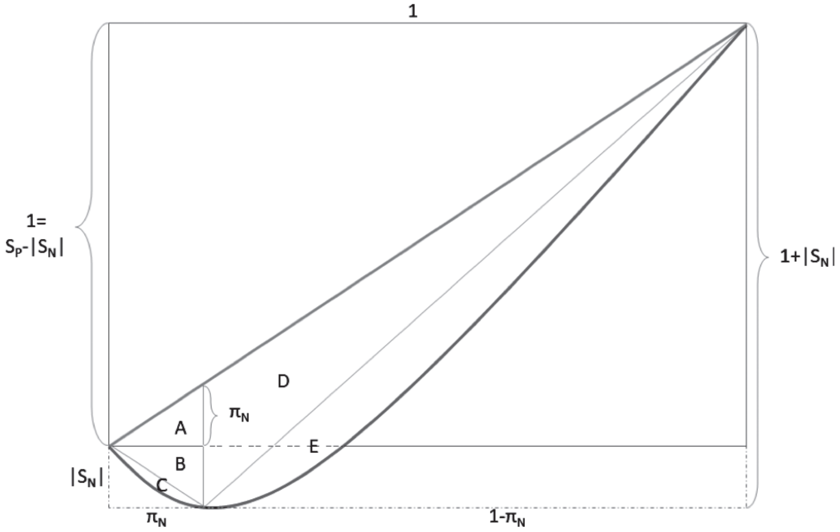


Figure 1. Gini Coefficient Decomposition Using Lorenz Curve

Legend: Thick dark gray line shows the actual Lorenz curve, and thick light gray 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 nonnegative 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 nonparametrically or parametrically. $G_(1 - N)$ is the Gini estimated nonparametrically among nonnegative incomes. The overall Gini can thus be computed as: $G = ((A+B+C+D+E))/0.5 = \pi_N^2 + \pi_N|S_N| + G_N\pi_N|S_N| + (1-\pi_N)(\pi_N+|S_N|) + G_(1-N)(1+|S_N|)(1-\pi_N) = -G_N\pi_N S_N + \pi_N - S_N + G_(1-N)(1-\pi_N - S_N + \pi_N S_N)$

France and Italy, as prevalent as negative incomes in Greece, Serbia, and Spain, and entirely or near nonexistent in Israel and Slovenia (respectively). Among the southern Mediterranean surveys, only the Egyptian 2012 data set in the LIS database and the Iraqi 2007 data set contain negative incomes, but zero incomes appear in Iraq 2012, and in the Palestinian and Sudanese data sets 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 and type of recall on interviews) and administrators' practices (e.g., bottom coding and imputation), rather than with countries' socioeconomic conditions. When negative incomes are present in a survey, their values 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, undue liabilities for taxes, and social security contributions ($DHI - HILS - HICID - HITP$), and the burden of social security contributions alone (see Figures 2 and 3). A quick review suggests that, empirically and among the various income components,

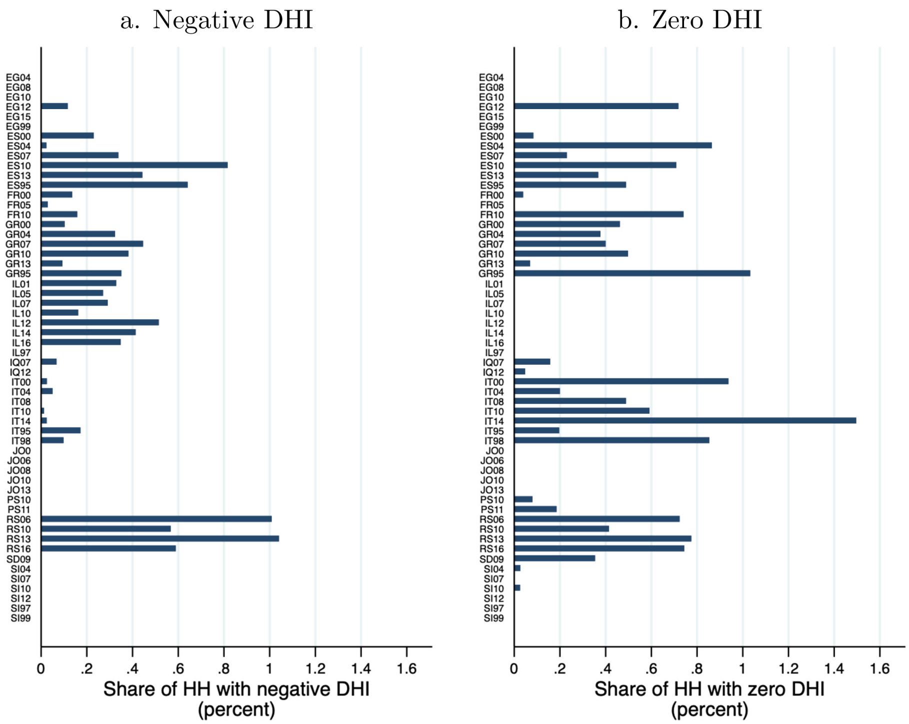
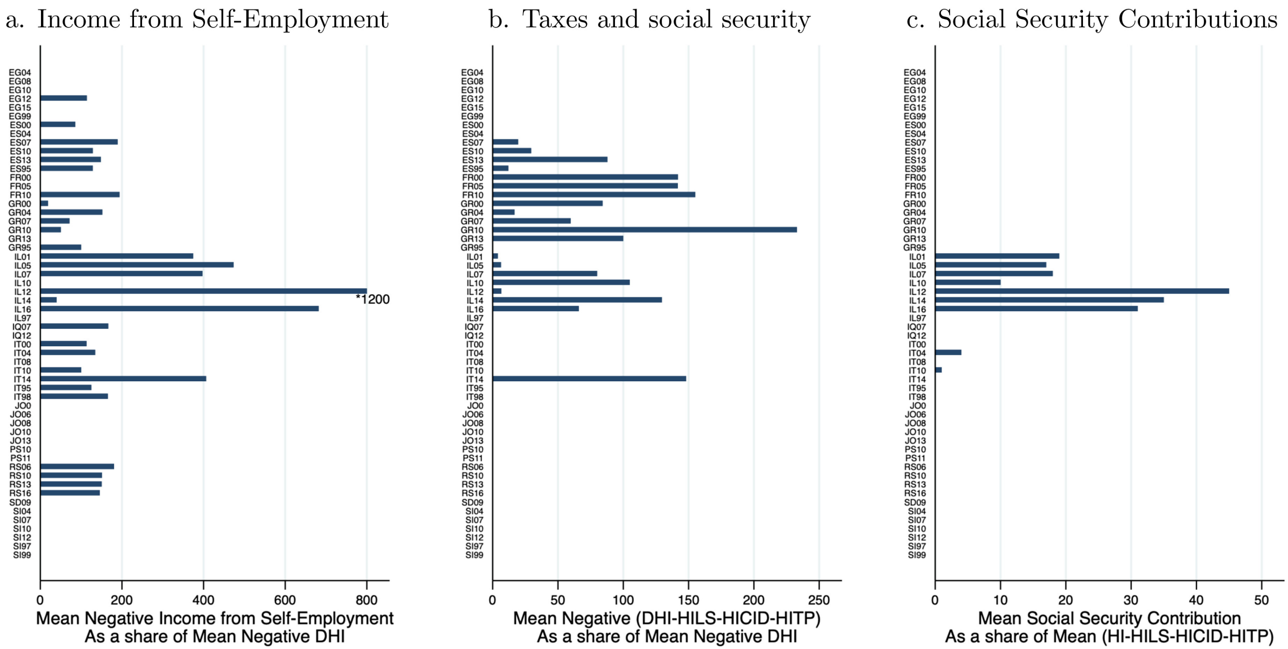


Figure 2. Share of Households with Nonpositive Disposable Household Income.
Source: Authors' elaboration from Table 1. [Colour figure can be viewed at wileyonlinelibrary.com]

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, and 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 nonnegative for all households in all data sets, and so it does not contribute to explaining negative *DHI*s (not reported in Table 1).

Data for Greece, Italy, and Serbia in the LIS database show that up to 1 percent of households report negative disposable incomes, linked to negative self-employment incomes (accounting for 50–150 percent of the reported negative *DHI*). In Greece 2013 and recent Italian surveys, tax and social security withholding also accounts for a handful of negative incomes (averaging 112 percent of the size of reported negative *DHI* in Greece, and 148–290 percent in Italy 2010–2014). In Israel, the count of negative *HILS* is lower, but the values are much larger (of 350–2000 percent of the size of the negative *DHI*). In Spain, the negative incomes are predominantly due to self-employment (120–200 percent), but in 2004 the three negative income values were due to large income-tax burdens. In Egypt 2012, 191 households recorded zero incomes and 10 recorded negative incomes. These 10 are on account of large negative *HILS*.



a. Income from Self-Employment

b. Taxes and social security

c. Social Security Contributions

Figure 3. Mean Negative Income from Different Sources as a Share of Mean Negative Disposable Household Income.
Source: Authors' elaboration from Table 1. [Colour figure can be viewed at wileyonlinelibrary.com]

In sum, the available evidence suggests that negative *HILS* is the primary source of negative *DHI* in three-quarters of all data sets, while in other data sets the problem is mainly due to high social security and other burdens. Interestingly, when country data sets are sorted by the frequency of negative *DHI*, negative *HILS* shows up as the top source of their prevalence. By contrast, when data sets 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 level 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.

Finally, it should be stressed that the data used in this paper are limited to high- and middle-income countries. This excludes low-income countries characterized by large agricultural subsistence sectors and large non-agricultural informal sectors. In those countries, the problem of negative and zero incomes may be expected to be larger than in richer countries. For example, agricultural income varies substantially across the year and recorded incomes largely depend on the time of the year the survey is administered. Even when income is recorded with recall questions, assessing the value of gross or net income may be very difficult for small subsistence farmers. Similarly, workers in urban informal activities tend to have occasional or irregular incomes which may vary across the year. While including low-income countries is a real challenge because of a lack of proper income data, we may expect the problem of negative and zero incomes to be greater than in higher income countries.

3.1. *Association of Incomes with Other Socioeconomic Outcomes*

Next, in each data set where it is available, we calculate mean household consumption, consumption of food, homeownership, self-reported health, and education among households with nonpositive *DHI*, and we relate these figures to those for households with positive *DHI*. This helps to identify the true welfare of households with nonpositive *DHI* across different data sets.[13] We also calculate mean outflows from mortgages, loans, and repayments, to proxy for 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 percent) than the respective national means among households with positive *DHI*. This is not new or unique to these countries and something that has been observed in other countries such as the UK in the past (Brewer *et al.*, 2017). 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 data sets, and the only observable pattern is that in France

[13] We also evaluate this on all 354 data sets in the LIS database. Consumption is available in 43 data sets and food consumption in 73 data sets. Negative-*DHI* households do not appear to have unduly low consumption. Mean food consumption of negative-*DHI* is not a cause for concern. Some negative-*DHI* households appear to be food-poor.

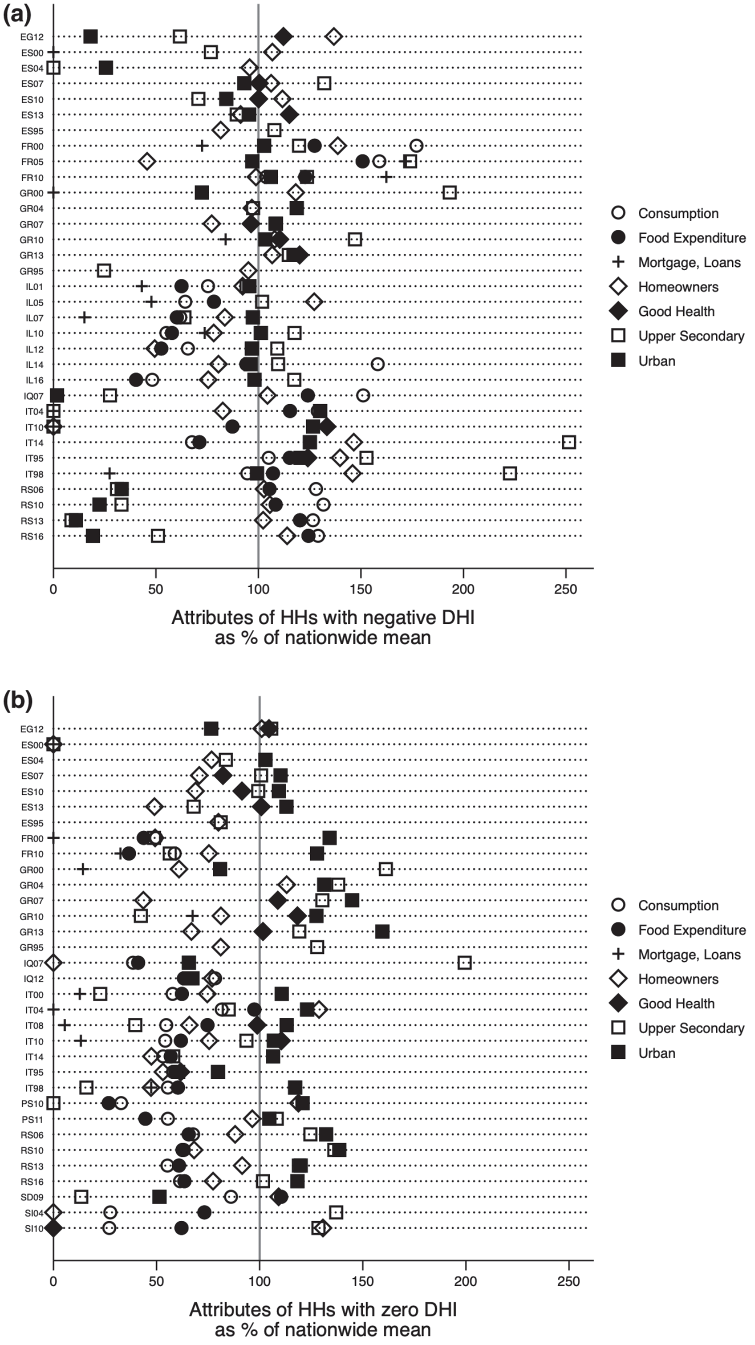


Figure 4. Socioeconomic Characteristics of Households with Nonpositive Disposable Household Income.

Source: Authors' elaboration from Table 2.

households with negative *DHI* are more burdened by debts than the national mean, while in Greece and Israel the opposite is true. In Italy and Spain, no clear patterns emerge. Regarding the completion of secondary education, it is not 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 data sets for Greece, Italy, and Spain, but households with negative *DHI* systematically outperform the respective national means.

These patterns differ clearly from those for zero-income households. Zero-income households in France, Italy, and Slovenia have a 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 contrary, 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).

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 than those with positive incomes, while zero-income households in France, Greece, and Serbia are more likely to be urban (and as likely as positive-income households 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 well-being, 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.

4. Adjusting Welfare Measures for Nonpositive Incomes

4.1. *Bottom-coding Negative Incomes*

Table 3 and Figure 5 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 nonpositives)—leads to a systematic monotonic fall in the inequality and poverty indexes.

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 have 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 also

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)	–	–	–	6.00 (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)	53.00 (1.18)	52.32 (1.19)	18.59 (0.35)	18.59 (0.35)	18.40 (0.35)	17.37 (0.35)
EG15	0	0	35.20 (1.48)	–	–	–	5.05 (0.20)	–	–	–
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)	8.90 (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)	34.10 (0.38)	34.04 (0.38)	9.99 (0.24)	9.99 (0.24)	9.80 (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)	17.70 (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)	13.80 (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 (0.50)	19.31 (0.50)	19.14 (0.50)	–
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	39.50 (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)	17.90 (0.41)	–
IT95	14	16	37.43 (0.48)	37.30 (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)

TABLE 3
Continued

			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 ($DHI > 0$)
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)	15.30 (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)	11.70 (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 (0.70)	–	–	–
JO06	0	0	40.32 (1.29)	–	–	–	11.64 (0.60)	–	–	–
JO08	0	0	40.33 (1.66)	–	–	–	11.26 (0.60)	–	–	–
JO10	0	0	41.01 (1.88)	–	–	–	11.57 (0.60)	–	–	–
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 (0.60)	19.13 (0. 60)	–	–	18.98 (0.60)
RS06	46	33	40.26 (0.51)	39.52 (0.45)	38.90 (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)	15.60 (0.54)	15.60 (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)	16.70 (0.46)	16.70 (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)	–	–	31.70 (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)

TABLE 3
Continued

			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 ($DHI > 0$)
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)	38.70 (0.54)	38.64 (0.54)	17.66 (0.55)	17.66 (0.55)	17.50 (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)	16.60 (0.33)	16.60 (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)	34.60 (0.28)	15.63 (0.32)	15.63 (0.32)	15.44 (0.32)	15.30 (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)	15.20 (0.33)	15.20 (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)	–	–	21.60 (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).

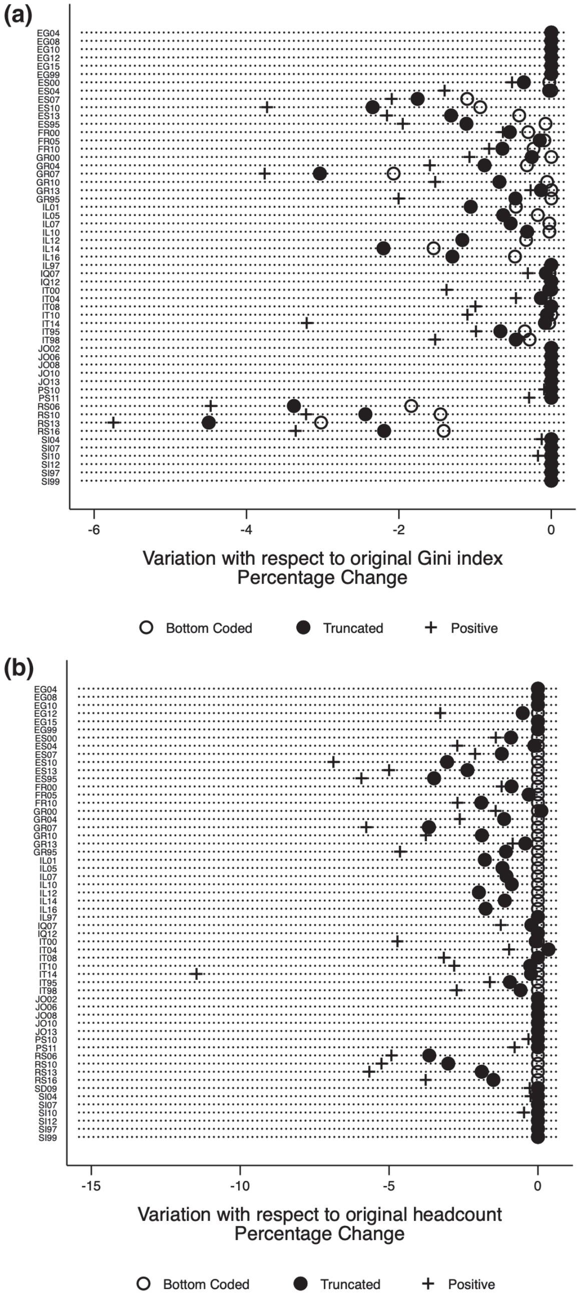


Figure 5. Inequality and Poverty on Uncorrected vs Truncation-Corrected *DHI* Distribution: Variation Over Time.

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+ data sets 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) 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 headcount by up to 1.0 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 nonpositive incomes have noticeable effects even on inequality and poverty measures known to be relatively robust to adjustments in the distribution tails. Between the uncorrected values and the values corrected by deleting all nonpositive 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, and ES10; and 0.8–0.9pt. in EG12, GR95, ES95, and ES13), while poverty headcount falls by up to 1.5 points (1.1–1.5pt. in EG12, IT14, and ES10; 0.8–0.9pt. in GR95, RS06, RS10, RS13, ES95, and ES13).

For countries with three or more time observations, we can evaluate how the correction affects the trend and volatility in inequality and poverty. Figure 6 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 visibly, except in the case of the Serbian Gini, which falls with the correction as may be expected.

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 (I) distribution, or the two-parameter generalized Pareto (II) distribution. We find that the Pareto (I) distribution does not offer 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 data sets (out of 33 containing 2+ negative incomes) do we get plausible results. In these data sets, combining the parametric Gini coefficients for negative incomes with nonparametric Gini coefficients for positive incomes yields trivial corrections to the Gini coefficients reported in Table 3, on the order of 0.01 percentage points of the Gini. This is due to the small number of negative incomes in the data. The overall Gini appears to be robust to the method for treating negative incomes.

The two-parameter generalized Pareto (II) distribution provides a somewhat better fit, thanks to the flexibility provided by the additional parameter. For 18 data sets out of 33, we estimate plausible coefficients, income shares, and parametric Gini coefficients. Combining the parametric Gini coefficients for negative incomes with nonparametric Gini coefficients for positive incomes yields small corrections to the Gini coefficients, of up to 0.70 percentage points (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 data.

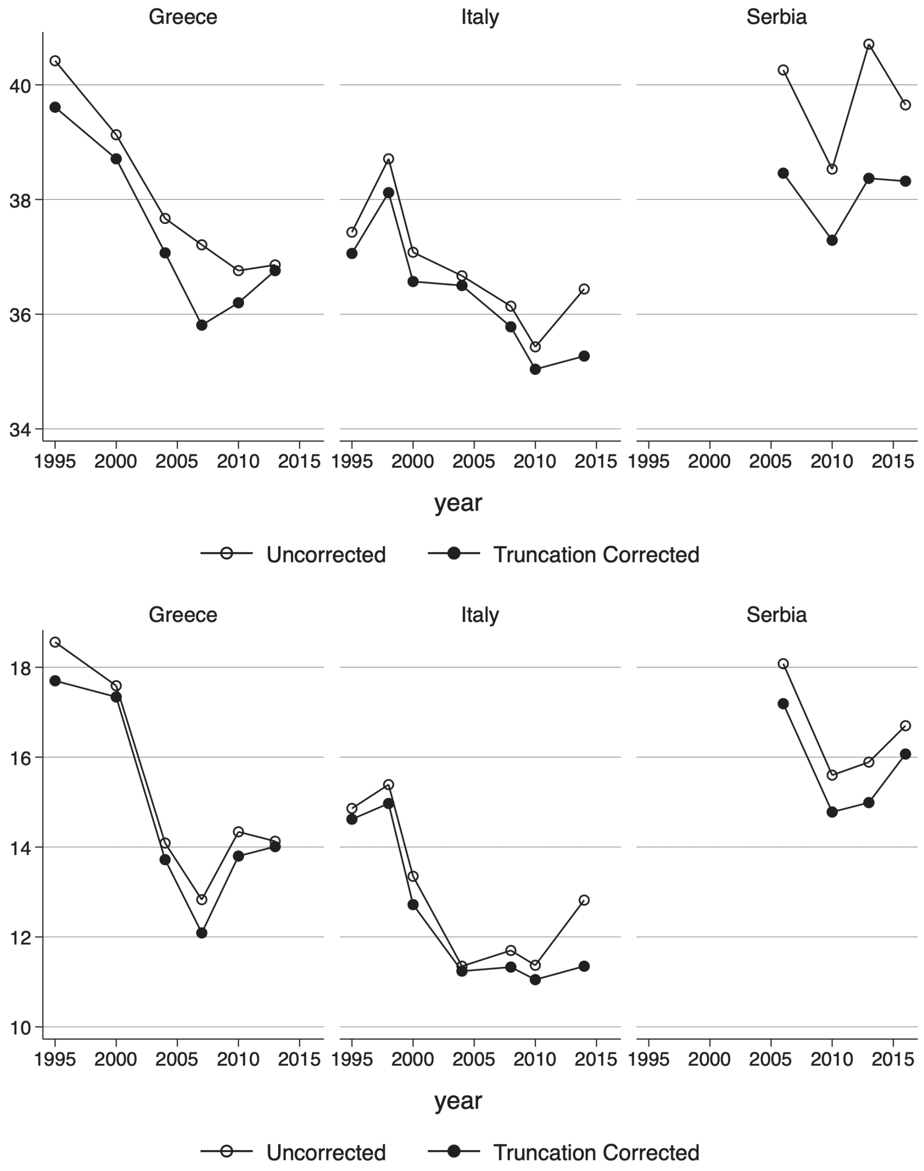


Figure 6. Distributional Changes with Different Corrections of Nonpositive Incomes *Source*: Authors' elaboration from Table 3.

It is important to note that the shape parameter is estimated exactly at the mean of the corrected parametric distribution (by design), and almost exactly at the mean of the uncorrected original distribution in all data sets with sufficient negative incomes. The parametric Gini of negative incomes is uniformly 0.5 by design. Therefore, in Equation 3, the first half of the expression is essentially the same with and without the parametric correction. The semi-parametric Gini of

TABLE 4
Estimates of Pareto and Generalized Pareto Distributions among Negative Incomes

			Pareto (I)						Pareto (II)					
	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(σ)	χ	Estim. Income Share	Estim. Neg-income Gini	Mean Neg. Income	Semipar. Pareto (II) Gini
EG12[a]	60.44	53.00	0.321 (0.048)	0.26	–[c]	–	–	–	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)	–[d]	–	–	–
FR05[b]	30.37	32.99	0.930 (0.397)	0.03	–	–	–	–	10.492 (–)	−1.222 (–)	0.02	–	−16,209	33.04
FR10	61.43	34.10	0.345 (0.038)	0.20	–	–	–	–	8.626 (0.373)	0.576 (0.329)	0.08	40.44	−13,151	34.34
GR95	0.00	40.23	1.443 (–)	0.32	−0.001	0.530	−48	40.42	–	–	–	–	–	–
GR00	61.23	39.03	1.273 (1.060)	0.16	−0.001	0.647	−137	39.13	3.827 (0.889)	1.293 (0.834)	0.00	–	–	–
GR04	54.87	37.34	0.323 (0.038)	0.32	–	–	–	–	8.357 (0.47)	0.303 (0.408)	0.10	17.87	−6,117	37.68
GR07	64.71	36.08	0.160 (0.008)	0.56	–	–	–	–	9.351 (0.321)	0.598 (0.28)	0.70	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)	0.00	–	–	–
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
IL10[b]	20.67	40.91	0.452 (0.046)	0.20	–	–	–	–	8.329 (–)	−1.151 (–)	0.00	–	−1,927	–
IL12	90.13	38.95	1.500 (0.436)	0.55	−0.013	0.500	−3,926	39.3	7.649 (0.211)	0.836 (0.19)	0.04	71.84	−12,804	39.34
IL14	75.10	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)	0.50	–	–	–	–	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	51.90	38.53	0.769 (0.263)	0.11	–	–	–	–	9.185 (0.588)	0.542 (0.457)	0.11	37.16	−21,280	38.76
IT00[b]	3.58	37.07	12.220 (10.800)	0.01	−0.001	0.043	−2,886	37.08	8.428 (–)	−1.49 (–)	0.00	–	−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	–
IT14[b]	12.26	36.41	2.843 (2.170)	0.04	−0.004	0.213	−2,609	36.44	8.342 (–)	−1.405 (–)	0.00	–	−1,746	–
RS06	63.50	38.90	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)	0.60	–	–	–	–	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)	0.50	–	–	–	–	11.892 (0.362)	0.993 (0.357)	18.52	98.68	−22,000,000	64.84
ES95[b]	14.25	39.06	0.510 (0.020)	0.66	–	–	–	–	6.838 (–)	−1.413 (–)	0.02	–	−388	–
ES00	46.67	38.70	0.840 (0.185)	0.20	–	–	–	–	6.155 (0.386)	0.429 (0.249)	0.01	27.28	−826	38.84
ES04	36.15	36.41	2.534 (2.580)	0.02	0.000	0.246	−680	36.42	5.421 (0.963)	1.581 (0.778)	0.00	–	–	–

TABLE 4
CONTINUED

			Pareto (I)						Pareto (II)					
	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(σ)	χ	Estim. Income Share	Estim. Neg-income Gini	Mean Neg. Income	Semipar. Pareto (II) Gini
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 (0.010)	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 (LIS database) are provided.

[b] For negative incomes in five data sets (FR05, IL10, IT00, IT14, and 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 data sets, estimation of the Pareto (I) model yielded α coefficients < 1, implying wide dispersion with an undefined parametric mean. For these data sets, 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 data sets, estimation of the generalized Pareto (II) model yielded implausible pairs of coefficients σ, χ giving positive income shares and Gini coefficients outside of the unit interval. For these data sets, we omit parametric estimates of the income share and the Gini, as they would be outside of reasonable bounds.

the entire income distribution is thus nearly identical to the nonparametric Gini. We conclude that the generalized Pareto (II), being the most flexible with two parameters, outperforms Pareto (I), which outperforms the negative exponential distribution in terms of providing a meaningful correction for the potentially mismeasured lower tail.

4.3. *Imputation of Nonpositive Incomes with Random Forest*

Next we perform a random forest ensemble classification of positive income observations, to replace nonpositive incomes with the households' most likely positive values based on other households with similar observed characteristics. The intuition is that while we cannot trust the nonpositive incomes, we can rely on households' other characteristics for imputing the most likely positive income values 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 in the presence of categorical variables (Zhao *et al.*, 2017). One pitfall is the possibility of 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 three quantiles), urban/rural residence, house ownership, and household head's education (binaries: none, primary, secondary, and tertiary), self-perceived health (binaries: bad or very bad, fair, good, and very good), age and age squared (binaries for three quantiles).[14] These variables are selected for their ability to proxy for households' earning capacity or economic status, and their availability across the majority of data sets.

For each household, we identify the node (or positive income value) with the highest probability, and we replace nonpositive incomes in the data with these best-matched positive values. First, we do this for self-employment income only, as the most prevalent driver of negative *DHI*s. (Zero *HILS* are very common among households not engaged in self-employment, and therefore these values are not replaced.) We recalculate *DHI* for households that initially had nonpositive *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 nonpositive *DHI* and negative ($DHI - HILS - HICID - HITP$), using the best-matched positive values

[14] The algorithm is based on the 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 using the following command: `chaidforest X, unordered(hlth_fair, hlth_good, hlth_vgood, edu_prim edu_sec edu_ter ownhouse rural) xtile(age age2 nhhmem, nquantiles(3)) ntree(100).`

TABLE 5

Random Forest Imputation of Negative Incomes: Gini and Poverty HCR Estimates

		(1)				(2)				(3)		
		Random Forest Classification and Imputation of $HILS < 0$				Random Forest Classification and Imputation of $DHI - HILS - HICID - HITP \leq 0$				Random Forest Classification and 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	–[b]		–	–	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	36.90	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	42.10	13.41
IQ12	12	–		–	–	–		–	–	12	41.28	16.68
IL01	19	6	6	38.55	16.76	14	15	38.67	16.60	19	38.46	16.60
IL05	17	7	7	39.51	18.37	12	16	39.56	18.40	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	17.60	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	15.30	–		–	–	68	37.95	14.95
IT00	77	2	2	37.07	13.34	–		–	–	77	36.42	12.57**
IT04	20	4	4	36.62	11.40	18	20	36.49	11.25	20	36.47	11.22

TABLE 5
Continued

		(1)				(2)				(3)		
		Random Forest Classification and Imputation of $HILS < 0$				Random Forest Classification and Imputation of $DHI - HILS - HICID - HITP \leq 0$				Random Forest Classification and 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
IT08	39	–		–	–	–		–	–	39	35.67	11.20
IT10	48	1	1	35.40	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	15.60
RS13	82	47	47	38.98**	15.44	–		–	–	82	39.36*	15.89
RS16	86	38	38	38.88*	16.55	–		–	–	86	39.01	16.70
SI04	1	–		–	–	–		–	–	1	31.71	11.97
SI10	1	–		–	–	–		–	–	1	34.30	14.91
ES95	67	39	39	39.05	13.96	60	67	38.62*	13.70	67	38.46**	13.59
ES00	15	9	11	38.69	17.53	–		–	–	15	38.56	17.40
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**	14.50**	97	37.16***	14.28***
SD09	28	–		–	–	–		–	–	28	54.30	21.58

Note: Years refer to income-reference years. Surveys were harmonized by LIS and ERF. Data sets restricted to those containing nonpositive DHIs.

*Corrected estimate within 90 percent (**95 percent) confidence interval of uncorrected estimate.

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

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

(column 2 in Table 5). Finally, we repeat the classification exercise for *DHI* itself, and we replace nonpositive *DHI* with the best-matched positive values according to the node with the highest probability for the household (column 3 in Table 5).

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 nonpositive *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 in other data sets. The mean Gini correction across all data sets 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 Gini coefficients, 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 data sets is 0.48 points, and the median is 0.21 points.

Finally, using the random forest classification method on *DHI* itself and converting all nonpositive *DHI* into positive values yield typically larger downward corrections to the national Gini coefficients, 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 nonpositive *DHI*s were being recomputed using the new *HILS*. Here, nonpositive *DHI*s 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 nonnegative *DHI*s rose significantly above zero, while in column 2 the correction to all nonpositive *DHI*s put them just above zero.

Our poverty measure is also sensitive to the random forest corrections. The 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 the corrections in column 3 are significant for EG12, GR07, IT00, IT14, ES04, ES10, and ES13. In all these cases, the corrections produce lower poverty estimates. This is evidently due to negative incomes being reclassified into positive ones by the random forest classifier. For completeness, Figure 7 shows the time trends in the Gini coefficients 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.

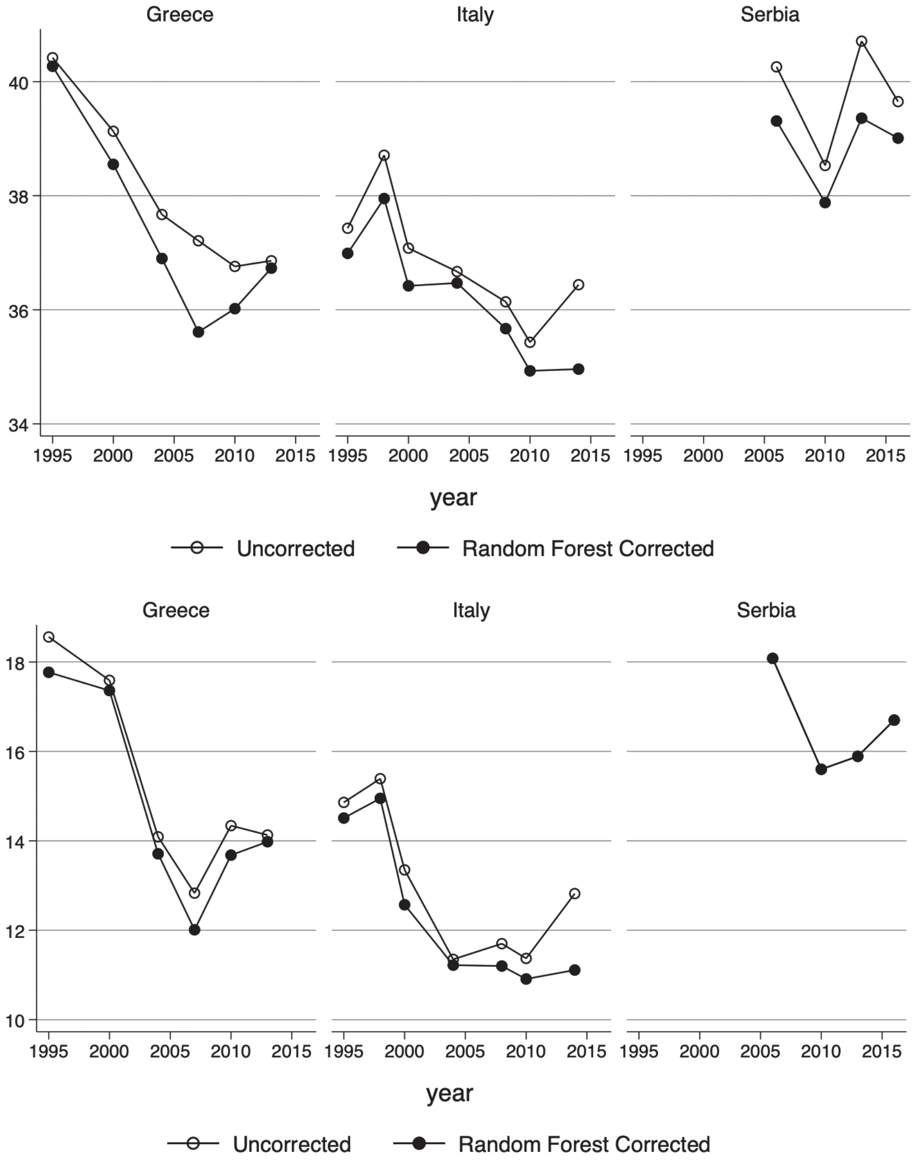


Figure 7. Inequality and Poverty on Uncorrected vs Random-Forest Corrected *DHI* Distribution: Variation Over Time

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+ data sets with corrected estimates. The corrected series in this figure are taken from Table 5, column 3

5. Discussion

This paper has reviewed the prevalence of negative and zero incomes in HI surveys, and their implications for the measurement of inequality and poverty. We have relied on 57 harmonized data sets from the LIS and Economic Research Forum databases, covering 12 Mediterranean countries over the period 1995–2016. We have found 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 withholding (France, Greece, Israel, and Spain), and unduly high self-paid social security contributions (Israel 2007–2010 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 well-being, 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 (in terms of health and educational attainment) are not clearly lower than those of their compatriots.

These findings point to an important observation: The income metric is not a good metric to measure household well-being or rank households when incomes are negative. Poverty and inequality measures are typically designed to capture household well-being, and household well-being can be measured with a variety of metrics of which income is only one. What we show in this paper is that, for at least some categories of households, well-being is not well measured by income and can be better assessed using alternative metrics such as consumption or expenditure. Indeed, in low- and middle-income countries and unlike in high income countries, poverty and inequality are almost invariably measured with consumption or expenditure, two measures that cannot be negative. Of course, low positive incomes that classify people as "poor" could also include persons who are not poor, like very rich people who report low income for tax purposes. The question here is where to draw the line between incomes that proxy well-being well and those that do not. This paper does not address this question because our focus is on negative and zero incomes which are either excluded from analyses by scholars or bottom coded by statistical agencies. Our aim is to improve on these traditional practices.

To correct income distributions for the unreliable nonpositive incomes, we have moved beyond traditional replacing through bottom-coding or reweighting through truncation by proposing the following two advanced replacing methods: the recently promulgated approaches of replacing parametrically extreme income observations with smooth distributions; and nonparametric imputation using random forest classification of incomes. The results of these estimations are summarized in Tables 3–5. We find that the traditional approaches produce nontrivial 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 rather weak, possibly because of the poor fit with the evaluated distribution functions, and because they are restricted to correcting 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 an unrealistically large dispersion among negative incomes. The two-parameter generalized Pareto distribution fits better, giving rise to realistic parametric means and Gini coefficients for negative incomes, but still yields trivial corrections to the overall Gini coefficients, 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; therefore, 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 nonpositive incomes, as it produces a continuous distribution of overall incomes without a point-density at zero, and converts nonpositive incomes into realistic positive values based on households' observed characteristics. This imputation has shown sensible results across multiple countries and multiple model specifications, and lowers the estimated Gini by up to 1.7 percentage points. It should be remarked that this approach can create issues when one works with panel data. For example, suppose that the negative income of a person in time *t* is replaced with an income above the poverty line, and that in time *t+1* this same person has a nonnegative income below the poverty line. This would be an individual who is classified as non-poor at time *t* and poor at time *t+1*. In this case, this person could have effectively moved from non-poor to poor, or it could simply be that our proposed method has incorrectly classified this person at time *t*. It is important to consider these types of issues if one is working with panel data.

These estimations, conducted under rather conservative assumptions and modeling specifications, suggest that the poverty-identification and inequality-measurement problems posed by negative and zero incomes are not trivial, and deserve attention and careful modeling by practitioners. In relation to the "static" problem of nonpositive incomes, our corrections produce more accurate inequality and poverty indexes for the majority of countries. However, in relation to the "dynamic" problem of nonpositive 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 very 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, because 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. As the problems of poverty and unequal economic opportunities in the region have been linked to civil discontent and uprisings, a better understanding of the scale of these problems can give policymakers the tools to bring political instability down, and even fix some traps and obstacles to economic growth.

References

Athey, S., and G. Imbens, *Machine Learning Methods Economists Should Know About*. 2019. arXiv, 1903.10075.

Atkinson, A. B., T. Piketty, and E. Saez, "Top Incomes in the Long Run of History," *Journal of Economic Literature*, 49, 3–71, 2011.

Blundell, R., and M. Costa Dias, "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 44, 565–640, 2009.

Breiman, L., "Random Forests," *Machine Learning*, 45, 5–32, 2001.

Brewer, M., B. Etheridge, and C. O'Dea, "Why are Households that Report the Lowest Incomes So Well-Off?" *The Economic Journal*, 127, F24–49, 2017.

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

Cowell, F. A., and E. Flachaire, "Income Distribution and Inequality Measurement: The Problem of Extreme Values," *Journal of Econometrics*, 141, 1044–72, 2007.

Cowell, F. A., and M. P. Victoria-Feser, "Income Distribution and Inequality Measurement: TheProblem of Extreme Values," *European Economic Review*, 40, 1761–71, 1996.

Dagum, C., "A Model of Net Wealth Distribution Specified for Negative, Null and Positive Wealth, a Case Study: Italy." in Dagum, C., and M. Zenga, (eds), *Income and Wealth Distribution, Inequality and Poverty*. Springer,Berlin, 42–56, 1990.

______, "A Study on the Distributions of Income, Wealth and Human Capital," *Revue Europeenne des Sciences Sociales*, 37, 231–68, 1999.

Davis, H. T., *The Theory of Econometrics*. Principia Press,Bloomington, 1941.

De Battisti, F., F. Porro, and A. Vernizzi, "The Gini Coefficient and the Case of Negative Values," *Electronic Journal of Applied Statistical Analysis*, 12, 85–107, 2019.

Eurostat, "Current Treatment of Taxes and their Implication on Negative Income and on Comparability Between Countries." *EU-SILC Documents TFMC-1*. European Commission, 2005.

______, "Self-Employment Income." *EU-SILC Documents TFMC-02/06*. European Commission, 2006a.

______, "Some Proposals on the Treatment of Negative Incomes." *EU-SILC Documents TFMC-15/06*. European Commission, 2006b.

Fitzpatrick, C. A., P. Bull, and O. Dupriez, *Machine Learning for Poverty Prediction: A Comparative Assessment of Classification Algorithms*. Technical Report. World Bank Knowledge for Change (KCP) Program, 2018.

Foster, J., J. Greer, and E. Thorbecke, "A Class of Decomposable Poverty Measures," *Econometrica*, 52, 761–66, 1984.

______, "The Foster-Greer-Thorbecke (FGT) Poverty Measures: 25 Years Later," *Journal of Economic Inequality*, 8, 491–524, 2010.

Haziza, D., and J.F. Beaumont, "On the Construction of Imputation Classes in Surveys," *International Statistical Review*, 75, 25–43, 2007.

Hlasny, V., "Top Expenditure Distribution in Arab Countries and the Inequality Puzzle," *Journal of Economic and Social Measurement*, 44, 177–201, 2020.

______, "Parametric Representation of the Upper Tail of Income Distributions: Options, Historical Evidence and Model Selection," *Journal of Economic Surveys*, 35, 1–25, 2021.

Hlasny, V., and P. Verme, “The Impact of Top Incomes Biases on the Measurement of Inequality in the United States.” *Technical Report 452*. Ecineq Working Paper, 2018a.

______, “Top Incomes and the Measurement of Inequality in Egypt,” *World Bank Economic Review*, 32, 428–55, 2018b.

Jäntti, M., E. Sierminska, and P. V. Kerm, “Modelling the Joint Distribution of Income and Wealth.” *Discussion Paper 9190*. IZA, 2015.

Jenkins, S. P., and M. Jäntti, “Methods for Summarizing and Comparing Wealth Distributions.” *Working Paper 2005-05*. ISER, 2005.

Jenkins, S. P., R. V. Burkhauser, S. Feng, and J. Larrimore, “Measuring Inequality Using Censored Data: A Multiple—Imputation Approach to Estimation and Inference,” *Journal of the Royal Statistical Society Series A*, 174, 63–81, 2011.

Luchman, J. N., “Random Forest Ensemble Classification Based on Chi-square Automated Interaction Detection (Chaid) as Base Learner.” *Statistical Software Components S457932*. Boston College Department of Economics, 2015.

Ostasiewicz, K., and A. Vernizzi, “Decomposition and Normalization of Absolute Differences, When Positive and Negative Values are Considered: Applications to the Gini Coefficient,” *Quantitative Methods in Economics*, 18, 472–91, 2017.

Scott, C. D., and J. A. Litchfield, “Inequality, Mobility and the Determinants of Income among the Rural Poor in Chile, 1968–1986.” *STICERD Discussion Paper 53*. London School of Economics, 1994.

Stich, A., “Inequality and Negative Income,” *Journal of the Italian Statistical Society*, 5, 297–305, 1996.

Van Kerm, P., “Extreme Incomes and the Estimation of Poverty and Inequality Indicators from EU-SILC.” *Working Paper 2007-01*. CEPS-Instead IRISS, 2007.

Victoria-Feser, M. P., and E. Ronchetti, “Robust Methods for Personal Income Distribution Models,” *Canadian Journal of Statistics*, 22, 247–58, 1994.

Zabala, F., “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, 2015.

Zhao, P., X. Su, T. Ge, and J. Fan, “Propensity Score and Proximity Matching Using Random Forest,” *Contemporary Clinical Trials*, 47, 85–92, 2017.