IARIW 2021
Monday 23 – Friday 27 August
Statistisk sentralbyrå
Statistics Norway

Revisiting Income Inequality in Greece: A Tale of Two Tails

Apostolos Fasianos

(Brunel University London)

Chrysa Leventi

(Greek Ministry of Finance)

Fidel Picos

(Joint Research Centre, European Commission)

Andreas Thiemann

(Joint Research Centre, European Commission)

Paper prepared for the 36th IARIW Virtual General Conference
August 23-27, 2021
Theme 13: Other

Revisiting income inequality in Greece: a tale of two tails

Apostolos Fasianos*[1], Chrysa Leventi[2], Fidel Picos[3], and Andreas Thiemann[3]

[1]Brunel University London
[2]Council of Economic Advisors, Greek Ministry of Finance
[3]JRC, European Commission

July 30, 2021

Work in progress: please do not circulate or cite without permission!

Abstract

This paper attempts the first systematic comparison of Greek inequality estimates derived from administrative data and household survey data from the Greek SILC survey. The administrative data we employ contain more than 650 thousand tax returns and carry the unique feature of including the full top 1 percent of Greek taxpayers. The aim of the paper is threefold: first, we perform a systematic comparison of the two data sources. Second, we propose different adjustments for the top tail, using both survey and administrative data. We replace the top tail from SILC with the corresponding one from the tax data and perform a Pareto top tail adjustment. Our preliminary distributional estimates reveal substantially higher levels of income inequality in the adjusted sample. Finally, we illustrate the income profile of the richest 1 percent of the Greek population.
Keywords : income inequality, administrative data, survey data

*Corresponding author: apostolos.fasianos@brunel.ac.uk. Disclaimer: The views expressed are those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission or the Council of Economic Advisors of the Greek Ministry of Finance.

1 Introduction

Survey and income tax data are two frequently used sources of information for research on inequality and the measurement of top incomes. A problem with measuring inequality using only household surveys, is that they under represent the richest individuals or households, standing at the right tail of the income distribution. This undercoverage of the rich masks the true level of inequality and often leads to misguided policy conclusions.

This paper makes use of a large representative sample of personal income tax returns provided by the Greek Ministry of Finance. The sample contains more than 650 thousand tax returns submitted to the authorities in 2018 and has the unique characteristic that it also contains the full top 1% of taxpayers. Using both the Eurostat and the Greek versions of the European Union's Statistics on Income and Living Conditions[1], we derive income concepts and units of analysis that are as comparable as possible to the tax data. We are thus able to estimate measures of income concentration at the bottom, middle and top parts of the SILC distribution and systematically compare them to those derived from the income tax returns. Income distribution can be either measured by survey or tax administration data, with each source carrying different advantages and disadvantages. The most prominent advantage of survey data is that they are constructed from a random sample of households, which is an economically-meaningful unit of observation (Bricker et al., 2016). Furthermore, they include a plurality of variables, which are originally designed to fit particular research aims, such as income conditions, consumption patterns, financial behaviour, etc. Yet, a serious caveat of household surveys when employed for distributional analysis, is their failure to capture the richest individuals or households, standing at the right tail of the income distribution. Tax administrative data, on the other hand, provide universal coverage at the top of the distributions, as tax filing is compulsory for the extremely rich. However, tax data may not capture well those at the bottom of the distribution who are not obliged to file a tax declaration, while their unit of analysis is the tax unit, which may not be suitable for economic

[1] EU-SILC UDB and PDB respectively

behaviour analysis (Bricker et al., 2016). In general, it is very difficult to obtain precise information on the top of the distribution from small-scale voluntary surveys. Furthermore, the non-response bias tends to increase with income, as the literature shows (Kennickell and McManus, 1993).

Several studies have shown that high income individuals or households are underrepresented in household surveys. For the US case, Atkinson et al. (2011); Burkhauser et al. (2012), suggests the share of total income held by the top 1% in surveys is substantially lower than when estimated from tax return data. For the UK, Jenkins (2017) suggests that the 99.5 centile's income in the UK household survey, can be as low as 77% the corresponding one in administrative tax data. The undercoverage is found substantially higher in emerging markets. For Colombia, the average income of the top 1% in tax data is found 50% larger than in surveys (Alvaredo and Vélez, 2013).

To overcome the problem of the missing rich at the top of income distribution one can rely on functional form assumptions regarding the top tail. Most studies apply a Pareto distribution when estimating top income or wealth concentration. Nevertheless, even when relying on the Pareto assumption for the top of the distribution, the problem of missing (income) rich households in survey data remains (Bach et al., 2019). A possible solution is to rely on income tax data that better represent the top of the income distribution. While administrative tax data generally does not provide a lot of socio-demographic information (often the household context is missing), it often covers the entire distribution of taxable income, including the very top.

The aim of the paper is threefold: first, we perform a systematic comparison of the two data sources. Having performed the comparison for the entire income distribution, we propose different adjustments for the top tail using both SILC and administrative data. We replace the top tail from the SILC data with the corresponding one from the tax data and perform a Pareto top tail adjustment. Our preliminary distributional estimates reveal substantially higher levels of income inequality in the adjusted sample. Third, we illustrate the income

profile of the richest 1 percent of the Greek population.

Our analysis carries important implications for policy making as it reveal a more realistic dimension of the actual level of inequality in the country.

2 Data: description and comparisons

The two key datasets used in this study include administrative data and the Greek component of the SILC survey, both for the year 2017.

The administrative data refer to a large sample of unaudited income tax returns filed in 2018 (incomes earned in 2017) provided in an anonymised form by the Greek Ministry of Finance. The sample of tax returns covers approximately 1.1 million individuals in 480,000 households (10.2% of the country's population). The tax returns sample's basic unit is the tax-filer; for each tax-filer the total income is available together with its breakdown into separate income components. The tax-filer also reports the income of his/her spouse (with the same breakdown), the number of children and other household members. The administrative data comprise highly detailed and disaggregated information, providing us with more than 550 variables on market incomes (taxable and non-taxable), imputed incomes, as well as information on pensions and social benefits, certain real estate and financial assets, and some expenditures. More importantly, our tax dataset comprises the full richest 1% of tax-filers, a characteristic that is crucial for the analysis performed in this study. Despite its multiple advantages, as typical in administrative data sources, our tax returns dataset also carries some shortcomings. A key one is the lack of detailed socio-economic information (particularly labour-related), which prevents the user from carrying out a number of economic behaviour experiments or control for the effects of public policy on various population groups with specific characteristics.

As a standard data source for measuring income conditions using survey data, we draw from the Greek component of the European Union Statistics on Income and Living Conditions

(EU-SILC), carried out in 2018 reporting household incomes for 2017. The sample covers approximately 56,000 individuals living in 24,300 households. SILC is the benchmark survey used by the European Union to report the continent-wise levels of inequality and poverty. To these ends, the survey collects comparable and multidimensional micro-level data on income, social exclusion, housing, work, education, and health. This dataset was enriched with information from the national version of the Greek SILC (Production Database) on individuals' social insurance fund. This additional information allowed us to split self-employment income into farming and business income, which are the income categories that are found in the administrative data.

Meticulous work has been carried out to make the two samples comparable. In particular, we first ensure that income variables in the two datasets are consistently defined: in SILC, incomes are reported net of income tax and social insurance contributions; in tax returns, incomes are reported gross of income tax and net of social insurance contributions. As in the tax returns' sample income taxes are also available as separate variables, we use the latter to construct income variables that are net of both taxes and social insurance contributions, that can then be compared to SILC.

Moreover, the population in the tax data had to be slightly adjusted in order to align with the population represented in SILC: households containing at least one non-resident tax return (i.e. individuals living abroad) were removed from the sample. The same approach was followed for homeless individuals and individuals living in institutions, since these population categories are absent from SILC. After these adjustments, the weighted sample of the tax data contains 10,449,756 individuals in 448,7013 households, whereas the respective numbers for the SILC data are 10,455,382 individuals in 412,5263 households. The relatively larger number of households found in the administrative data can be partly explained by the occurrence of 'household splitting' (i.e. reporting a household composition different from the actual one for social benefit or taxation reasons). As has been noted in Marini et al. (2019), households -especially those located at the bottom end of the income distribution-

have a high incentive to split, so as to receive the highest possible amount from the country's flagship social assistance benefit (Social Solidarity Income).

Accounting for the above-mentioned adjustments, the following comparable income variables were created in both datasets: (a) Salaries; (b) Self-employment income; (c) Farming income; (d) Investment income; (e) Property income; (f) Public pensions; (g) Private pensions; (h) Unemployment benefits; (i) Family benefits; (j) Welfare benefits; (k) Other incomes. The issue of negative values in income sources (b) and (c) requires some special attention. In the tax data individuals can report losses from exercising agricultural of business activities stemming from previous years; this is not the case in SILC, where individuals are only asked to report any losses that occurred during the income year of the survey. This discrepancy results in having many more (and much larger) negatives in the tax data. For this reason, negative values are excluded from our analysis, except when otherwise stated.

Figure 1 depicts a first illustration of the differences in the distributions of the two data sources by plotting mean total net incomes by individual centile in the SILC and the tax administrative data. Although for the most part the two sample distributions look remarkably similar, mean total income for the top 1% in the survey data is estimated at more than 80,000 EUR per year, while the corresponding amount in the administrative tax data is estimated at a bit less than 70,000 EUR per year. Moreover, the left tails of the two distributions seem to diverge, with mean total incomes in SILC being higher than those depicted in the tax data. This divergence between the two extreme tails serves as an initial motivation for the adjustment of the survey data performed in the paper.

Figures 2 and 3 attempt to dig deeper into each of the income sources of interest, showing the way these are distributed by individual centile and by income group respectively. As expected, we observe differences in all income sources; however, the ones with the most diverging patterns are investment income (especially at the bottom of the distributions), self-employment income (especially at the middle/upper part of the distributions), farming, property and other income. The latter, which seems to be playing a very important role at

Figure 1: Mean total net income by individual centile

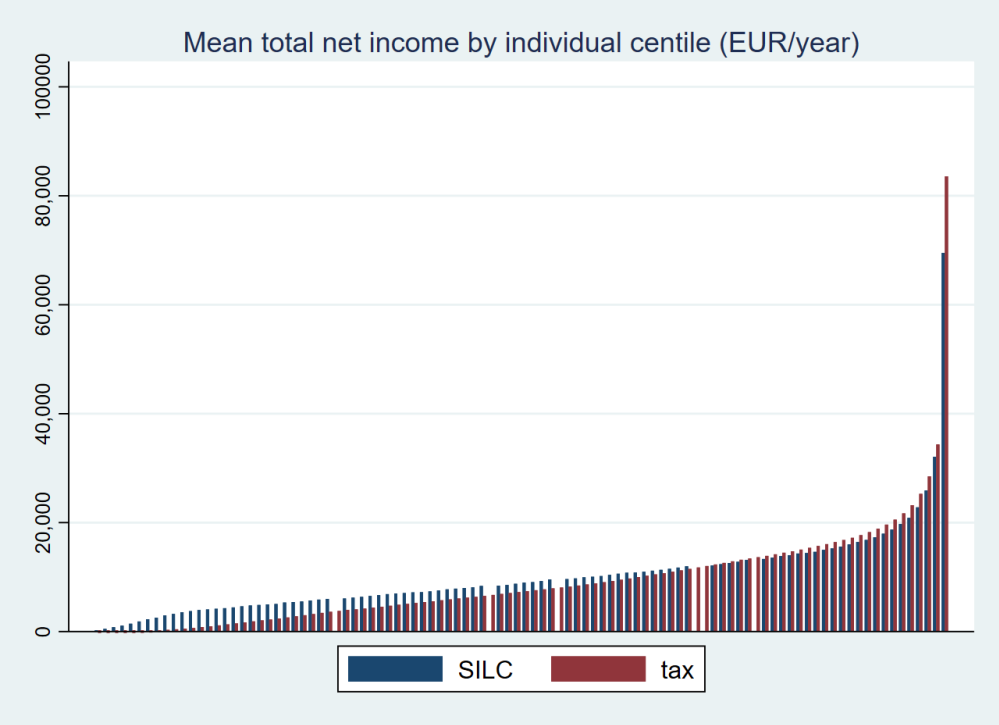


Figure 2: Income concepts by centile

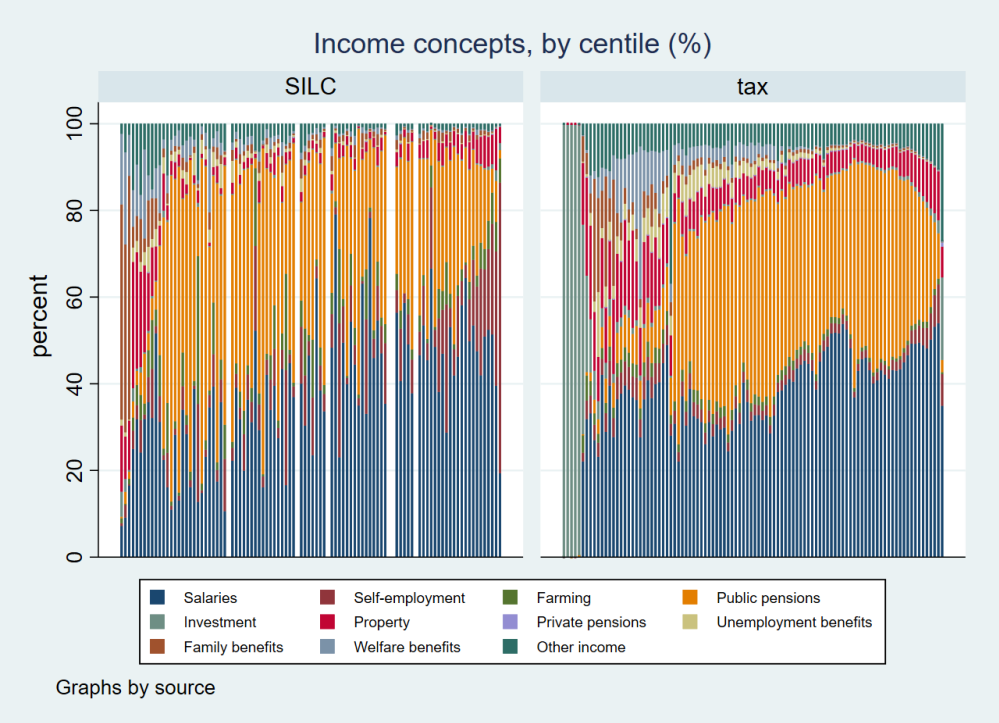
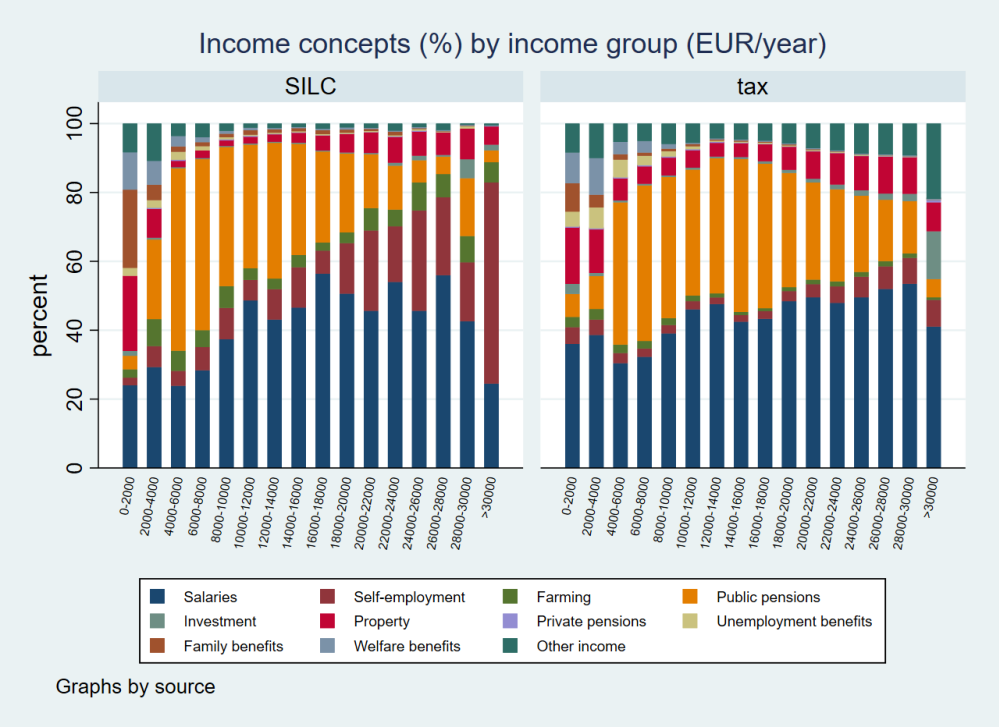


Figure 3: Income concepts by income group

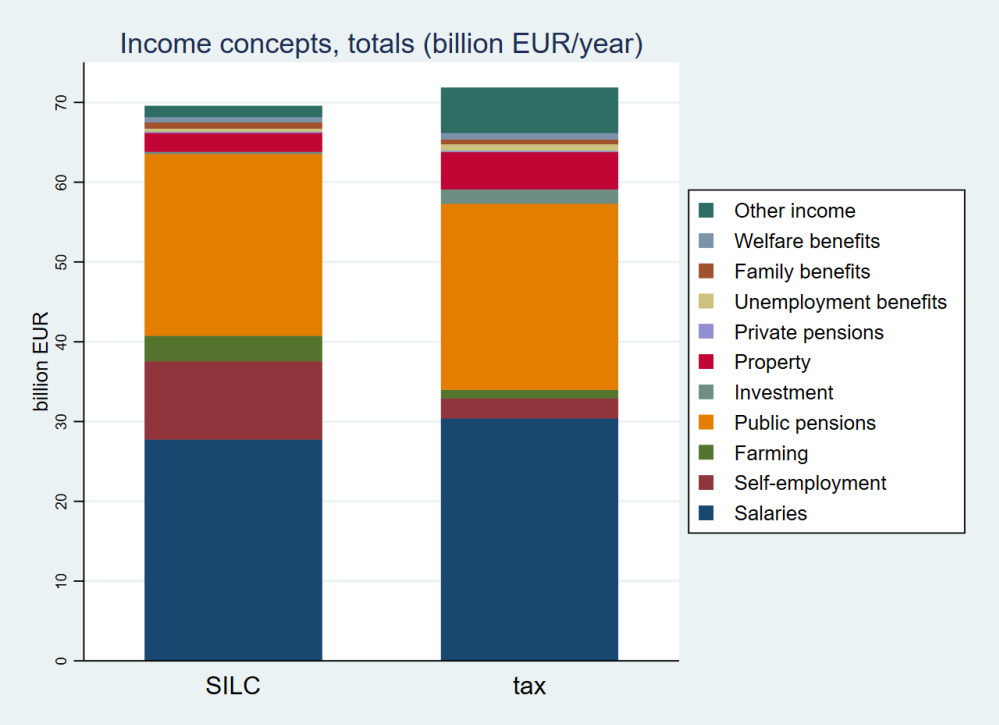


the top centile of the survey data, will be analysed in more detail at the next section of the paper. On the other hand, the distributions of salaries, public pensions, unemployment and welfare benefits depict more similar patterns.

Figure 4 seems to confirm what was depicted in the two previous graphs. In aggregate terms, our two data sources are very close with respect to total income, with yearly amounts for 2017 summing up to approximately 70 billion EUR. In SILC, however, self-employment and farming income sum up to 13 billion EUR, approximately 3.5 times higher than what is depicted in the tax data.

The plausible reasons for this discrepancy are manifold. First, the presence of tax evasion. In Greece, tax evasion is known to be rife (Artavanis et al., 2016; Leventi et al., 2013). The relevant research confirms that income under-reporting to the tax authorities is much more pronounced in the case of farming and self-employment income, and it is mostly located at the tails of the distributions. The 2017 personal income tax schedule, under which self-employment income was taxed at a non-negligible 22% from the first euro earned, is believed

Figure 4: Income concepts: totals



to have further exacerbated this behaviour. Second, the large discrepancies between our two data sources might be also related, besides the already mentioned losses from previous years declared in the tax data, to the way individuals perceive the relevant question on self-employment and farming income in SILC (i.e. 'annual profit or loss from business or activity after the deduction of business expenses'), as well as to the way accountants, tax authorities and individuals classify these incomes in each data source. For example, depreciation allowances are deducted in the tax data, while individuals may ignore them when reporting in SILC. Also, self-employed individuals working for up to three clients are allowed to declare this income as 'employment income' to the tax authorities. Finally, as it is generally the case in surveys, small/irregular amounts of income tend not to be reported in SILC; on the contrary, these amounts are duly deported in the tax data.

Figures 5 and 7 depict the number of individuals with positive incomes for each of the 11 income sources in question, as well as the mean value of those incomes. In order to shed light to the issue of non-reporting of small/irregular amounts in SILC, these figures are also

presented for a restricted sample of individuals earning more than 10 EUR/month (Figures 6 and 8).

The most striking result of these comparisons concerns investment income. As can be seen in Figure 5, the number of individuals in receipt of investment income in the tax data reaches almost 5.5 million people; in SILC, their respective number slightly exceeds 300 thousand individuals. This picture changes drastically when we only account for individuals receiving more than 10 EUR/month (Figure 6). The number of investment income recipients in the tax data is reduced to approximately 1 million individuals. As expected, the average yearly investment income amounts depicted in the tax data are also increased significantly once the sample is bottom-coded. The reason behind this discrepancy lies in the way this income source is reported in the two datasets. In SILC, investment income is self-reported by interviewees; in the tax data, this information is directly provided by banks, and hence, it also includes the very small amounts of interest income from bank deposits.

Apart from investment income, bottom-coding significantly affects farming and self-employment income. Again, it becomes obvious that individuals fail to report small (or one-off) remunerations in SILC. On the contrary, such amounts are an integral part of tax declarations.

As expected, all the discrepancies analysed so far have an impact on overall inequality. The first row in Figure 9 shows a strikingly higher value of the Gini coefficient in tax data (0.4488) than in SILC (0.3214). However, what would be the distributional impact of making the left tails of the two distributions more comparable? We try to reply to this question by calculating Gini coefficients after posing several comparability-enhancing restrictions to the left tails of our samples. Our results suggest that the only restriction that has an impact on our selected inequality measure is the exclusion of negatives from the tax data, reducing the Gini coefficient from 0.4488 to 0.4078. Restricting the samples to individuals reporting more than 10 to 50 EUR/month has a close-to-zero impact. As most negative values in the tax data are caused by self-employment and farming income losses, which we are not able to disentangle into those that occurred in 2017 or in previous years, we believe that attempting

Figure 5: Number of individuals with positive values

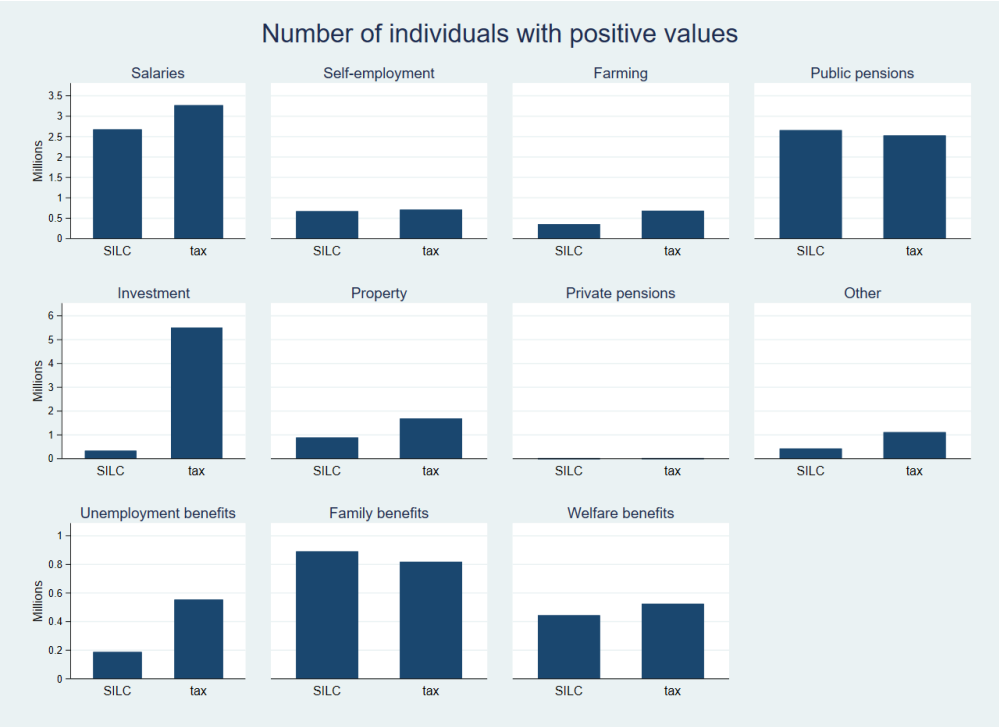


Figure 6: Number of individuals with positive values: restricting samples to individuals earning more than 10 EUR/month

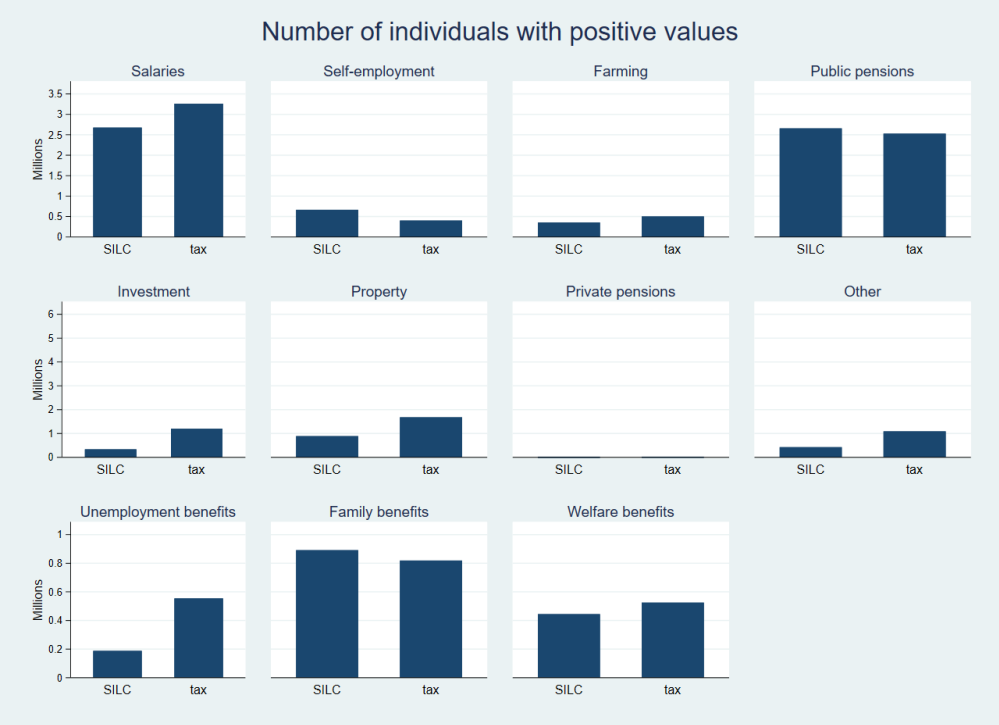


Figure 7: Means of positive values

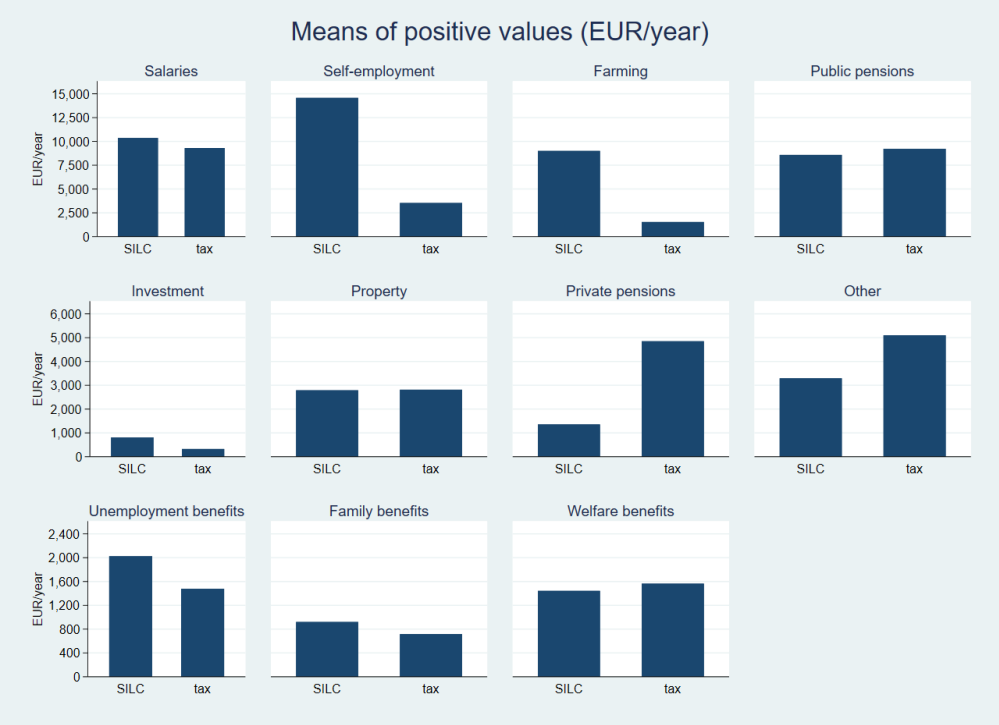


Figure 8: Means of positive values: restricting samples to individuals earning more than 10 EUR/month

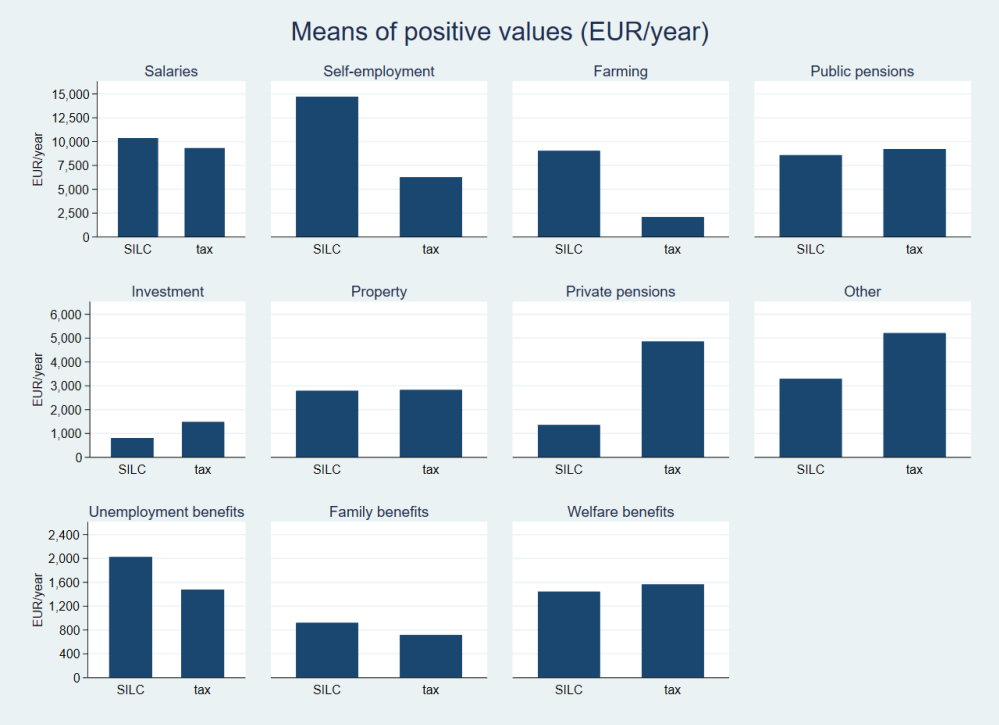


Figure 9: Gini indices

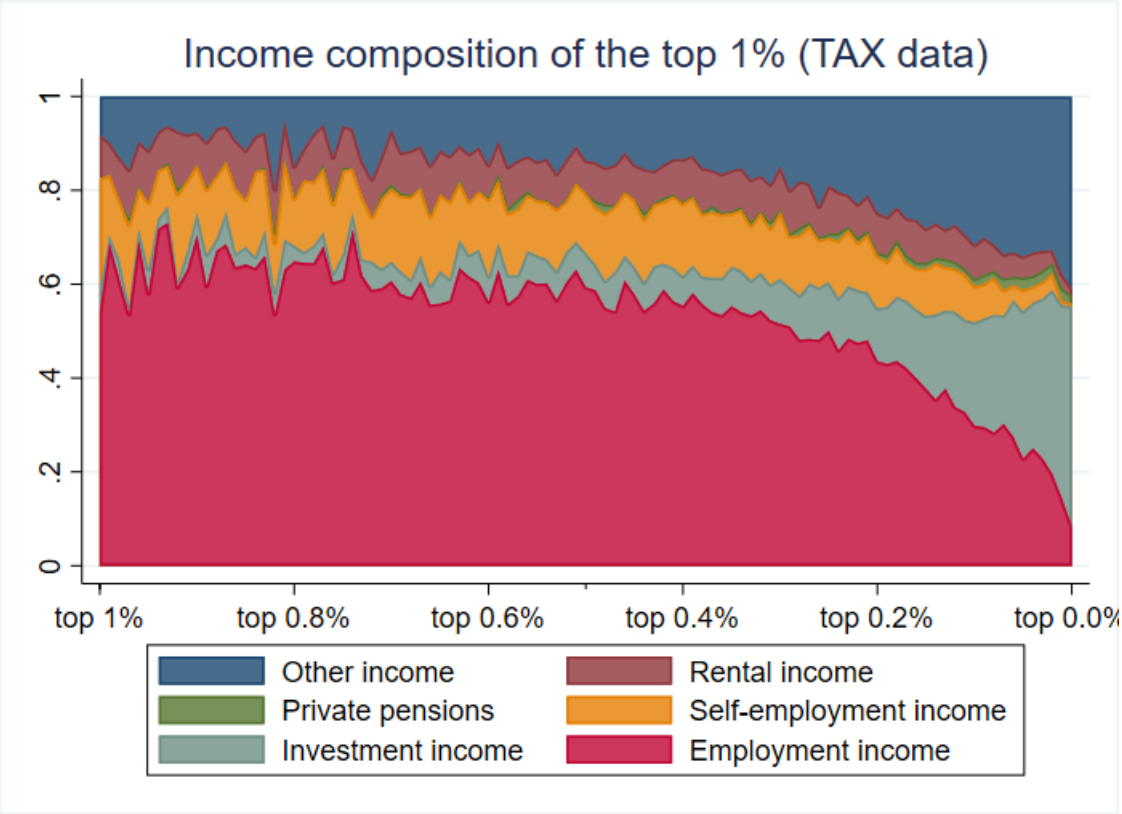
Total net equivalised income	SILC data	tax data
original values	0.3214	0.4488
…after setting negative values to zero	0.3213	0.4078
…after setting values less than 10 EUR/month to zero	0.3213	0.4079
…after setting values less than 20 EUR/month to zero	0.3213	0.4080
…after setting values less than 30 EUR/month to zero	0.3215	0.4083
…after setting values less than 40 EUR/month to zero	0.3215	0.4086
…after setting values less than 50 EUR/month to zero	0.3218	0.4090

to impute these values in SILC would not be an improvement for the survey data. Imputing the missing (very) small amounts of (mostly investment) income would move SILC data closer to reality but it would have a negligible impact on the Gini coefficient. Hence, we are now ready to turn our attention to the top tail of the two distributions.

3 Exploring the income distribution of the top 1%

One of the most striking and debated issues of the inequality literature is the role of the super-rich in driving the overall levels of income and wealth disparities (see, for example, Kopczuk and Saez (2004)). In this context, a lively research interest on the profile of the super-rich has emerged, focusing on the socio-economic behaviour and characteristics of those standing at the highest centile of the income distribution. As pointed by Roine et al. (2009), the top 1% and the top 10% comprise substantially different types of individuals: while the former concentrates individuals receiving large shares of capital income, the latter contains high labor income earners. Yet, due to serious limitations on rich individuals, as also discussed earlier, only a few studies managed to provide comprehensive results, and those studies are confined to the case of the US. For example, (Bakija et al., 2012), using data on US individual income tax returns, demonstrate that executives, managers, supervisors, and financial professionals account for about 60 percent of the top 0.1 percent of income earners. More recently, Smith et al. (2019) identifying business income by linking tax administrative data with firm level data, examined the relative importance of human and financial capital

Figure 10: Top1 % of the income distribution in tax administrative data



in shaping the distribution of the super-rich. Using our information on the full top 1% of Greek income earners, we are able to track the distribution of the income components of this group.

Figure 10 shows the income composition of the top 1 percent of Greek taxpayers based on their 2018 tax declarations. While employment income is the main income component for most taxpayers, its relative importance decreases substantially at the very top. In particular, among the top 0.1 percent of taxpayers investment income and other income are the main sources of income. Other income contains more than 20 sub-components, varying from voting compensations and compensations due to termination of employment, to interest from treasury bonds or treasury bills and profits from the transfer of listed securities. Investment income comprises of domestic and foreign income from dividends, interest and loyalties. As a next step, we intend to look into the sub-components of each of these income sources, to

identify the main driving factors of this finding.

4 Top tail adjustment

Inspired by Jenkins (2017) among other papers and considering the data we have, we perform the following adjustments on the survey data. In the first approach, we swap the SILC income top tail by the tax data-based top tail. In the second approach, we replace the SILC top tail according to the estimated Pareto distribution.

4.1 Top tail swap

The top tail swap, our first approach, directly replaces the highest incomes in the survey with the corresponding observations for the 1 percent in the tax return data. In this approach, we use the top tail of the income tax data, however without performing any adjustment in the lower part of the income distribution. Although this approach allows us to get a more realistic estimate of overall income inequality, than the one estimated using only survey data, its biggest caveat is that we miss the richness of socio-demographic characteristics present for top incomes. This is because we can only retain information reported by the tax collecting agency, which is much more limited than that of the SILC survey. No distributional assumptions are made in this approach.

4.2 Pareto adjustment

Our second approach replaces the top of the income distribution, based on SILC, by a Pareto distribution which is estimated exploiting both SILC and tax data. In our methodological approach, we follow closely Bach et al. (2019) who perform a similar estimate for net wealth in selected EU countries.

4.2.1 Theoretical background

Our approach relies on the Pareto distribution to adjust the top tail of SILC-based income distribution. The approach is commonly accepted to provide a good fit.[2] The Pareto distribution can be defined for any level of income higher than a certain threshold, y_{min}, formalized by its complementary cumulative distribution function (ccdf)

$$P(Y > y_i) = (\frac{y_{min}}{y_i})^{\alpha}; \forall\, y_i \geq y_{min} \tag{1}$$

The ccdf, shown in Equation 1, illustrates the relationship between household i's income y_i, the threshold y_{min}, and the Pareto coefficient α. It represents the probability of earning y_i or more, defined on the interval $[y_{min}, \infty[$. The α coefficient shows the degree of concentration in the top tail: the smaller α, the larger the overall income concentration.

By applying the Zipf's law, we express the ccdf in terms of each household's ranking in the top tail being above y_{min} (see for instance Vermeulen, 2017). In particular,

We then assign the rank one to the income-richest household and the lowest rank n to income-poorest household in the top tail. $n(y_i)$ reflects the individual household rank of observation i:

$$\frac{n(y_i)}{n} \cong (\frac{y_{min}}{y_i})^{\alpha}; \; y_i \geq y_{min} \tag{2}$$

Next, we are able to approximate the Pareto distribution by the ranking of the sample households, taking the assumption that the sample is sufficiently large to make such an approximation. After taking the logarithm and re-arranging, we get:

$$ln(i) = C - \alpha ln(y_i) \tag{3}$$

[2]For example, the following papers employ a similar approach: Dalitz (2016); Vermeulen (2017); Cowell (2011); Gabaix (2009); Gabaix and Ibragimov (2012); Clauset et al. (2009); Kleiber and Kotz (2003); Chakraborty and Waltl (2018).

with $C = ln(n) + \alpha ln(y_{min})$.

As pointed by Gabaix and Ibragimov (2012), the log-log-rank-size regressions are biased in finite samples. We follow the suggestion and shift the rank by 0.5. Furthermore, we follow Vermeulen (2017); Bach et al. (2019) adjust the setting to incorporate survey weights and end up with the following relationship:

$$ln((i - \frac{1}{2})\frac{\overline{N_{fi}}}{\bar{N}}) = C^* - \alpha ln(y_i) \tag{4}$$

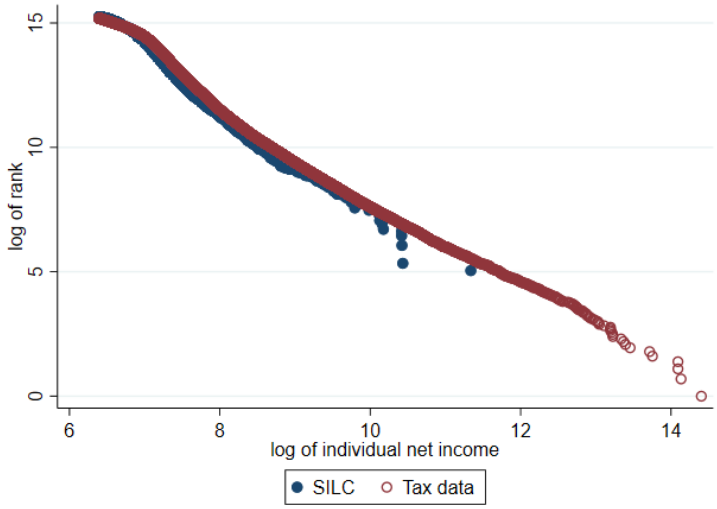
where the person with the highest individual net income, $i = 1$, has a survey weight of N_1, the one with the second highest income a weight of N_2, etc. $\bar{N} = \frac{\sum_{j=1}^{n} N_j}{n} = \frac{N}{n}$ gives the average survey weight of all observations, N is the total of all survey weights in the income top tail, $\overline{N_{fi}} = \frac{\sum_{j=1}^{i} N_j}{i}$, the average weight of the first i observations, and $C^* = ln(\frac{\bar{N}}{N}) + \alpha\, ln(y_{min})$. We derive α by estimation Equation 4 using plain Ordinary Least Squares (OLS).

After determining y_{min} and estimating α, we follow Bach et al. (2019, online appendix, section 5) and construct synthetic observations to represent the top tail and to assess the impact on distributional statistics. In particular, we replace all individuals in the SILC data above the chose income threshold y_{min} by new synthetic observations according to the estimated Pareto income top tail distribution. While the estimated Pareto function is continuous our synthetic observations will provide a discrete representation of the Pareto top tail. We therefore distribute the synthetic observations such that they match total income according to the continuous distribution function along the distribution.

4.2.2 Estimating the shape of the Pareto distribution

To estimate the shape parameter α, we have to choose how to best combine the available data, i.e. EU-SILC and the tax data. Before doing so, we visually investigate whether the income distribution resembles the characteristics of the Pareto distribution, Figure 11 plots the log-log-rank-income plot relationship for monthly individual incomes above 600 EUR for the two data sources.

Figure 11: Log rank - log income plot

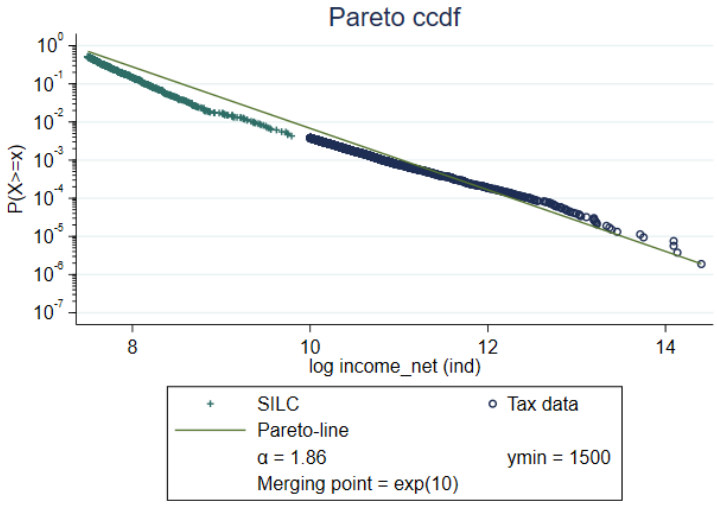


We see the linear relationship between the logarithmic representation of an individual's rank and the logarithm of here income, both for SILC and the tax data. The plot also shows that the tax data covers much better the very top of the distribution since there are several taxpayers reporting a higher income than the richest person in SILC, which underlines the benefit of relying not only on survey data but also on tax data to obtain a good coverage of the entire income distribution.

We have to determine the optimal lower lower bound, y_{min}, before estimating α. Here, we face a clear trade-off: When we choose an income level, which is too low, we might end up with observations that do not follow a Pareto distribution. However, if we choose a income threshold, which is too high, then we might end up with a small number of observations to estimate the Pareto shape parameter.

Figure 12 illustrates graphically the estimation of the Pareto coefficient α, when choosing the (monthly) income threshold of y_{min} equal to 1,500 EUR. The sample used to estimate α, which is equal to the negative slope of the Pareto line, consists of individuals with a monthly net income of 1,500 EUR or more but less than 13,330 ($= e^{10}$) EUR based on EU-SILC. Above that merging point, we replace SILC observations by the ones from the tax data, keeping total weights in the top tail constant according to the SILC data.

Figure 12: Cumulative complementary distribution function (ccdf)



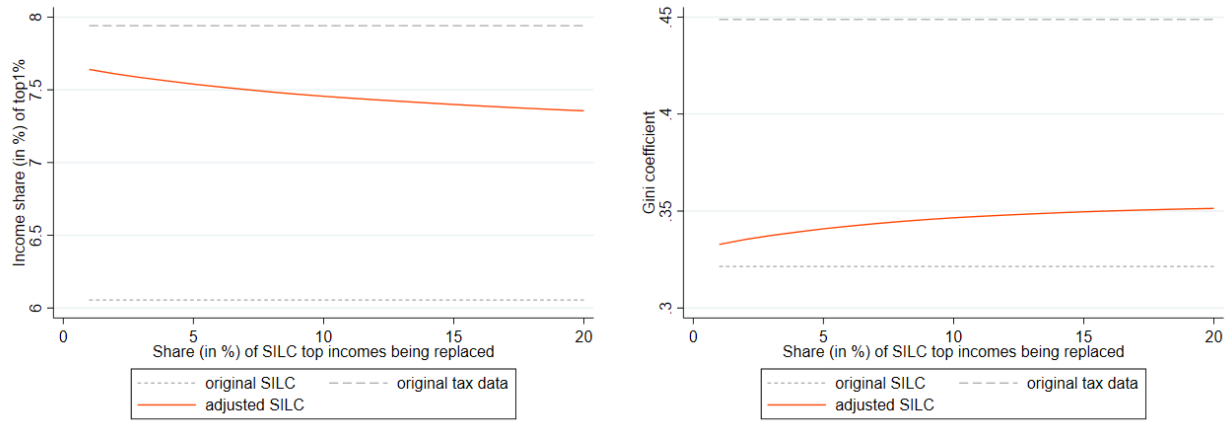
In the next version of this paper, we will provide a systematic analysis of how to best making use of the different data sets and finding the optimal level of y_{min}. The preliminary findings suggest that the estimated α coefficient depends more on the choice of y_{min} than the choice of the merging point. Figure A1 illustrates the α estimation for alternative merging points keeping y_{min} constant. In finding the optimal level of y_{min}, we plan to determine the optimal level of y_{min} relying visual inspection (Bach et al., 2019) and Goodness-of-fit tests (Dalitz, 2016; Krenek and Schratzenstaller, 2017).

5 Preliminary Results

This section sheds light on the impact of the two top tail adjustment alternatives. Figure 13 illustrates the distributional impact of replacing the top tail of the SILC (individual equivalised) income distribution by the corresponding one from the tax data. The left figure (13a) shows the impact of the top tail swap on the share of total income held by the top 1% by the size of the SILC top tail, which is swapped by the corresponding one from the tax data.[3]

[3]We replace exactly as many weighted observations from the original SILC data such that the total number of weights is kept constant. However, when replacing the SILC top tail by the tax data simply based on an income threshold the results are very similar.

Figure 13: Impact of the top tail swap on inequality by size of the top tail, being replaced



(a) Impact on the top1% income share

(b) Impact on the Gini coefficient

Note: The horizontal axis reports the share of the SILC-based income distribution, which is replaced by the corresponding tax data top tail. For instance, at 10%, the richest 10% in SILC are replaced by the richest 10% from the tax data, maintaining total weights constant according to the SILC sample. Calculations based on equivalised individual disposable income.

Swapping the richest percent of the SILC income increases the income share of the top1% from 6% to 7.6%, which is a quite large increase of inequality by 26%. Increasing the size of the top tail being replaced, i.e. going further to the right on the horizontal axis, implies two different channels. First, we replace a larger share of the SILC income distribution which is likely to increase inequality. Secondly, however the larger the SILC top tail being replaced, the larger is the impact of the top income distribution within the tax data top tail, which might have an ambigous effect on the after-adjustment income concentration. As a result, the top1% income share increases slightly to about 7.4% when replacing the top quintile of the SILC income distribution, still substantially higher than according to the original SILC data.

The right figure (13b) illustrates the corresponding impact on the Gini coefficient. When performing the top tail swap, the Gini coefficient increases from 0.321 (original SILC) to 0.333, replacing the richest percent of the SILC income distribution. In contrast to the left figure, however, the Gini coefficient increases steadily with the size of the SILC top tail,

which is swapped for the corresponding tax data tail.

Still pending is the systematic analysis of how a Pareto top tail adjustment affects the adjusted income distribution compared to the one based on the original SILC data. The next version of the paper will provide a thorough assessment and robustness checks for different choices of y_{min} and different samples.

6 Conclusion

Survey and administrative data are frequently used as sources of information for research on income inequality. In this paper we attempt the first systematic comparison of Greek inequality estimates derived from these two data sources.

Our administrative data come from a large representative sample of 2018 personal income tax returns (2017 incomes), provided by the Greek Ministry of Finance. The sample, which contains approximately 1.1 million individuals, has the unique characteristic that it also contains the richest 1% of taxpayers. Our survey data come from the Greek component of the EU-SILC, carried out in 2018 (2017 incomes).

Careful work was carried out to make sure income concepts and units of analysis are consistently defined between the two data sources. Our comparisons suggest that, although for most of their part the two sample distributions look remarkably similar, the tails are the ones that diverge the most. The income sources with the most diverging patterns were found to be investment income (especially at the bottom of the distributions), self-employment income (especially at the middle/upper part of the distributions), farming, property and other income. Another striking result of these comparisons concerned investment income. The number of investment income recipients in the tax data reaches almost 5.5 million people; in SILC, their respective number slightly exceeds 300 thousand individuals. The reason for this discrepancy lies in the way this income source is reported in the two datasets. In the tax data, this information is directly provided by banks, and hence, it also includes the very

small amounts of interest income from bank deposits (i.e. less than 10 EUR/month), which are completely missing from SILC.

As expected, the above-mentioned discrepancies have an important impact on overall inequality; the Gini coefficient in the tax data is estimated to have a strikingly higher value than in SILC (0.449 vs 0.321).

We first attempt to quantify the impact of making the left tails of the two distributions more comparable, by posing several restrictions to the left tails of our samples. Our results suggest that the only restriction that has an impact on the Gini coefficient is the exclusion of negatives from the tax data, reducing it from 0.449 to 0.408. Restricting the two samples to individuals reporting more than 10 to 50 EUR/month has a close-to-zero impact on the Gini coefficient.

We then turn our attention to the left tails of the income distributions. We perform two alternative top-tail adjustments on the survey data. First, we swap the SILC income top tail by the tax data-based top tail. Second, we replace the SILC top tail according to the estimated Pareto distribution. Our preliminary estimates suggest that the level of income inequality increases substantially when the tail swapping approach is performed.

These results might not sound surprising if we consider the established finding of survey under-coverage in the right tail of the distribution. Yet, our paper is the first to estimate such figures for the case of Greece, whose main inequality estimates have been solely provided by survey data up to now. These estimates, complemented with the forthcoming systematic analysis on how a Pareto top tail adjustment affects the SILC data, should be informative to policy makers concerned with the level of income inequality, as they unmask a more realistic dimension of how income is distributed in Greece.

References

Alvaredo, F. and Vélez, J. L. (2013), High incomes and personal taxation in a developing economy: Colombia 1993-2010, Commitment to Equity (CEQ) Working Paper Series 12, Tulane University, Department of Economics.
URL: *https://ideas.repec.org/p/tul/ceqwps/12.html*

Artavanis, N., Morse, A. and Tsoutsoura, M. (2016), 'Measuring income tax evasion using bank credit: Evidence from greece', *The Quarterly Journal of Economics* **131**(2), 739–798.

Atkinson, A. B., Piketty, T. and Saez, E. (2011), 'Top incomes in the long run of history', *Journal of economic literature* **49**(1), 3–71.

Bach, S., Thiemann, A. and Zucco, A. (2019), 'Looking for the missing rich: Tracing the top tail of the wealth distribution', *International Tax and Public Finance* **26**(6), 1234–1258.

Bakija, J., Cole, A. and Heim, B. T. (2012), "jobs and income growth of top earners and the causes of changing income inequality: Evidence from u.s. tax return data.".

Bricker, J., Henriques, A., Krimmel, J. and Sabelhaus, J. (2016), 'Estimating top income and wealth shares: Sensitivity to data and methods', *American Economic Review* **106**(5), 641–45.

Burkhauser, R. V., Feng, S., Jenkins, S. P. and Larrimore, J. (2012), 'Recent trends in top income shares in the united states: reconciling estimates from march cps and irs tax return data', *Review of Economics and Statistics* **94**(2), 371–388.

Chakraborty, R. and Waltl, S. R. (2018), Missing the Wealthy in the HFCS: Micro Problems with Macro Implications. European Central Bank, Working Paper Series, 2163.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009), 'Power-law distributions in empirical data', *SIAM Review* **51**(4), 661–703.
URL: *http://dx.doi.org/10.1137/070710111*

Cowell, F. (2011), *Measuring inequality*, Oxford University Press.

Dalitz, C. (2016), Estimating wealth distribution: Top tail and inequality, Technischer Bericht Nr. 2016-01, Hochschule Niederrhein.

Gabaix, X. (2009), 'Power Laws in Economics and Finance', *Annual Review of Economics* **1.1**, 255–294.

Gabaix, X. and Ibragimov (2012), 'A simple way to improve the ols estimation of tail exponents', *Journal of Business & Economic Statistics* **29**(1), 24–39.

Jenkins, S. P. (2017), 'Pareto models, top incomes and recent trends in uk income inequality', *Economica* **84**(334), 261–289.

Kennickell, A. B. and McManus, D. A. (1993), Sampling for household financial characteristics using frame information on past income, *in* 'Proceedings of the Section on Survey Research Methods'.

Kleiber, C. and Kotz, S. (2003), *Statistical Size Distribution in Economics and Actuarial Sciences*, Wiley Interscience.

Kopczuk, W. and Saez, E. (2004), Top wealth shares in the united states: 1916-2000: Evidence from estate tax returns, Technical report, National Bureau of Economic Research.

Krenek, A. and Schratzenstaller, M. (2017), Sustainability-oriented future eu funding: A european net wealth tax, Working Paper-Series 10, FairTax.

Leventi, C., Matsaganis, M. and Flevotomou, M. (2013), Distributional implications of tax evasion and the crisis in greece, Technical Report EUROMOD Working Paper EM17/13, University of Essex.

Marini, A., Zini, M. D., Kanavitsa, E., Millan, N., Leventi, C. and Umapathi, N. (2019), A quantitative evaluation of the greek social solidarity income, Technical Report No. 133962, The World Bank.

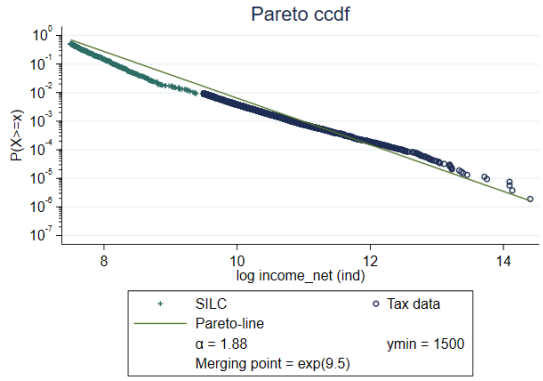
Roine, J., Vlachos, J. and Waldenström, D. (2009), 'The long-run determinants of inequality: What can we learn from top income data?', *Journal of public economics* **93**(7-8), 974–988.

Smith, M., Yagan, D., Zidar, O. and Zwick, E. (2019), 'Capitalists in the twenty-first century', *The Quarterly Journal of Economics* **134**(4), 1675–1745.

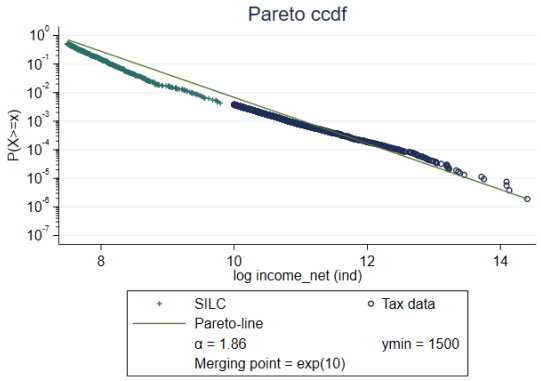
Vermeulen, P. (2017), 'How fat is the top tail of the wealth distribution?', *Review of Income and Wealth* . Forthcoming.
URL: *http://dx.doi.org/10.1111/roiw.12279*

Appendix

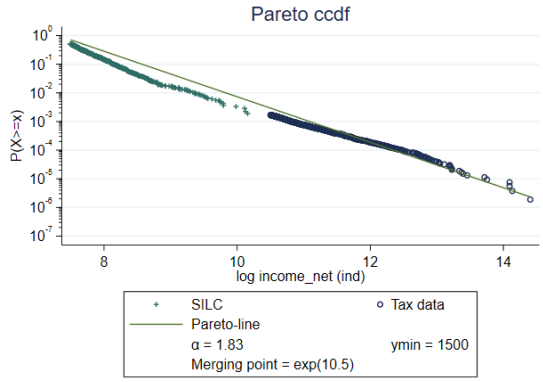
Figure A1: Cumulative complementary distribution functions by variation of the merging point (mp)



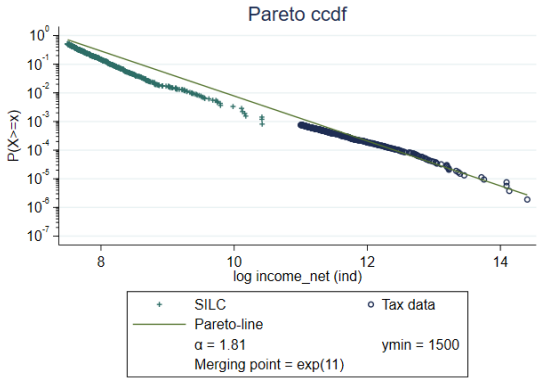
(a) $y_{min} = 1500$, $mp = 13360$



(b) $y_{min} = 1500$, $mp = 22026$



(c) $y_{min} = 1500$, $mp = 36316$



(d) $y_{min} = 1500$, $mp = 59874$

Note: y_{min} determines the lower bound of the Pareto top tail, while the merging point (mp) reports the income cutoff above which the SILC sample is replaced by the tax data in the sample used to estimate Pareto α.