Working Paper Series

Wealth Survey Calibration Using Income Tax Data

Daniel Kolar

ECINEQ 2023 659

ECINEQ 2023 659
2023 November
www.ecineq.org

Wealth Survey Calibration Using Income Tax Data

Daniel Kolar
Charles University in Prague

Abstract

Wealth surveys tend to underestimate wealth concentration at the top due to the missing rich problem. I propose a new way of improving the credibility of wealth surveys by making them consistent with tabulated income tax data. This is possible with the harmonized triannual Household Finance and Consumption Survey (HFCS), which collects data on both income and wealth. I achieve consis- tency by calibrating survey weights using the income part of HFCS. I apply the calibration method of Blanchet, Flores, and Morgan (J Econ Inequal 20(1):119- 150, 2022) in a new context and propose a new, intuitive way to determine the merging point where the calibration starts. I then use the calibrated weights with HFCS wealth values. Tested on Austria, calibration aligns the survey totals closer to the National Accounts, with wealth inequality increasing in the second and third survey waves. I also find a strong downward bias in the Austrian HFCS income distribution. Following the calibration, I test other top tail adjustments: replacing the survey top tail with a Pareto distribution and combining the data with a magazine rich list.

Keyword: inequality, wealth surveys, calibration, Household Finance and Consumption Survey

JEL Cassification: D31, C83

Wealth Survey Calibration Using Income Tax Data

Daniel Kolář*

Institute of Economic Studies, Faculty of Social Sciences,
Charles University in Prague, Czech Republic.

Abstract

Wealth surveys tend to underestimate wealth concentration at the top due to the "missing rich" problem. I propose a new way of improving the credibility of wealth surveys by making them consistent with tabulated income tax data. This is possible with the harmonized triannual Household Finance and Consumption Survey (HFCS), which collects data on both income and wealth. I achieve consistency by calibrating survey *weights* using the income part of HFCS. I apply the calibration method of Blanchet, Flores, and Morgan (J Econ Inequal 20(1):119-150, 2022) in a new context and propose a new, intuitive way to determine the merging point where the calibration starts. I then use the calibrated weights with HFCS wealth values. Tested on Austria, calibration aligns the survey totals closer to the National Accounts, with wealth inequality increasing in the second and third survey waves. I also find a strong downward bias in the Austrian HFCS income distribution. Following the calibration, I test other top tail adjustments: replacing the survey top tail with a Pareto distribution and combining the data with a magazine rich list.

Keywords: inequality, wealth surveys, calibration, Household Finance and Consumption Survey

JEL Classification: D31 , C83

*E-mail: daniel.kolar@fsv.cuni.cz. I express gratitude to Petr Janský, my advisor; to Thomas Blanchet, Rafael Carranza, Ignacio Flores, Arthur Kennickell, Emanuel List, Marc Morgan, Thomas Piketty and Sofie Waltl; to participants at the 2022 Annual Congress of the IIPF in Linz, the 37th IARIW General Conference in Luxembourg, the SEA Meeting 2022 in Bratislava, the Tenth ECINEQ Meeting 2023, the EEA-ESEM Congress 2023, and the Public Economics Seminar in Prague. The HFCS data are provided by the European Central Bank and the EU-SILC data by Eurostat, neither of whom is responsible for the conclusions drawn from the data. Any errors are my own. This work was supported by Charles University, project GA UK No. 321921.

1 Introduction

Wealth inequality is a challenge for countries worldwide, yet its exact level can only be estimated. As opposed to income, household wealth is typically not taxed and therefore not recorded. One way to estimate wealth inequality is by using wealth surveys, but those suffer from the "missing rich" problem (Lustig, 2019). Since wealth distribution is highly skewed to the right, the richest households may not be adequately represented in the survey. They are less likely to participate (Kennickell, 2019) and even if they do, they may underreport their true wealth more often or leave sensitive questions unanswered.

I propose a novel method that increases the credibility of wealth surveys by making their income distribution consistent with income tax data. This is possible thanks to the triannual Household Finance and Consumption Survey (HFCS), which covers 23 European countries in its last wave and collects data on both income and wealth. I achieve consistency by applying the calibration method of Blanchet et al (2022b) to the income part of the survey. As only survey *weights* are adjusted, they can then be used together with the wealth data to estimate wealth inequality. Within the calibration framework, I propose a new method to determine where the merging should start, which is based on a visual comparison of survey and tax income distributions.

A large body of literature aims to mitigate the missing rich problem in wealth surveys (see e.g., Kennickell et al, 2022, for an overview). One stream combines wealth surveys with external data sources: Vermeulen (2014, 2018) replaces the survey's top tail with a Pareto distribution estimated from survey observations combined with the Forbes World's Billionaires List (the Pareto distribution may also be estimated using survey data alone, without the rich list [Eckerstorfer et al, 2016; Hlasny and Verme, 2018]). Subsequent research extends this work by using more detailed national rich lists (Bach et al, 2019), by scaling up different asset classes to match the National Accounts totals (Vermeulen, 2016; Waltl, 2022), by exploring different estimators of the Pareto coefficient (Waltl and Chakraborty, 2022), by using a more flexible Generalized Pareto distribution (Kennickell, 2021; Disslbacher et al, 2020), or by determining a non-arbitrary lower bound of the Pareto distribution (Eckerstorfer et al, 2016; Brzeziński et al, 2020).

However, there is no consensus on the optimal approach. The quality of magazine rich lists with opaque methodology may be contested, as may be the choice of the Pareto estimation technique. Results can be sensitive to both these factors (Kennickell et al, 2022). Perhaps also due to such limitations, the World Inequality Database prefers the income capitalization method as a starting point for their wealth inequality estimates (Alvaredo et al, 2021).[1]

My method improves the representation of high-income households in a wealth survey, insofar as their income is recorded in the tax statistics. It makes use of a high-quality external source, income tax data, and does not depend on arbitrary assumptions. It also results in a data set of identical shape as the original, with only the weight column changed. Any analysis performed on the original HFCS data can thus be extended to the calibrated sample without complication. In addition, calibration has the potential to mitigate the problem that different countries use different (if any) strategies for oversampling of the rich in HFCS, which hinders their comparability. If the oversampling is of high quality and leads to a representative sample at the top of the income and wealth distributions, the impact of calibration will be minimal.

[1] The income capitalization method uses more reliable tax data on capital income. It scales each category up to match the aggregate value of the corresponding asset class in the National Accounts' household balance sheet. Assets that do not generate taxable income flows may then be imputed from surveys (e.g., Garbinti et al, 2021). The income capitalization method requires income tax microdata and assumes a constant rate of return for each asset class across wealth groups.

In contrast, a significant effect can be expected if no oversampling strategy whatsoever has been implemented. Calibration also preserves the main socio-demographic characteristics of the survey.

I test the method on Austrian data and find that calibration aligns the survey totals closer to the official macroeconomic statistics. Top 1 % wealth shares increase from 26 % to 37 % in 2014 and from 23 % to 27 % in 2017. In 2011, even though calibration increases the net worth of the top 1 %, the denominator (i.e., total wealth) increases even more, leading to a slightly negative impact. An oversampling strategy was employed in the 2011 wave, which led to a smaller discrepancy between the survey and tax income distributions. At the same time, calibration highlights the issue of insufficient coverage at the top, where only a few survey observations determine the top wealth shares. This leads to large standard errors of survey estimates both before and after calibration. As part of my merging point algorithm, I find a large downward bias in the Austrian HFCS income distribution, which starts as early as before the 80^{th} percentile. This bias is much larger than in the EU Statistics on Income and Living Conditions (EU-SILC) data, which I also show.

After calibration, I combine my adjustment with those in the literature. I replace the survey top tail with a Pareto distribution and combine the data with a magazine rich list. The inclusion of the rich list in the Pareto estimation has the largest impact. The top 1 % share increases to around 40 % in all three years and the impact of survey weight calibration is negligible. This is because the rich list "dominates" the objective function of the Ordinary Least Squares (OLS) Pareto estimator and its quality is therefore crucial. If the rich list is incorrect, so will be the top shares. Fitting a Pareto tail using only survey data, without the rich list, produces erratic results. They have very large standard errors and are sensitive to the choice of the estimation method as well as the lower bound of the Pareto distribution. Calibration is suitable especially when the rich list is considered unreliable and when it is desirable to preserve all properties of the survey dataset for subsequent analysis.

The drawback of the presented wealth survey calibration method is shared with most existing adjustments: its limited theoretical foundation. It cannot be guaranteed that "fixing" the income distribution will also correct the entire bias in the wealth distribution. In the Austrian context, the "fixing" relates mainly to households with high labor and self-employment income, since this income is well captured in the income tax statistics. The rich with low taxable income or with capital income taxed at source remain missing even after calibration, just like when applying the income capitalization method. Standard errors remain high, which can only be mitigated by increasing the number of surveyed rich households. The proposed method is thus another exercise in "the art of the possible" (Sen, 1997). Ideally, a similar exercise would be incorporated directly into the survey design: high-income households would be oversampled based on income tax records and their survey weights computed to ensure their representativeness of the high-income strata. This would not only ensure consistency of survey and tax income distributions, but also increase the absolute number of surveyed rich households. An approach along these lines has recently been applied to the fourth wave of the Italian HFCS (Barcaroli et al, 2021).

In addition to proposing a new method for improving wealth surveys, I make one contribution to the income calibration framework of Blanchet et al (2022b, hereafter BFM). I construct a new method to determine the *merging point*, which is the optimal percentile from which calibration should start. While I share the aim of BFM to preserve the continuity of the new income density, my approach relies on a (perhaps more intuitive and informative) graphical comparison of survey and tax income distributions. In the context of the HFCS, such comparison is an interesting external check of the survey's quality. In contrast to BFM, my method may also lead to more than one candidate merging point, giving researchers more flexibility. In a Monte Carlo simulation, my method performs comparably or better, depending on the specification.

Moreover, it does not rely on the assumption of a monotonically decreasing survey to tax density ratio and is not sensitive to the choice of the lower bound from which the tax data are considered reliable. However, my method is more computationally demanding and less suitable when the merging point should be above the 99^{th} percentile. The latter disadvantage can nevertheless be mitigated by using more granular income brackets.

The remainder of the paper is organized as follows. Section 2 introduces the data that I use to apply the method: wealth surveys, income tax data and rich lists. I also discuss the main income and wealth concepts. The methodological Section 3 describes the survey weight calibration and the adjustments to the top of the wealth distribution. In Section 4, I report the results, including those of a Monte Carlo simulation that evaluates the new merging point algorithm. Section 5 concludes.

2 Data and main concepts

2.1 Wealth surveys

The wealth surveys to be corrected come from the triannual Eurosystem Household Finance and Consumption Survey (HFCS), which is harmonized and coordinated by the European Central Bank (ECB). While the first wave in 2010 covered 15 euro area countries, the fourth wave will include all 19 euro area countries as well as Croatia, Hungary, Poland, and for the first time also the Czech Republic. Questions regarding assets, liabilities, consumption and savings are collected at the household level, which is the main unit of analysis. The HFCS also collects data on seven main categories of income, three at the individual level (wages, pensions and self-employment income) and four at the household level (real estate income, income from financial investment [i.e., interest and dividends from publicly traded companies], income from private business, and other income [which includes capital gains]). Income is recorded as gross, including taxes and contributions to social insurance paid by employees. The income tax data have to be carefully matched with these concepts to ensure accurate calibration. For wealth inequality measurement, I use the net wealth concept of the HFCS: For each household, it is the sum of their real assets (real estate, vehicles, valuables and self-employment businesses) and financial assets (e.g., deposits, mutual funds, bonds and shares; excluding public and occupational pension plans) minus its liabilities.

The main issue with wealth (as well as income) surveys is that they may not capture well the top of the distribution. Lustig (2019) calls this the "missing rich problem", while BFM refer to the "non-sampling error". First and foremost, it has been observed that when approached by the interviewer, wealthier households are more likely to refuse to participate (Kennickell, 2019; Kennickell and Woodburn, 1999). One obvious reason is the opportunity costs: While the median interview length in HFCS countries is rarely below 40 minutes (in Austria, for example, it is 55 minutes), the reward is mostly symbolic (for example, the Czech Statistical Office rewards participating households with commemorative coins with a total face value of approximately 5 euro). Second, the rich may not answer questions regarding asset classes or income types they consider sensitive. In this case their value is imputed five times, leading to five dataset replicates for each HFCS wave. A third source of bias arises if the rich are more prone to underreport wealth, which may happen especially in the case of offshore wealth that is strongly concentrated at the top (Alstadsæter et al, 2019). In addition, surveys may be plagued with "sampling error" (Blanchet et al, 2022a), meaning that the number of rich households in the survey is too small to produce accurate results (as demonstrated, for example, by Eckerstorfer et al, 2016).

To mitigate the missing rich problem, most HFCS countries oversample the wealthy. The most precise oversampling method would utilize individual data from tax

registers to determine the "rich" strata in the population. Alternatively, oversampling can be based on income in a given geographic area, street address, dwelling characteristics or even electricity consumption. Austria, the country analyzed in this paper, oversampled Vienna in the first wave, but did not record any oversampling strategy in waves 2 and 3. Vermeulen (2018) finds a correlation between the oversampling strategy and the number of wealthy respondents in the net sample: "In practice, successful oversampling leads to many wealthy households in the sample, all with relatively low survey weights" (Vermeulen, 2018, p. 362). Consequently, I note that oversampling reduces the bias of wealth inequality estimates not so much because the richest are sampled more (this mainly reduces the variance, although small sample bias may also be an issue [Taleb and Douady, 2015]), but because the rich strata are more accurately defined and the adjustment of survey weights for non-response can be more specific.

Figure 1 reports two survey statistics that serve as a proxy for survey quality: the oversampling rate and the (net) sample size relative to the total population. The oversampling rate is based on the number of survey households in the top 10 % of the survey wealth distribution. In a simple random sample all weights would be equal, the number of such households would be 10 % and the oversampling rate 0 %. In HFCS countries other than Austria the median oversampling rate is 46 % in the third wave. This means that the top 10 % of the wealth distribution consists of 14.6 % survey households, indicating the existence of a meaningful oversampling strategy. In contrast, Austria only records a small and positive oversampling rate in the first survey wave (when Vienna was oversampled), after which the statistic even turns negative. The number of survey households that constitutes the top 10 % is only 8.5 % in the third wave.

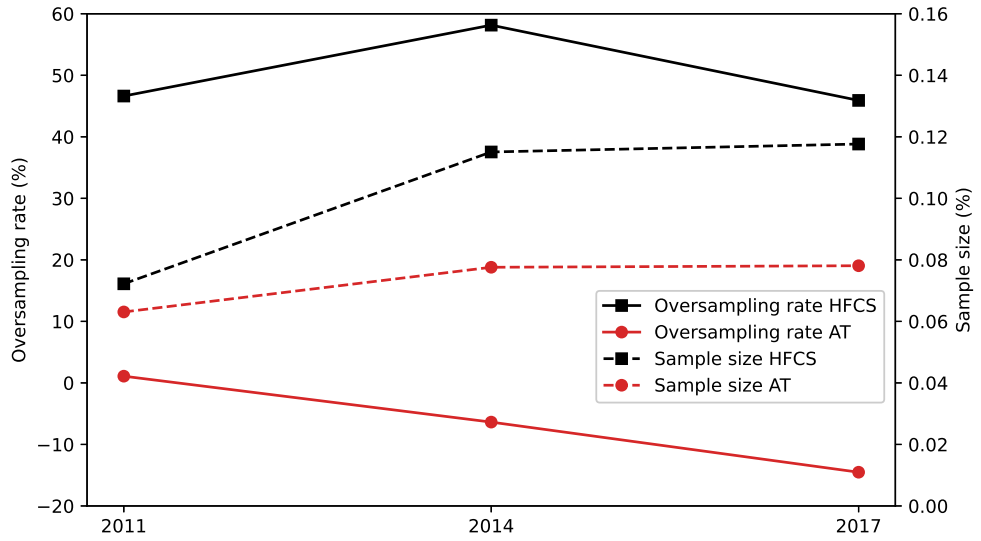

Fig. 1 Oversampling rate and net sample size relative to the population in HFCS. Values for Austria are compared with the median of the remaining HFCS countries that participated in all three survey waves. Values are averaged over implicates.

While the oversampling rate indicates a decline in the quality of the Austrian HFCS, the relative sample size gives the opposite impression. It increases by 23 % (0.015 p.p.) between the first and second waves, and grows slightly even in the third wave. One can thus expect lower variance of the inequality measures in the second and third waves. However, the sample size still remains smaller than the median of the remaining HFCS countries.

Calibration must also take into account the potentially different time periods for which income and wealth data was collected. In the Austrian case, as in most other countries, the reference period for wealth information was the time of interview. Survey fieldwork could span a period of two calendar years, in which case I assign wealth data to the year in which fieldwork predominantly took place. In contrast, data on income was collected for a preceding calendar year. As a result, the first HFCS wave in Austria contains income data for 2009 and wealth data for 2011. For the second wave the income data is for 2013 and wealth data for 2014, and for the third wave the respective years are 2016 and 2017. To assess the quality of the HFCS income distribution, I compare it with the EU Statistics on Income and Living Conditions (EU-SILC) data for the respective reference years.

2.2 Tax data

Data from income tax returns represents the "true" income distribution in the weight calibration process. It is not perfect and may be incomplete due to tax evasion, tax exemptions or withholding of certain taxes at source. For the income types that are included in the tax returns, this data nonetheless represents the best available information about their distribution. Tax data are typically published in a tabulated form: The population that files tax returns is divided into income brackets, and each bracket lists information on the number of people and their total or average income. A complete income distribution can be obtained using the Generalized Pareto interpolation method of Blanchet et al (2022c).

For Austria, tabulated tax data are published annually in the Integrierte Statistik der Lohn-und Einkommensteuer (Statistics Austria, 2016). However, taxes from capital investment income, including capital gains, are withheld at source and typically not recorded in tax returns.[2] Based on the documentation, I match tax data with wages, pensions, self-employment income, real estate income and public transfers in the survey. As the tax unit in Austria is the individual, I split real estate income and public transfers in the survey (recorded at the household level) equally between adult household members aged 20 and over. In addition, tax data are reported net of social contributions (Jestl and List, 2020), which are part of the gross income concept in the survey. Following Blanchet et al (2022a), I estimate and deduct social contributions from survey data where appropriate, based on rates recorded by the OECD.[3]

2.3 Rich lists

Rich lists provide information about the very top of a country's wealth distribution. By suitably combining them with wealth survey data, one can "anchor" (Vermeulen, 2018) the top of the distribution and estimate the wealth of those "too poor to be in the rankings, but too rich to be in the survey" (Blanchet, 2016). Austrian rich lists are compiled by *Trend* magazine. They have 100 entries, but an exact wealth estimate is assigned only to the first 60 of them. I follow Eckerstorfer et al (2016), Appendix III and adjust the rich list so that each entry represents one household rather than a family or clan. For example, the top entry in all studied years lists the Piëch and Porsche families, which Eckerstorfer et al (2016) divide into seven households.

[2] Exceptions exist for foreign capital income or for realized capital gains offset with realized capital losses in the same period; however, this is likely to be a small part of total capital gains income.

[3] For wages, I use rates recorded in the annual OECD Taxing Wages publication, for pensions I use the OECD Pensions at a Glance publication and for self-employment income the OECD Tax Statistics. In addition, I set the limit for maximum social security contributions from self-employment (17,793 euro in 2016) as the limit for the total contributions of an individual.

3 Methodology

The core of the presented method lies in adjusting survey weights to impose consistency of the survey's income distribution with the tax data. I follow the reweighting approach of BFM, but propose a new method to determine the merging point – the percentile from which consistency is imposed. The new weights are then used with sampled households' wealth to estimate top shares.

3.1 Income distribution calibration

3.1.1 The calibration formula

Survey weight calibration, as applied by BFM, consists of several steps. First, the tax data are divided into many brackets based on fractiles: From 0 to 0.99, from 0.99 to 0.999, from 0.999 to 0.9999 and from 0.9999 to 0.99999. Survey observations are then matched to these brackets. The tax data usually cover only part of the population, in which case the lowest bracket starts at a higher percentile than 0 and the left-bounded interval where the tax data are reliable is referred to as the *trustable span*. In simple terms, the goal of calibration is that weights of survey observations that are matched to a bracket sum up to the tax population size of that bracket. BFM describe it as a "histogram approximation"; it is essentially a scaling up (or down) of histogram bins of survey data to match their tax counterpart. As it is generally not desirable to approximate the entire tax distribution, only brackets above the predefined *merging point* are calibrated. I discuss the choice of the merging point in Section 3.1.2.

After matching survey observations to fractile-based brackets, I merge brackets where the number of corresponding survey observations is below x. In this case, calibration would either not be possible at all (if the number of matched observations is 0) or the weight adjustment would be too large. For similar reasons, brackets are merged if the ratio of survey to tax frequencies is below $1/y$ or above y. BFM choose $x = y = 5$, which I follow. These parameters reflect a trade-off between calibration accuracy and survey distortion. The new weights that satisfy the defined conditions are obtained by solving the linear calibration problem:

$$\min_{w_1,\dots w_n} \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i} \quad \text{s.t.} \quad \sum_{i=1}^{n} w_i \boldsymbol{x}_i = \boldsymbol{t}, \tag{1}$$

where d_i is the original weight and w_i the new weight of individual i; n is the number of surveyed individuals. $\boldsymbol{x}_i$ is a k-dimensional vector characterizing individual i and $\boldsymbol{t}$ is a k-dimensional vector of corresponding population totals from an external source. For example, assume k is the number of fractile brackets obtained in the previous step and x_{1i} is a dummy variable denoting whether individual i belongs to the highest bracket. The condition $\sum_{i=1}^{n} w_i x_{1i} = t_1$ means that the calibrated weights of individuals in the highest bracket will sum up to the population total from the tax data, t_1. The remaining $k-1$ conditions are constructed accordingly.

Thanks to the flexibility of linear calibration, other conditions can be imposed on the calibrated weights by expanding the vectors $\boldsymbol{x}_i \forall i$ and $\boldsymbol{t}$. Like BFM, I preserve the main socio-demographic characteristics of the survey population: age, gender and household size distribution, as well as total population size. An infinity of solutions will generally satisfy these conditions, and Equation 1 is solved by one which minimizes the χ^2 distance between the original and new weights. As long as all the constraints are neither incompatible nor perfectly collinear, Equation 1 has a closed-form solution (BFM, equation 5). In addition, no new weight should be lower than 1, which I enforce with a simple algorithm.[4]

[4] If the linear calibration leads to some weights lower than one, I repeat the process but add conditions that these weights must be equal to one. If the problem remains, meaning that some *other* observations now have a weight lower than 1, I correct weights from the first iteration directly by setting them to one and

One specific problem arises in the presented application: While income is recorded at the individual level in surveys and in Austrian tax data, wealth is only recorded at the household level. Original HFCS weights are the same for the household as well as for each household member, but that will no longer be the case once individual weights are calibrated to the tax data. One solution would be to use individual members' average weight as the household weight, but if I then move back from household to individual weights assuming the same weight for each household member, tax and survey income distributions will no longer be consistent. Instead, I add $m-1$ additional linear constraints for each m-member household so that the difference between each pair of household members' weights is equal to zero.

3.1.2 Optimal merging point

The merging point is the fractile from which merging of the survey and tax data begins. It is, in principle, possible to start the calibration as soon as the tax data are available. However, as BFM argue, this would unnecessarily distort the survey, which should be corrected only once the bias at the top starts. BFM propose a new approach to determine the merging point with the aim of preserving the continuity of the density function. Their method is based on a theoretical framework and applied by comparing the ratio of survey and tax densities with the ratio of survey and tax cumulative distribution functions. They assume that the ratio of survey to tax density is monotonically decreasing, presumably to avoid multiple solutions to their algorithm.

I propose a new way of choosing the merging point while also aiming to preserve the continuity of the calibrated density function. The starting point is a visual comparison of tax and survey frequencies at different percentile-based brackets, illustrated in Figure 2 using Austrian third-wave HFCS data. As the income intervals are computed using tax data, the tax frequency is constant and equal to 1 % of the population. In contrast, survey observations are matched to these brackets and the frequency may thus be higher than 1 % (if a bracket is overrepresented in the survey) or lower (if it is underrepresented). In what follows, I work with an adaptive kernel estimator of the survey density (Cowell and Flachaire, 2015) with a Gaussian kernel and extended to incorporate survey weights (Buskirk, 1998). I integrate it over each bracket's income interval and multiply the integral by the population total, obtaining a smoother frequency estimate for each bracket.[5]

I search for the *best* rather than *optimal* merging point: I test each percentile and observe whether the new, calibrated survey distribution can be considered continuous. The test statistic is the absolute distance between the frequency where merging begins (equal, by construction, to the tax frequency) and the frequency at the neighboring bracket which was not calibrated. The latter frequency, however, is not simply equal to the original survey frequency, but is adjusted so that the population total remains the same. This adjustment is assumed to be uniform for all brackets below the merging point.

A straightforward choice for the merging point would be the percentile for which the test statistic, that is, the distance between the two neighboring frequencies, is lowest. This is indeed correct if the ratio of survey to tax density is monotonically decreasing (as assumed by BFM) and the frequencies thus cross only once. However, if the relationship between the densities is more complex, there may be more merging points leading to a visually continuous density.

I propose the following algorithm for choosing *candidate* merging points: Consider all percentiles for which the test statistic is lower than 3% of the tax frequency – in

adjusting (i.e., slightly decreasing) all other weights to keep the population total constant. BFM instead use an iterative method described in Singh and Mohl (1996).

[5] To produce informative results, this integration requires the tax distribution to be reasonably smooth as well. This will be the case if the tabulated income tax data are interpolated using the Generalized Pareto interpolation method of Blanchet et al (2022c).

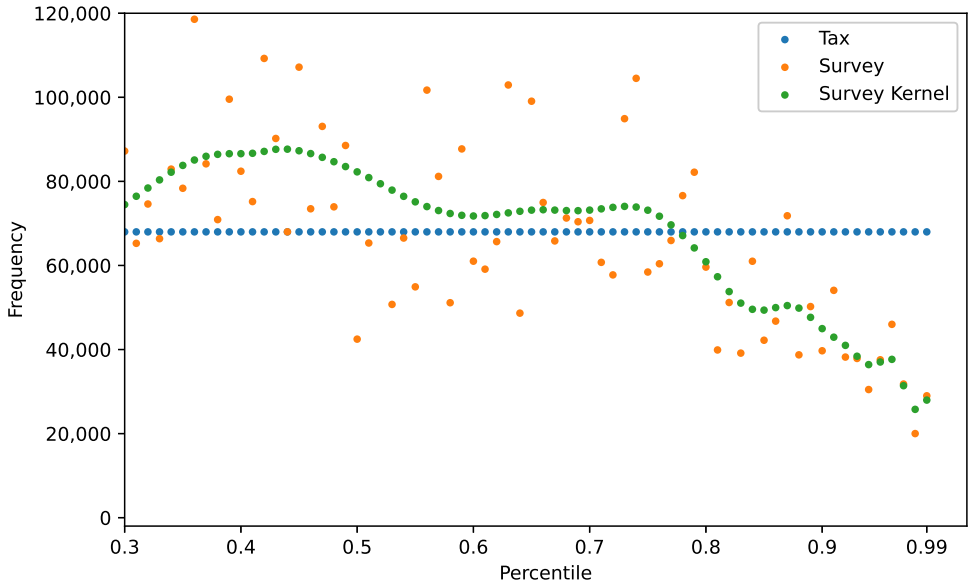

Fig. 2 Tax and survey frequencies.

this case, the new density can be considered visually continuous. Consider also all percentiles with test statistic lower than 130 % of the minimum value – this condition guarantees at least one candidate merging point if no percentile satisfies the first condition. Finally, disregard percentiles where a neighboring percentile has a lower test statistic – in this case, that neighboring percentile is clearly preferred. If the first condition does not identify any merging point, it is a warning sign that there might be an issue with the data.[6] The algorithm should always be accompanied by a visual inspection of tax and survey frequencies, which is informative on its own.

I illustrate the merging point choice in Figure 3, using the same data as in Figure 2. Setting the trustable span to start at the 30^{th} percentile, the algorithm identified two candidate merging points: percentiles 0.56 and 0.74; both with the test statistic below 3 % of the tax frequency. In general, all candidate merging points should be considered. For example, a graphical comparison may reveal uneven oversampling at the top where the top 2-5 % is overrepresented but the top 2 % underrepresented. In that case, one may want to correct also the overrepresentation of the top 2-5 %. But my general recommendation, which I also apply, is to choose the highest candidate merging point; in the illustrative case this will be percentile 0.74. This corresponds to the goal of correcting the survey distribution only once the downward bias at the top starts.

Comparing my optimal merging point method to the algorithm of BFM, I argue that my method should be considered due to its intuitiveness and simplicity. It avoids the assumption of a monotonically decreasing survey to tax density ratio and, furthermore, allows for an informative visual comparison of frequencies. It is also more flexible thanks to the possibility of more candidate merging points. On the other hand, the method of BFM allows for the merging point to be above the 99^{th} percentile. If the frequency comparison suggests this could be the case, my method can be easily extended to more granular frequency brackets (e.g., 1/2 or 1/4 of a percent). I describe and illustrate this extension in Appendix A, where I also note that the BFM merging point can be sensitive to the choice of the trustable span, which is arguably undesirable.

[6] For example, concepts in tax and survey data may not be correctly matched. Or tax data may only cover a fraction of the population (at the top) that is too small. In the latter case, BFM extrapolate the ratio of survey to tax density for percentiles not covered.

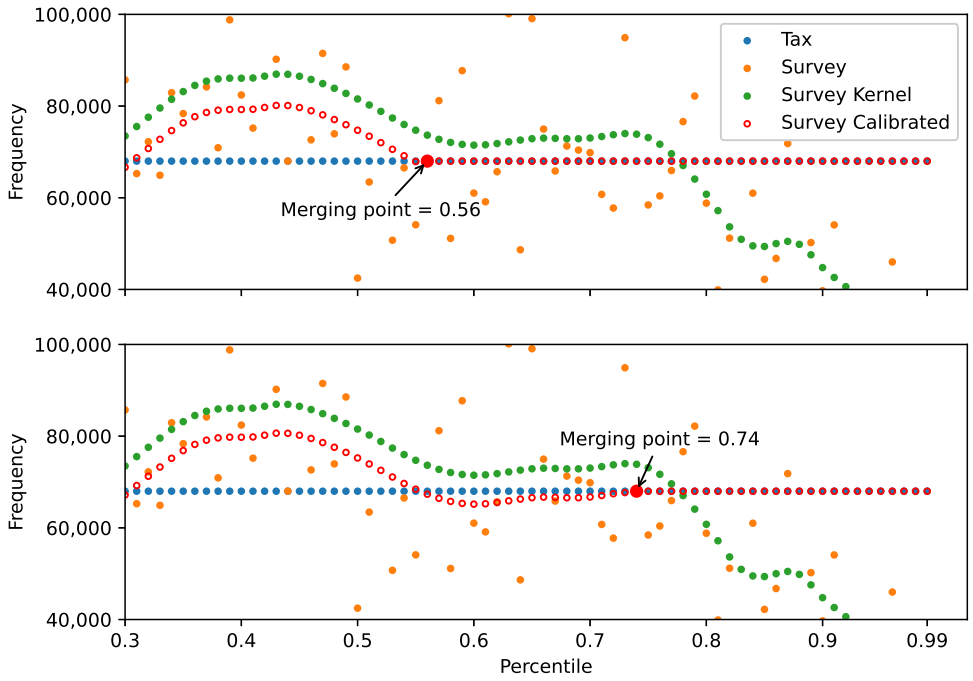

Fig. 3 Choosing the optimal merging point. 0.56 and 0.74 are two candidate merging points, each leading to a visually continuous density after calibration. I choose 0.74 as the optimal one as it is the highest. Same dataset as in Figure 2.

As the aim of the two methods is the same (i.e., preserving the continuity of the density function), they can lead to similar results. That is also the case with the Austrian data used to illustrate my method in Figures 2 and 3: My method leads to a merging point at the 74$^{\text{th}}$ percentile, just like the method of BFM. An additional comparison based on simulated datasets is provided in Section 4.1.

3.1.3 HFCS-specific adjustments

As the HFCS data are provided in five dataset replicates due to the imputation of missing values, my method must be adjusted to take this into account. Rather than estimating a potentially different optimal merging point for each replicate, I estimate one merging point based on the average test statistic of each percentile.[7] A further issue concerns surveyed individuals who are young but exhibit relatively high incomes. When top wealth or income shares are reported in the literature, the unit of analysis is usually *adult* individuals, frequently defined as aged 20 or over (Alvaredo et al, 2021). However, if a survey records younger high-income individuals, excluding them will not lead to an accurate comparison of tax and survey distributions (as they likely also file tax returns). On the other hand, constructing the fractile-based brackets from the entire population will unnecesarily widen these brackets. I therefore choose to treat young individuals with income above the threshold of 10,000 euro as adults. This does not lead to any conceptual issues as I study the inequality of wealth, which is reported at the household level. However, if one was interested in income inequality estimates of adult individuals, "young rich" people should be removed from the adult population after calibration, as they would be if one had perfect tax microdata.

[7] Another possibility is to apply the maxi-min criterion (Wald, 1945; Eckerstorfer et al, 2016), which in this context means choosing the percentile for which the maximum (i.e., worst) test statistic across five implicates is the lowest.

3.2 Existing wealth survey adjustments: The Pareto tail

3.2.1 Fitting a Pareto tail

One approach to mitigate the "missing rich" error is to replace the top of the survey wealth distribution with a Pareto distribution. Monte Carlo simulations have shown that this approach can decrease or even eliminate bias caused by the more likely non-response of the wealthy (Vermeulen, 2018) as well as decrease error resulting from a small sample size (Eckerstorfer et al, 2016). In brief, the method consists of setting a wealth threshold, estimating a Pareto coefficient from survey observations above that threshold (optionally including individuals from a rich list) and replacing the population these observations represent with a Pareto distribution. For a more exhaustive overview of wealth survey adjustment methods, see Kennickell et al (2022).

Pareto coefficient α

The Pareto distribution is characterized by the following complementary cumulative distribution function (ccdf) and density:

$$P(X > w) = \left(\frac{w_{min}}{w}\right)^{\alpha} \tag{2}$$

$$f(w) = \left(\frac{\alpha w_{min}^{\alpha}}{w^{\alpha+1}}\right), \tag{3}$$

where w_{min} is the threshold at which the Pareto distribution starts and α is the Pareto coefficient, $\alpha > 0$. Vermeulen (2018) shows how to extend estimators of the α parameter from simple random samples to surveys where observations are weighted and not i.i.d. The maximum likelihood estimator of α that takes into account complex survey weights can be defined as

$$\hat{\alpha}_{MLE} = \frac{n-1}{n}\left[\sum_{i=1}^{n}\frac{N_i}{N}\ln\left(\frac{w_i}{w_{min}}\right)\right]^{-1}, \tag{4}$$

where n is the number of survey respondents with wealth above threshold w_{min}, N_i is the survey weight of respondent i, w_i their wealth and $N = \sum_{i=1}^{n} N_i$ is the top tail population to be replaced. Vermeulen (2018) calls this the pseudo-maximum likelihood estimator. Note that if all weights are equal to 1, the estimator becomes a maximum likelihood estimator that is minimum-variance unbiased in a simple random sample and when w_{min} is known (Rytgaard, 1990). The extension means that observations that represent more households have a larger impact on the estimate.

The second approach to estimating Pareto coefficient α exploits the property that a Pareto distributed sample approximately follows a straight line on a log-rank log-wealth graph. The relationship can be derived by replacing the complementary cumulative distribution (Equation 2) by its empirical counterpart and manipulating it:

$$\frac{N(w_i)}{N} \approx \left(\frac{w_{min}}{w_i}\right)^{\alpha} \tag{5}$$

$$\ln\left(\frac{N(w_i)}{N}\right) \approx -\alpha \ln\left(\frac{w_i}{w_{min}}\right), \tag{6}$$

where $N(w_i)$ is the population (i.e., the sum of survey respondents' weights) with wealth at or above w_i.

An estimate of α can be obtained from Equation 6 with a linear regression, but it will be biased. Intuitively, the source of this bias is that for a continuous distribution $P(X > w) = P(X \geq w)$. The empirical ccdf can thus be represented by both $\frac{N(w_i)}{N}$ and $\frac{N^*(w_i)}{N}$, where $N^*(w_i)$ is the population with wealth strictly above w_i. Gabaix and Ibragimov (2011) show that in a simple random sample (i.e., when all weights

are one and all observations i.i.d), bias can be removed by subtracting 1/2 from an observation's rank, which in the present notation corresponds to taking the average of $\frac{N(w_i)}{N}$ and $\frac{N^*(w_i)}{N}$. Wildauer and Kapeller (2019) propose to keep this correction also in the complex survey setting, which I do as well. A similar correction is applied by Vermeulen (2018), but the difference is that the correction of Wildauer and Kapeller (2019) allows estimating α without the intercept. This improves the fit of the Pareto line to the data because the intercept does not enter the estimated Pareto distribution. I denote this estimator $\hat{\alpha}_{OLS}$.

Vermeulen (2018) shows in a Monte Carlo simulation that with survey data alone, the maximum likelihood estimator generally performs better than the OLS estimator: Its estimates of α tend to be closer to the truth and with a slightly lower variance, especially in the presence of oversampling. On the other hand, the weighted maximum likelihood estimator is insensitive to adding a rich list to the survey data, since each rich list entry has a weight of only 1. Pareto coefficients based on survey data combined with a rich list are therefore estimated only using $\hat{\alpha}_{OLS}$.

Wealth threshold w_{min}

There is a trade-off in setting w_{min}, the threshold where the Pareto tail starts. By setting it too low, there is a risk of dealing with observations that are not high enough to be well approximated by a Pareto distribution. By setting w_{min} too high, the sample size is decreased, leading to higher variance of $\hat{\alpha}$. One way to determine the threshold is a visual inspection of the data, by observing where the log-rank log-wealth relationship appears to be linear (e.g., Cowell, 2011a) or where van der Wijk's law appears to hold (e.g., Bach et al, 2019).[8] Vermeulen (2018) sets three thresholds at 0.5, 1 and 2 million euro and reports results for all three values.

A non-arbitrary method to determine w_{min} was proposed by Clauset et al (2009) and applied to HFCS data by Eckerstorfer et al (2016) and Brzeziński et al (2020). It consists of estimating a Pareto coefficient for many thresholds and selecting the one where the data provide the best fit to the estimated Pareto distribution. Clauset et al (2009) choose the Kolgomorov-Smirnov (KS) goodness-of-fit test: The KS test statistic is the maximum absolute distance between the empirical and estimated cumulative distribution functions (or, equivalently, the complementary cumulative distribution functions). The threshold for which the KS statistic is lowest is then chosen as w_{min}. Eckerstorfer et al (2016) use a conceptually similar Cramér–von Mises test.

I considered applying the KS statistic to determine w_{min} for the benchmark results but found that it fails to identify obvious deviations from a Pareto tail at the very top of the distribution. I show this in Figure 4 using the first implicate of Austrian 2011 data. The application of the KS method to this dataset leads to a very low optimal threshold at 275,000 euro, covering the top 25 % of Austrian households. However, a visual inspection of the log-rank log-wealth relationship reveals a visible break from the Pareto straight line at around 1.5 million euro (corresponding to the top 4 % of households). The KS test misses the break due to the graph's logarithmic scale, which I illustrate in Figure 4 on the first and fifth largest survey observations. The difference between the largest observation's empirical complementary cumulative distribution function and the one fitted at $w_{min} = 275,000$ is only 0.0027. I compare this with the value for the fifth largest observation, which is visibly closer to the Pareto line but has a higher difference between ccdfs of 0.0045. The maximum difference between the two ccdfs at this threshold, which is the KS statistic, occurs much lower in the distribution. Therefore, observations at the top would not influence the KS statistic for the 275,000 threshold even if they were further away from the Pareto line.

[8] Van der Wijk's law is a property of the Pareto distribution that average wealth above any wealth threshold is a constant multiple of that threshold (Cowell, 2011b).

Due to deviations from the straight line even above 1,500,000 euro, the KS statistic at this threshold is still larger than at 275,000 euro. This is because the ccdfs for the KS test are computed only from the top tail population, i.e., the ccdf at the threshold is always 1. For this reason, Figure 4 is only illustrative; the included values relate to the difference of ccdfs of the entire population.

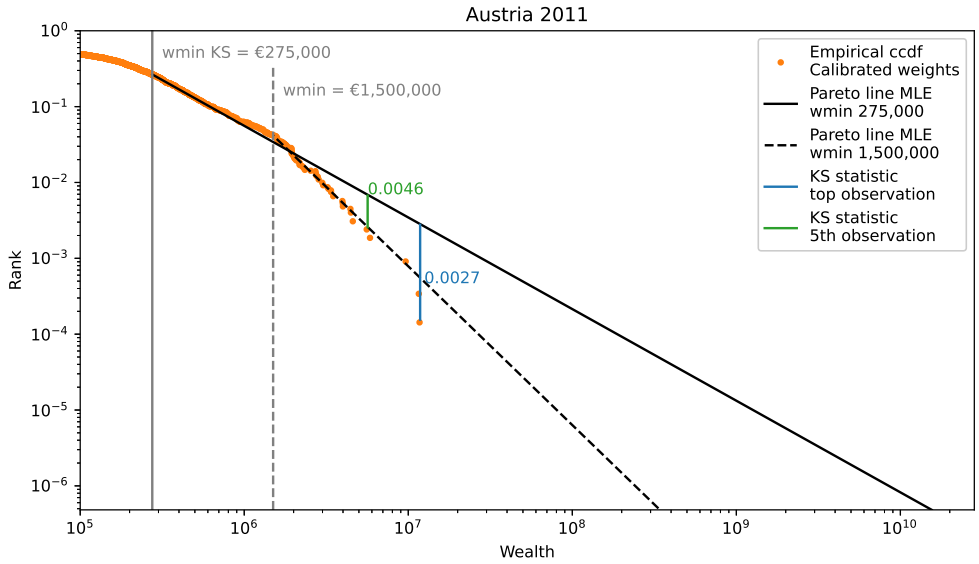

Fig. 4 Estimating the Pareto distribution at different thresholds. 275,000 euro is the optimal threshold according to the Kolgomorov-Smirnov test. The blue and green lines illustrate how the KS statistic is evaluated for two observations at the top. Because of the logarithmic scale, visual deviations from the Pareto line at the top are not accounted for sufficiently in the KS test. Data: HFCS, Austria 2011, first implicate, calibrated weights.

I experimented with adjusting the setup of the KS test but it did not produce reliable results. First, I tried comparing the logarithms of ccdfs in the KS test, which corresponds more closely the graphical comparison in Figure 4. Second, I compared ccdfs at the top computed from the entire population rather than just from the top tail. Each of these adjustments (as well as both in combination) led to implausibly high estimates of optimal thresholds, always near the top of the range of tested values, even if only a small number of survey observations were left.

As neither the approach used by Clauset et al (2009) nor any subsequent adjustments produced estimates of the Pareto threshold that would be consistent with the log-rank log-wealth graph, I resort to setting w_{min} based on a visual examination of the distribution. As a result, the benchmark estimates for all three years are based on a threshold of 1.5 million euro and I also report the sensitivity of results to setting the threshold at 1 and 2 million euro. Clauset et al (2009) recommend that the number of observations from which the Pareto distribution is estimated be at least 50, otherwise the sampling error may be too high. This is satisfied in Austrian data for 2011 and 2017, but not for 2014, with only 39 to 45 observations above the 1.5 million euro threshold, depending on the implicate. However, given the visual fit to the data, I proceed with this threshold nonetheless and report results under alternative thresholds in Appendix C.

Finally, I note that judging the Pareto tail based on its fit to the data may be problematic in principle. If one believes that the data are biased due to differential non-response, it is difficult to justify that the best Pareto tail is the one that fits these biased data. The same argument applies to methods evaluating whether the Pareto distribution is the right distribution to fit. As I explain in Section 3.2.2, the

rationale for combining survey data with rich lists is that the latter "anchors" the Pareto distribution which would otherwise be biased due to differential non-response in the former. In the context of Figure 4, there is a possibility that the break from the Pareto line is due to biased survey data and that the Pareto distribution estimated at the 275,000 euro threshold would in fact fix the bias rather than create it. However, as I estimate the Pareto tail using data that have been (at least partially) corrected by calibration, I consider my approach justified.

3.2.2 Adding rich lists

Estimating the Pareto distribution from the combination of survey data and lists of wealthiest individuals is conceptually simple: The rich list observations are appended to the survey with a weight of 1.[9] $\hat{\alpha}_{MLE}$ is not a suitable estimator of the Pareto coefficient because of how it weighs the observations: The rich list entries' weight of 1 means that their influence on the estimate is minimal. In contrast, in $\hat{\alpha}_{OLS}$ survey weights are only accounted for indirectly, in the rank of observations, and each observation enters the OLS minimization problem with equal relevance. Moreover, since rich list wealth values are extremely large, they will "dominate" the objective function which OLS minimizes (even though they are evaluated as logarithms). I illustrate this point in Figure 5 on the same Austria 2011 dataset as in Figure 4: The Pareto line estimated at the 1.5 million euro threshold is now aligned with the Pareto line estimated at $w_{min} = 275,000$, even though survey observations above 1.5 million euro are not.

Consequently, the quality of rich lists is fundamental for the reliability of wealth inequality estimates. If the rich list is incorrect, so will be the wealth inequality estimates even if there was no bias in the survey data.

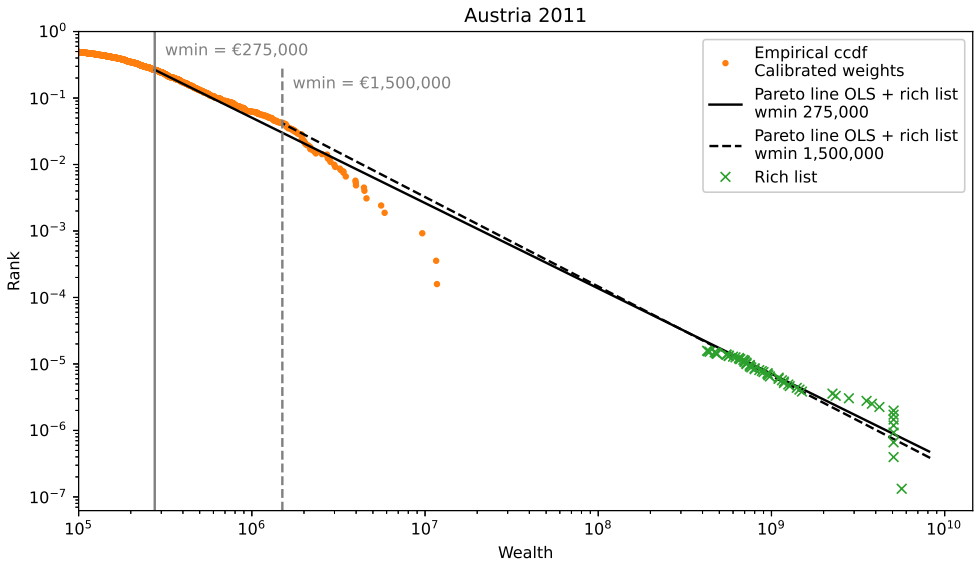

Fig. 5 Estimating the Pareto distribution from survey data combined with a rich list. Once the rich list is included in the estimation, it becomes the main determinant of the slope of the Pareto line. Data: HFCS, Austria 2011, first implicate, calibrated weights.

[9] To preserve the original population size, I also decrease the weight of each survey observation by a small constant. This has no impact on the results.

3.2.3 The Pareto population

Once the Pareto distribution at the top is fully characterized (with or without utilizing the rich list in the estimation), the final question is how to draw the population that it represents. One can obtain top tail wealth directly, using the expected value times the population size at the top. When only part of the top tail belongs to the population of interest (e.g., the top 1 %), conditional expected values may be used. I believe that this approach is used by Vermeulen (2018) because I obtain identical results for Austrian first wave data which are also analyzed in his work.

In this paper, I apply a different approach: I construct a new, synthetic survey population consistent with the Pareto distribution. This approach is used in some form by Bach et al (2019) and Brzeziński et al (2020). In addition to being intuitive, it is also less sensitive to extreme values because as the Pareto coefficient tends to one, the expected value tends to infinity. In general, however, the results achieved using either of the two methods will be very similar, since resampling is essentially numerical integration (Dalitz, 2016). I construct a synthetic population which has an empirical ccdf identical to the Pareto one. Synthetic population's size follows from the survey: It is the sum of weights of survey observations with wealth above the Pareto threshold. When a Kolgomorov-Smirnov goodness-of-fit test is applied to this population, the resulting statistic is 0, the lowest possible.

As a final step of drawing the Pareto population, I replace observations at the very top with corresponding values from the rich list when it is used in the Pareto estimation. This can further prevent implausibly large values at the top but contains a small complication: The largest non-replaced synthetic household may have higher wealth than the poorest household on a rich list. The difference is generally not very large (if this problem is present at all) and I disregard it, implicitly assuming that the journalists may have omitted some households when compiling the rich list.

3.2.4 Variance estimation

To estimate the sampling error component of variance, replicate bootstrap weights are provided in the HFCS dataset. The bootstrap procedure involves sampling with replacement from different population strata and adjusting replicate weights in the same manner as in the original survey (European Central Bank, 2020). The problem in the present context is that these weights are not suitable for variance estimation if I use the new, calibrated weights. There is not enough information to replicate the bootstrap procedure; mainly it is unknown how to divide survey observations into population strata. I therefore replicate it only partially, using the same Rao-Wu rescaled bootstrap (European Central Bank, 2020, Section 7.2) but working with the entire population as the only stratum and not performing any additional adjustments. I create 500 new bootstrap weights for each set of weights. For the non-calibrated weights, I can compare the standard errors of top 1 % share estimates based on the provided weights with those based on my replication. In all studied years, the difference is less than 2 % (0.2 percentage points) when the top share is computed using survey data alone and less than 9 % (1.1 percentage point) when a Pareto tail is fitted using $\hat{\alpha}_{MLE}$.

In addition to sampling error, total variance must take into account variance due to missing values, which are imputed five times. For this, I apply the formula by European Central Bank (2020, Section 7.3).

4 Results

4.1 Monte Carlo simulation: The optimal merging point

In Section 3.1.2, I developed a new method to determine the merging point, i.e., the percentile where the calibration of survey and income tax data starts. I presented a different framework than BFM but with the same aim of preserving the continuity of the new, calibrated density function. Because the goal is the same, the two methods can lead to the same or similar results, as was the case for the illustrative dataset in Section 3.1.2. Here I present a more systematic comparison using simulated Monte Carlo data.

The setup of the Monte Carlo simulation largely, but not entirely, follows BFM. A population of 9 million is obtained by taking the exponent of a draw from standard normal distribution (which corresponds to a lognormal distribution). In each iteration, 1 % of the population is sampled. The probability of response is 50 % until the 90^{th} percentile and then decreases linearly with rank until nearly reaching 0 %. The probability of misreporting is 20 % until the 95^{th} percentile and then increases linearly with rank until almost 100 %. The distribution of misreported income is again lognormal but independent of the true income. In addition, the true income distribution is recorded in the income tax data and available in an aggregated tabulated form.[10]

The deviation from the setup of BFM lies in the tax data quality. The tabulated tax data in my setup are accurate for the entire distribution, while BFM assume accuracy only from the 90^{th} percentile onward and a downward bias until that point. I discuss the sensitivity of results to this change later in this Section.

I perform 1,000 iterations of the setup and apply both merging point approaches to the simulated data. My approach is applied in two variants which differ in handling situations with more than one candidate merging point. The first variant, denoted K, is the one I describe in Section 3.1.2 and apply in the empirical part: I determine candidate merging points, i.e., percentiles for which the test statistic is considered low, and, if there are more than one, I choose the highest candidate merging point. In the second variant, denoted $K direct$, I proceed directly with the percentile which has the lowest test statistic.

Table 1 reports the results. Had the non-response profile of the population been publicly known, the hypothetical researcher should set the merging point at the 90^{th} percentile. Such merging point would correct the bias at the top in its entirety while minimizing the survey distortion below this point. The benchmark method K generally identifies the merging point closer to the 90^{th} percentile. In 43.5 % of simulations, the K optimal merging point was between percentiles 0.89 and 0.91, while the estimator of BFM, denoted BFM, was within this range in 29 % of simulations. On the other hand, my method located the merging point at percentile 0.92 in two iterations of the simulation, in which case it did not correct the entire bias.

Table 1 Distribution of optimal merging point estimates.

MP method	0-0.79	0.80-0.88	0.89-0.91	0.92-1	Total
BFM	32.1	38.9	29.0	0.0	100
K	15.4	40.9	43.5	0.2	100
K direct	85.9	11.2	2.9	0.0	100

Note: Table 1 reports the share of merging point estimates which fall within each range. The bias is set to start at percentile 0.9 in the Monte Carlo simulation.

[10] Defining the income brackets of the tabulated tax data introduces a degree of arbitrariness into the simulation. I define reasonably fine brackets of size 0.2 from income values 0 to 2, of size 0.4 from income values 2 to 6.2, complemented by 5 brackets at the top with lower bounds at income values 7, 8, 10, 12, and 15.

The *Kdirect* method, which directly chooses the percentile with the lowest test statistic, performed poorly. Intuitively, this is due to the the simulation's setup in which there is no differential bias in the survey nor in the tax data below the 90^{th} percentile. Consequently, the two densities could cross multiple times due to sampling error in the survey, leading to many candidate merging points. This result highlights the need to consider all candidate merging points, as discussed in Section 3.1.2.

Figure 6 illustrates the results by plotting different optimal merging point approaches against each other. The *BFM* and *K* methods are visually similar, with the exception of several cases where the *BFM* merging point lies around the 15^{th} percentile. One potential explanation is that the tabulated (and subsequently interpolated) income tax data are not accurate at the bottom of the distribution. In any case, the *K* method is not sensitive to such an issue. Comparing the *K* and *K direct* methods, it is apparent that the *K direct* method serves as a lower bound to the *K* merging point. This follows from the definition of the two approaches, where *K direct* is defined as the candidate merging point with the lowest test statistic, and the *K* merging point is the largest candidate merging point.

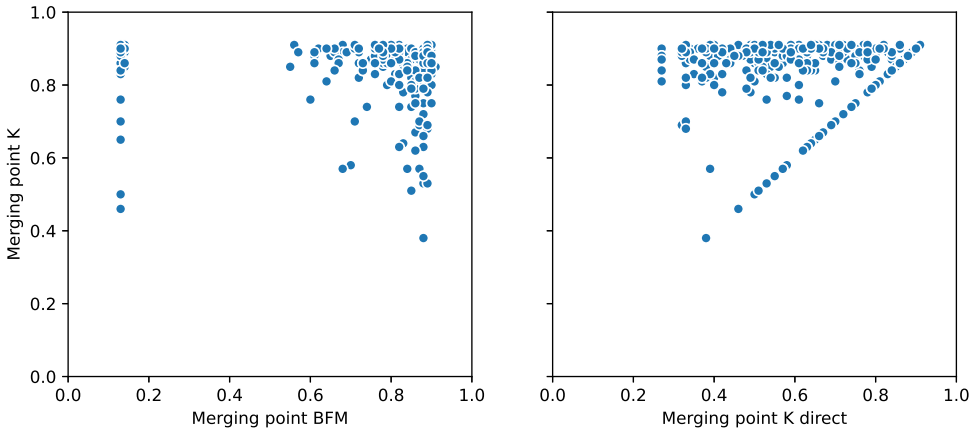

Fig. 6 The correlation of optimal merging point approaches. For visualization purposes, only 300 simulations are shown.

Merging point choice only impacts income inequality estimates to a small degree: The correlation coefficient between the top 1 % shares based on *K* and *BFM* merging points is 0.999. For the Gini coefficient the figure is slightly lower (0.953) because Gini considers the entire distribution. This merging point irrelevance is due to assumptions that the tax distribution is accurate everywhere and that the survey is not systematically biased below the 90^{th} percentile. These assumptions may not hold in real life and I still consider it desirable to aim for as low a survey distortion as possible, which in this setup means a merging point at the 90^{th} percentile. Finally, it was already established by BFM that the estimates themselves are significantly closer to the true value and with less variance than the original survey estimates.

In this simulation, I assume that the tabulated tax data accurately represent the entire distribution. BFM instead assume that they are downward biased up until the top 10 %. Under this assumption, the *K* and *BFM* methods perform comparably. However, I consider this assumption quite extreme because it implies that the unbiasedness of survey and tax data overlaps at precisely one point, the 90^{th} percentile. When the survey and tax data are unbiased over a larger interval, say from the 70^{th} to the 90^{th} percentile, my method will again tend to estimate the merging point closer to the optimal value, which is the 90^{th} percentile. Detailed results of Monte Carlo simulations under these alternative assumptions are provided in Appendix B.

4.2 Empirical data

4.2.1 Survey and tax income distributions

The merging point approach developed in this paper allows for an intuitive visual comparison of survey and tax distributions. This is shown in Figure 7, where I also include the income distribution of the EU-SILC survey. While the main aim of HFCS is to record the distribution of wealth, EU-SILC is the benchmark EU survey for income. The comparison of Austrian EU-SILC and HFCS income distributions is nevertheless striking: EU-SILC fits the tax data quite well (albeit not perfectly as it appears to overrepresent several percentiles near the top). In contrast, the downward bias in HFCS is observable as early as before the 80$^{\text{th}}$ percentile.

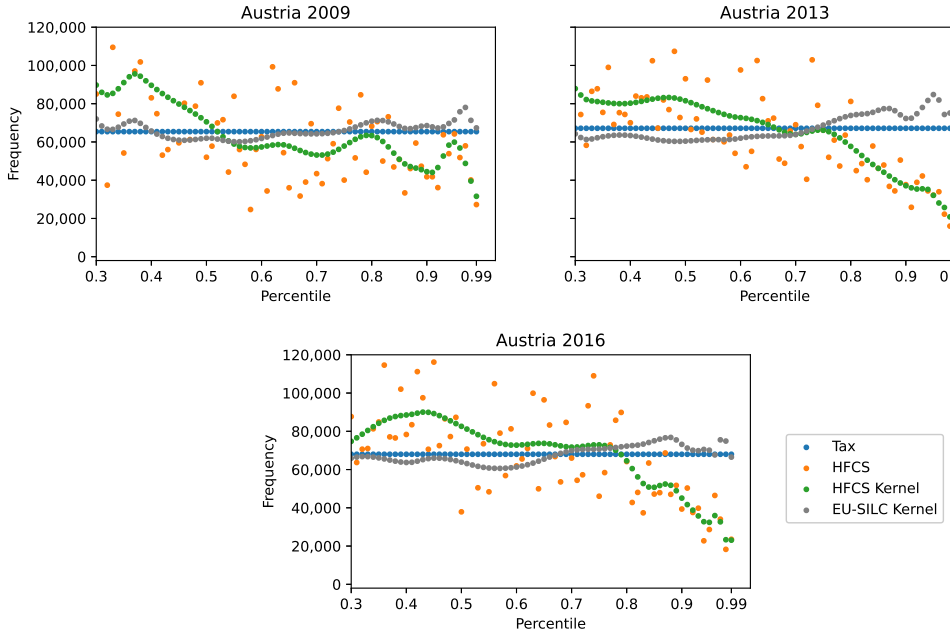

Fig. 7 A comparison of survey (EU-SILC, HFCS) and tax distributions. Percentile-based brackets are computed using the HFCS adult population size. Reference years for income and wealth differ in HFCS. HFCS data are averaged over implicates.

Austria is a country with no recorded oversampling strategy in the second and third waves and with a basic oversampling of Vienna in the first wave. Indeed, the first-wave survey appears to have the best fit to the income tax data, although the EU-SILC distribution remains superior. Another important difference is the sample size. The HFCS distribution is based on 4,140, 5,074 and 5,223 adult individuals in the first, second and third waves, respectively. The corresponding sample sizes in the Austrian EU-SILC data are more than twice as large, between 10,500 and 11,000.[11] EU-SILC also takes labor, pension and unemployment income data from public registers (Heuberger et al, 2013), which eliminates the bias arising from the untruthful reporting of these variables.

The comparison of survey and tax densities, which is part of my optimal merging point approach, constitutes an informative external check of HFCS data quality. The main aim of HFCS is to measure wealth and not income, but if the wealth distribution is captured correctly (which cannot be checked externally), then the income

[11] As explained in Section 3.1.3, the sample size referred to in this paragraph consists of individuals aged 20 or over, as well as younger individuals with income over the threshold of 10,000 euro.

distribution should be as well. Figure 7 strongly suggests that this is not the case with Austrian HFCS data.

4.2.2 Coverage ratios

Coverage ratios express the aggregate value of a wealth or income concept in a survey as a percentage of the corresponding total in the official macroeconomic statistics, the National Accounts. Totals in the National Accounts are obtained by combining various administrative and other data sources, and are thus perceived to be of higher quality than totals in the survey. However, the net wealth concept in the HFCS does not have a highly comparable counterpart in the National Accounts.[12] I compute coverage ratios for Austria using two wealth concepts adopted from EG-LMM (2020). The *narrow financial wealth concept* consists of deposits, bonds, listed shares, mutual funds, minus liabilities. These are assessed by EG-LMM (2020) to have high conceptual comparability. The *adjusted broad wealth concept* additionally includes unlisted shares, dwellings, or the value of self-employment businesses, but the conceptual comparability between the survey and the National Accounts is lower.

Figure 8 shows coverage ratios in the Austrian HFCS before and after calibration (without any Pareto adjustment). Calibrating survey weights generally puts the survey total closer to the National Accounts. The impact of calibration is the smallest in the first wave, when the coverage ratio of the broad wealth concept increases from 88 % to 102 %. This suggests that the oversampling of Vienna in the first wave has helped to achieve a more representative survey sample, at least in terms of total wealth. In contrast, the largest impact occurs in the second wave, when the calibrated survey even exceeds its National Accounts counterpart: the coverage ratio of the broad wealth concept rises from 79 % to 122 %. The difference from the ideal ratio of 100 % thus remains approximately constant in absolute terms. Calibration increases the coverage ratio also of the narrow financial wealth concept, although it still remains notably undercovered. Overall, the results support the claim that calibration can improve the reliability of wealth surveys.

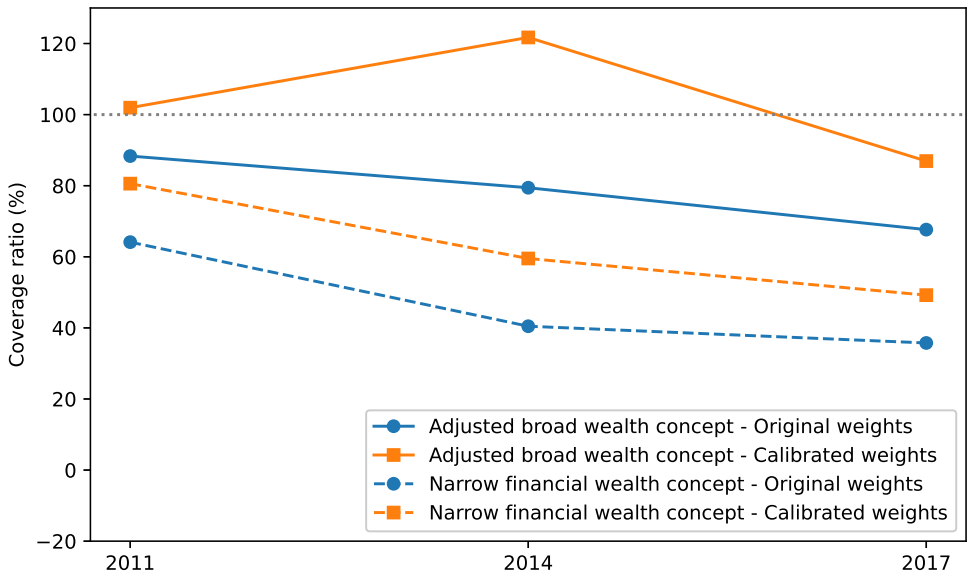

Fig. 8 Coverage ratios before and after calibration. Coverage ratios compare the survey totals with those in the National Accounts. Wealth concepts are adopted from EG-LMM (2020). Values are averaged over implicates.

[12]The most problematic components are dwellings and land, parts of non-listed business wealth, and consumer durables (see e.g. Waltl, 2022, Appendix A).

4.2.3 New wealth inequality estimates

Figure 9 illustrates the impact of calibration on the wealth distribution, which is the smallest in the first Austrian survey from 2011, when Vienna was oversampled. In that first wave, the downward bias of the survey income distribution is the smallest, as shown in Figure 7 in Section 4.2.1. In addition, there appears to be a weak correlation between top income holders in the survey (whose weights have been increased) and top wealth holders in 2011. In contrast, income calibration visibly increases survey weights of wealthy families in 2014 and 2017, as would be expected given the hypothesised underrepresentation of the rich. The distribution shifts to the right, meaning that estimated wealth is larger for each rank at the top (for example, 10^{-3}, which corresponds to a household in the top 0.1 %). Figure 9 also reveals the issue of insufficient coverage at the top, where only a few survey observations represent the wealthiest households.

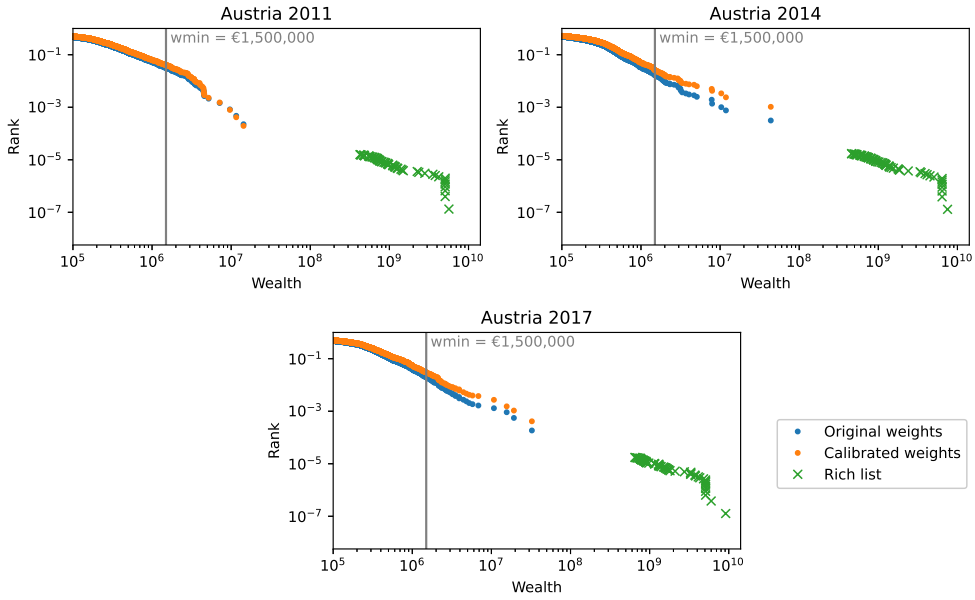

Fig. 9 Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, average over implicates. Figures for all five implicates are provided in Appendix D.

The impact of calibration on the Pareto tail estimation is highlighted in Figure 10 using Austrian 2017 data. The Pareto line estimated using calibrated weights is shifted to the right, implying larger inequality. In addition, the Pareto line for the calibrated weights starts at a lower rank than for the original weights, even though the w_{min} threshold is the same. This is because calibration has increased the population size above the threshold, meaning that the tail to be replaced is larger. This is encouraging because if the probability of response starts to decline below the Pareto threshold, the size of the top tail will originally be underestimated, biasing the results (Westermeier, 2016). When the list of the wealthiest individuals is included, the difference in the slope of the Pareto tail becomes negligible: the rich list "dominates" the OLS objective function, as explained in Section 3.2.2.

Table 2 presents a systematic comparison of the top 1 % wealth share estimates based on different weights and methods. As reported in the first column, survey weight calibration increases the top 1 % wealth share from 25.5 % to 36.6 % in 2014 and from 22.8 % to 27.3 % in 2017. In 2011, even though calibration increases the net worth of the top 1 %, the denominator (i.e., total wealth) increases even more, leading to a decrease in the top 1 % share from 23.2 % to 21.5 %. All these values, especially the

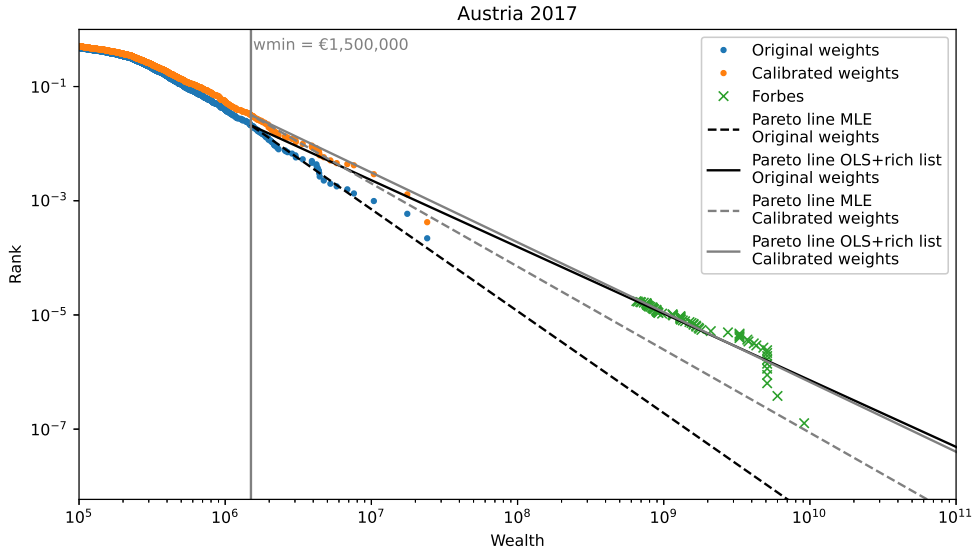

Fig. 10 Pareto distribution estimation. Data: HFCS, Austria 2017, average over implicates.

top share in 2014 with calibrated weights, are associated with a high level of uncertainty quantified by the bootstrap standard errors. In general, the precision of survey estimates depends on survey design: whether the sample size is sufficiently large and whether the wealthy are among the respondents in adequate numbers. Consequently, the observed drop in inequality between 2014 and 2017 could be due to measurement issues. A similar decrease is found in the German HFCS data, with the Deutsche Bundesbank suggesting such an explanation.[13]

Fitting a Pareto tail using survey data alone, without the rich list, leads to erratic results. They are sensitive to the choice of the estimation method and have large standard errors, likely due to the small number of wealthy households in the Austrian HFCS survey. As I report in Appendix C, the results are also sensitive to the choice of the estimation threshold w_{min}. Using the $\hat{\alpha}_{MLE}$ estimator, the impact of Pareto tail estimation is visible mainly in 2011 and in 2014 with calibrated weights. In 2014 with original weights and in 2017, the impact is smaller and there is even a slight decrease in the estimated top 1 % wealth share. The small sensitivity of estimates to the Pareto tail without a rich list is not unique to Austria, as reported in (Vermeulen, 2018, Table A3).[14] Including the rich list in the Pareto estimation confirms what is apparent in Figure 10: The top 1 % share increases and variance is almost eliminated. The increase is quite large, even compared to results based on calibrated weights. The only exception is the year 2014, where the survey with calibrated weights leads to similar results as the Pareto tail with a rich list. The drop in variance is optimistic but only to the extent that the rich list can be trusted. As explained in Section 3.2.2, if the rich list is not reliable, it will bias the results even if the survey data were accurate.

In Table 3, I present calculations of the Gini coefficient, which are broadly in line with the top 1 % shares. The impact of survey calibration is again positive in 2014 and 2017 and small and negative in 2011. The impact of Pareto fitting does not change either since it is the same Pareto tail that is being fitted. The Gini coefficient is a

[13] "In the wave 2017 it appears, in particular, that business assets in the top tail of the distribution were under-recorded. In addition, fewer very wealthy households participated in the survey compared with the survey waves in 2010 and 2014" (Deutsche Bundesbank, 2019; Albers et al, 2022).

[14] For completeness, I list reasons why my Pareto estimates may differ from those in Vermeulen (2018) even when non-calibrated weights are used and the w_{min} threshold is the same. First, after estimating the Pareto tail, I create synthetic households for the top tail rather than working with (conditional) expected values – see Section 3.2.3. Second, $\hat{\alpha}_{OLS}$ is estimated using a regression *without* intercept, which provides a better fit in the log-rank log-wealth graph. Third, due to calibration, I must work with my own bootstrap weights rather than those provided by HFCS – see Section 3.2.4.

Table 2 Top 1 % share estimates.

	Survey	Pareto + Survey MLE	Pareto + Survey OLS	Pareto + Survey + Rich list
2011, original weights	23.2	32.1	29.2	41.3
	(7.3)	(18.5)	(16.5)	(1.3)
2011, calibrated weights	21.5	29.3	26.9	39.6
	(7.2)	(17.8)	(16.2)	(0.9)
2014, original weights	25.5	24.1	28.7	39.9
	(8.0)	(12.8)	(15.8)	(1.2)
2014, calibrated weights	36.6	47.6	53.8	38.4
	(14.9)	(29.2)	(32.6)	(1.3)
2017, original weights	22.8	20.8	23.7	44.0
	(5.8)	(5.1)	(7.1)	(0.9)
2017, calibrated weights	27.3	26.4	29.3	43.0
	(6.8)	(12.0)	(12.0)	(0.8)

Note: Pareto threshold 1.5 million euro. Bootstrap standard errors in parentheses. Results for different thresholds are provided in Appendix C

more robust measure than the top 1 % share and the differences between methods are generally lower, as are the standard errors. In Appendix C, I provide the top 1 % share estimates for different thresholds, namely 1 million euro and 2 million euro. The sensitivity of Pareto estimates to the threshold choice is quite high, the only exception again being the inclusion of the rich list.

Table 3 Gini coefficient estimates.

	Survey	Pareto + Survey MLE	Pareto + Survey OLS	Pareto + Survey + Rich list
2011, original weights	76.2	78.7	77.7	81.9
	(4.0)	(6.8)	(6.3)	(1.9)
2011, calibrated weights	75.0	77.2	76.4	81.0
	(4.3)	(7.0)	(6.6)	(1.9)
2014, original weights	73.1	72.6	74.3	78.4
	(3.0)	(4.7)	(5.8)	(0.9)
2014, calibrated weights	76.7	80.9	83.2	77.6
	(6.0)	(11.0)	(12.2)	(1.2)
2017, original weights	73.0	72.3	73.4	80.6
	(2.2)	(2.0)	(2.7)	(0.8)
2017, calibrated weights	74.8	74.7	75.8	80.7
	(2.7)	(4.6)	(4.5)	(0.9)

Note: Pareto threshold 1.5 million euro. Bootstrap standard errors in parentheses.

Estimates based on calibrated weights tend to exhibit larger standard errors than those based on original weights. This is because calibration includes one additional source of uncertainty: imputed missing values in the income part of the survey. Variance in the original estimates, on the other hand, is only due to sampling and due to imputed wealth values. It is a positive sign that standard errors tend to decrease with each HFCS wave, suggesting that imputed missing values are becoming less of a concern. In Appendix D, I provide the log-rank log-wealth graphs for all five implicates, in which the only difference is the imputed missing values.

5 Conclusion

The problem of the missing rich in wealth surveys prevents their use for reliable wealth inequality estimates. This paper contributes to the existing body of literature that seeks to improve surveys using external sources and statistical adjustments. I develop a new approach that makes wealth surveys consistent with income tax data, a high-quality external source. The method avoids arbitrary assumptions to the extent possible, preserves the main socio-demographic characteristics and keeps the survey structure intact.

I apply the new method to three waves of the Austrian Household Finance and Consumption Survey, combined with the tabulated tax data. Calibration increases wealth inequality estimates in the second and third waves of the survey: the top 1 % wealth share rises from 26 % to 37 % in 2014 and from 23 % to 27 % in 2017. The effect is small and negative in the first wave, where the top 1 % share declines from 23 to 22 %. A potential explanation for this small effect is the existence of an (albeit basic) oversampling strategy in the first wave, which should reduce the missing rich problem. In contrast, there was no oversampling strategy recorded in the subsequent two waves. My merging point algorithm also reveals a strong bias in the Austrian HFCS income data when compared with the tax distribution and even with another survey, EU-SILC.

Combining my method with other estimation techniques, I find that fitting a Pareto tail produces results that are sensitive to the estimation method and the Pareto lower bound. While including a rich list eliminates these issues, this is because the rich list overshadows the survey data and "dominates" the estimation of the Pareto coefficient. Any errors in the rich list thus have direct consequences for the inequality estimates.

Application of the presented calibration method consists of three main steps. First, income tax data must be carefully matched with the income concepts in the survey. This involves considering the definition of the tax unit, the types of income included in the tax data, or whether taxes or social security contributions have been deducted. Adjustment may also be made to the tax data themselves, but they should remain continuous, as will be the case if interpolation is applied to tabulated data. Once the income concepts are harmonized, the second step is to apply the calibration method of BFM to the survey's income distribution. This only changes survey *weights*, not the reported income, and makes the income distribution in the survey consistent with the tax data. I use my own algorithm to determine the optimal merging point from which calibration should start. This algorithm includes an intuitive visual comparison of the survey and tax income distributions. Finally, the newly calibrated weights can be used along with the recorded wealth data to estimate wealth inequality.

The presented method can be useful for researchers who work with surveys to study the top of the wealth distribution (e.g., Garbinti et al, 2021; Palomino et al, 2022). In my optimal merging point algorithm, I also propose a visual comparison of survey and tax income densities that can be utilized in research on how these distributions differ (e.g., Yonzan et al, 2022; Bartels and Metzing, 2019). A key prerequisite for the successful application of my method is the availability of reliable income tax data and the correct matching of this data to income concepts in the survey.

References

Albers T, Bartels C, Schularick M (2022) Wealth and its Distribution in Germany, 1895-2018. CESifo Working Paper 9739. https://doi.org/10.2139/ssrn.4103952, URL https://papers.ssrn.com/abstract=4103952

Alstadsæter A, Johannesen N, Zucman G (2019) Tax Evasion and Inequality. American Economic Review 109(6):2073–2103. https://doi.org/10.1257/aer.20172043, URL https://www.aeaweb.org/articles?id=10.1257/aer.20172043

Alvaredo F, Atkinson AB, Bauluz L, et al (2021) Distributional National Accounts Guidelines: Methods and Concepts Used in the World Inequality Database. WID-world Working Paper

Bach S, Thiemann A, 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. https://doi.org/10.1007/s10797-019-09578-1, URL https://doi.org/10.1007/s10797-019-09578-1

Barcaroli G, Ilardi G, Neri A, et al (2021) Optimal sampling design for household finance surveys using administrative income data. Rivista di statistica ufficiale 2021(2)

Bartels C, Metzing M (2019) An integrated approach for a top-corrected income distribution. The Journal of Economic Inequality 17(2):125–143. https://doi.org/10.1007/s10888-018-9394-x, URL https://doi.org/10.1007/s10888-018-9394-x

Blanchet T (2016) Wealth inequality in Europe and in the United States: estimations from surveys, national accounts and wealth rankings. Paris School of Economics Master Thesis

Blanchet T, Chancel L, Gethin A (2022a) Why Is Europe More Equal than the United States? American Economic Journal: Applied Economics 14(4):480–518. https://doi.org/10.1257/app.20200703, URL https://www.aeaweb.org/articles?id=10.1257/app.20200703

Blanchet T, Flores I, Morgan M (2022b) The weight of the rich: improving surveys using tax data. The Journal of Economic Inequality 20(1):119–150. https://doi.org/10.1007/s10888-021-09509-3, URL https://doi.org/10.1007/s10888-021-09509-3

Blanchet T, Fournier J, Piketty T (2022c) Generalized Pareto Curves: Theory and Applications. Review of Income and Wealth 68(1):263–288. https://doi.org/10.1111/roiw.12510, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12510

Brzeziński M, Sałach K, Wroński M (2020) Wealth inequality in Central and Eastern Europe: Evidence from household survey and rich lists' data combined. Economics of Transition and Institutional Change 28(4):637–660. https://doi.org/10.1111/ecot.12257, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecot.12257

Buskirk TD (1998) Nonparametric density estimation using complex survey data. In: Proceedings of the Survey Research Methods Section at JSM1998. American Statistical Association, pp 799–801

Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. SIAM Review 51(4):661–703. https://doi.org/10.1137/070710111, URL

https://epubs.siam.org/doi/abs/10.1137/070710111

Cowell F (2011a) Inequality Among the Wealthy. LSE STICERD Research Paper CASE/150

Cowell F (2011b) Measuring Inequality. Oxford University Press

Cowell F, Flachaire E (2015) Statistical Methods for Distributional Analysis. In: Atkinson AB, Bourguignon F (eds) Handbook of Income Distribution. Elsevier, p 359–465, URL https://www.sciencedirect.com/science/article/pii/B9780444594280000072

Dalitz C (2016) Estimating Wealth Distribution: Top Tail and Inequality. Hochschule Niederrhein Technische Berichte 2016-01

Deutsche Bundesbank (2019) Vermögen und Finanzen privater Haushalte in Deutschland: Ergebnisse der Vermögensbefragung 2017. Monthly Report 13

Disslbacher F, Ertl M, List E, et al (2020) On Top of the Top: Adjusting wealth distributions using national rich lists. INEQ Working Paper Series 20

Eckerstorfer P, Halak J, Kapeller J, et al (2016) Correcting for the Missing Rich: An Application to Wealth Survey Data. Review of Income and Wealth 62(4):605–627. https://doi.org/10.1111/roiw.12188, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12188

EG-LMM (2020) Understanding household wealth: Linking macro and micro data to produce distributional financial accounts. Statistics Paper Series 37. URL https://www.econstor.eu/handle/10419/234507

European Central Bank (2020) The Household Finance and Consumption Survey: Methodological report for the 2017 wave. Statistics Paper Series, European Central Bank

Gabaix X, Ibragimov R (2011) Rank — 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents. Journal of Business & Economic Statistics 29(1):24–39. URL https://www.jstor.org/stable/25800776

Garbinti B, Goupille-Lebret J, Piketty T (2021) Accounting for Wealth-Inequality Dynamics: Methods, Estimates, and Simulations for France. Journal of the European Economic Association 19(1):620–663. https://doi.org/10.1093/jeea/jvaa025, URL https://doi.org/10.1093/jeea/jvaa025

Heuberger R, Glaser T, Kafka E (2013) 10. The use of register data in the Austrian SILC survey. In: Jäntti M, Törmälehto VM, Marlier E (eds) The use of registers in the context of EU–SILC: challenges and opportunities. Publications Office of the European Union, p 359–465

Hlasny V, Verme P (2018) Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data. Econometrics 6(2):30. https://doi.org/10.3390/econometrics6020030, URL https://www.mdpi.com/2225-1146/6/2/30

Jestl S, List E (2020) Distributional National Accounts (DINA) for Austria, 2004-2016. wiiw Working Paper 175. URL https://wiiw.ac.at/p-5220.html

Kennickell AB (2019) The tail that wags: differences in effective right tail coverage and estimates of wealth inequality. The Journal of Economic Inequality 17(4):443–459. https://doi.org/10.1007/s10888-019-09424-8, URL https://doi.org/10.1007/s10888-019-09424-8

Kennickell AB (2021) Chasing the Tail: A Generalized Pareto Distribution Approach to Estimating Wealth Inequality. Stone Center Working Paper Series 37. URL https://ideas.repec.org/p/osf/socarx/u3zs2.html

Kennickell AB, Woodburn RL (1999) Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth. Review of Income and Wealth 45(2):193–215. https://doi.org/10.1111/j.1475-4991.1999.tb00328.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4991.1999.tb00328.x

Kennickell AB, Lindner P, Schürz M (2022) A new instrument to measure wealth inequality: distributional wealth accounts. Monetary Policy & the Economy 21(Q4). URL https://ideas.repec.org/a/onb/oenbmp/y2022iq4-21b3.html

Lustig N (2019) The "Missing Rich" in Household Surveys: Causes and Correction Approaches. Commitment to Equity (CEQ) Working Paper Series 75. URL https://ideas.repec.org/p/tul/ceqwps/75.html

Palomino JC, Marrero GA, Nolan B, et al (2022) Wealth inequality, intergenerational transfers, and family background. Oxford Economic Papers 74(3). https://doi.org/10.1093/oep/gpab052, URL https://doi.org/10.1093/oep/gpab052

Rytgaard M (1990) Estimation in the Pareto Distribution. Astin Bulletin 20(2):201–216. https://doi.org/10.2143/AST.20.2.2005443, URL https://www.cambridge.org/core/journals/astin-bulletin-journal-of-the-iaa/article/estimation-in-the-pareto-distribution/FB32B606B03054E8B08378FD896EEDF0

Sen AK (1997) From Income Inequality to Economic Inequality. Southern Economic Journal 64(2):384–401. https://doi.org/10.2307/1060857, URL https://www.jstor.org/stable/1060857

Singh AC, Mohl CA (1996) Understanding calibration estimators in survey sampling. Statistics Canada URL https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X19960022973

Statistics Austria (2016) Integrierte Statistik der Lohn-und Einkommensteuer

Taleb NN, Douady R (2015) On the super-additivity and estimation biases of quantile contributions. Physica A: Statistical Mechanics and its Applications 429:252–260. https://doi.org/10.1016/j.physa.2015.02.038, URL https://www.sciencedirect.com/science/article/pii/S0378437115001429

Vermeulen P (2014) How fat is the top tail of the wealth distribution? European Central Bank Working Paper Series 1692. URL https://ideas.repec.org/p/ecb/ecbwps/20141692.html

Vermeulen P (2016) Estimating the Top Tail of the Wealth Distribution. The American Economic Review 106(5):646–650. URL http://www.jstor.org/stable/43861099

Vermeulen P (2018) How Fat is the Top Tail of the Wealth Distribution? Review of Income and Wealth 64(2):357–387. https://doi.org/10.1111/roiw.12279, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12279

Wald A (1945) Statistical Decision Functions Which Minimize the Maximum Risk. Annals of Mathematics 46(2):265–280. https://doi.org/10.2307/1969022, URL https://www.jstor.org/stable/1969022

Waltl SR (2022) Wealth Inequality: A Hybrid Approach Toward Multidimensional Distributional National Accounts In Europe. Review of Income and Wealth 68(1):74–108. https://doi.org/10.1111/roiw.12519, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12519

Waltl SR, Chakraborty R (2022) Missing the wealthy in the HFCS: micro problems with macro implications. The Journal of Economic Inequality 20(1):169–203. https://doi.org/10.1007/s10888-021-09519-1, URL https://doi.org/10.1007/s10888-021-09519-1

Westermeier C (2016) Estimating top wealth shares using survey data - An empiricist's guide. Discussion Papers Free University Berlin, School of Business & Economics 2016(21). URL https://ideas.repec.org/p/zbw/fubsbe/201621.html

Wildauer R, Kapeller J (2019) A comment on fitting Pareto tails to complex survey data. ICAE Working Paper Series 102. URL https://www.econstor.eu/handle/10419/206424

Yonzan N, Milanovic B, Morelli S, et al (2022) Drawing a Line: Comparing the Estimation of Top Incomes between Tax Data and Household Survey Data. The Journal of Economic Inequality 20(1):67–95. https://doi.org/10.1007/s10888-021-09515-5, URL https://doi.org/10.1007/s10888-021-09515-5

Appendix A Extension of the merging point algorithm to more granular frequency brackets

In some cases, especially when the merging point is presumed to be near the top of the income distribution, it may be desirable to use narrower frequency brackets than those based on percentiles. I illustrate such extension using Austrian EU-SILC data. Figure A1 introduces the dataset. The survey kernel density follows the tax density quite closely until around the 90th percentile, after which it starts to appear "bumpy" as the income intervals underlying each bracket become larger. Although the survey and tax distributions cross at around the 97th-98th percentile, my optimal merging point algorithm does not identify any candidate merging points there. This is because while the algorithm aims to preserve the continuity of the density function, the distribution was not visually continuous even before the calibration.

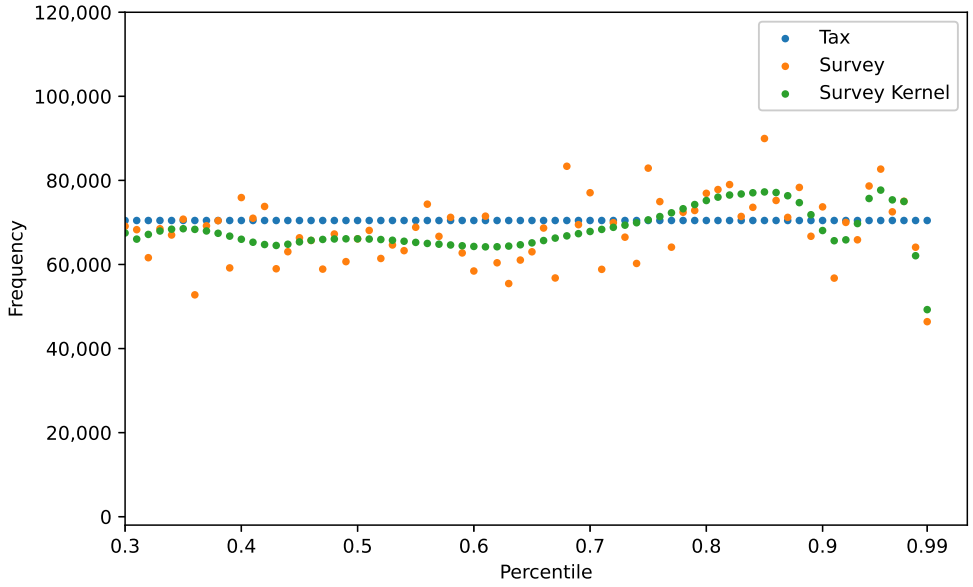

Fig. A1 Tax and survey frequencies. Gaps in the survey kernel distribution suggest more granular frequency brackets are desirable.

Figure A2 shows how to make the comparison of survey and tax distributions more informative and suitable for my optimal merging point algorithm. I apply the Generalized Pareto interpolation to the tax data again, creating finer brackets of the size 1/4 of a percent. All the other steps of the merging point method, including the algorithm for choosing candidate merging points, remain the same as in the original setup. Narrowing the bracket size solves the problem of the survey kernel density's "bumpiness". The merging point algorithm now identifies a candidate merging point at the 98th percentile, more specifically at its first quartile. In contrast, the largest candidate merging point using the standard bracket size was the 94th percentile, in addition to the 90th and 75th percentiles. Researchers wishing to calibrate this illustrative dataset can thus choose between four candidate merging points, all of which lead to a visually continuous density after calibration.

When the approach of BFM is applied to this dataset, it exhibits sensitivity to the choice of the trustable span, i.e., the left-bounded interval on which the tax data

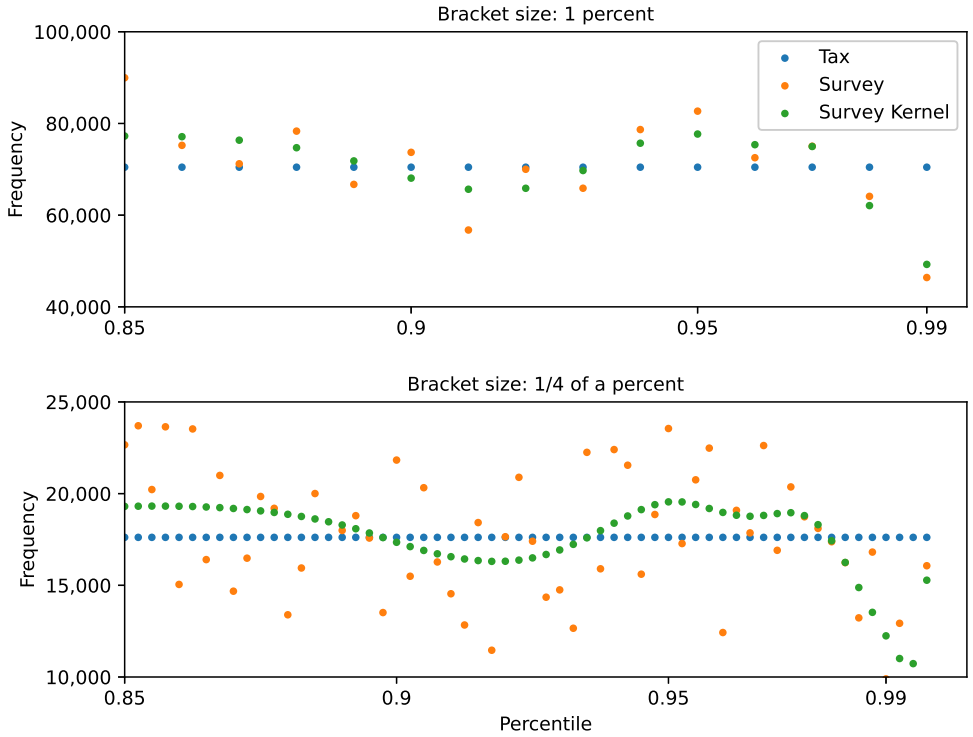

Fig. A2 Comparing tax and survey distributions using more granular frequency brackets. Y-axis scale in the bottom graph is 4 times lower than in the top graph. Same dataset as in Figure A1. Small random noise is added to the survey distribution in the bottom graph to ensure confidentiality.

are considered reliable.[15] My method does not suffer from this issue (of course, the trustable span must remain large enough to include the candidate merging point), which is another reason why it should be considered an alternative to the merging point algorithm of BFM.

[15] For example, if the trustable span (*ts*) starts at percentile 0.4, the optimal merging point (*omp*) is the percentile 0.42. If *ts* starts at 0.5, *omp* is 0.76; if *ts* starts at 0.6, *omp* is 0.97.

Appendix B Monte Carlo simulation: Alternative scenarios

In my benchmark simulation in Section 4.1, I have deviated from the setup of BFM by assuming that the tabulated tax data accurately represent the entire distribution. Here I report results of two Monte Carlo simulations which assume that the tax data are biased. These two simulations are conducted using the population size of 1 million due to computational demands associated with using the benchmark 9 million population. I report the sensitivity to the population size in the third simulation.

For the first specification, bias is introduced by multiplying each tax bracket by a coefficient that is 0 until the 50^{th} percentile and then increases linearly with rank until the 90th percentile, at which it reaches and sustains the value of 1. This is in line with the simulation in BFM and the results are reported in Table B1. The difference between the three optimal merging point approaches, *BFM*, *K* and *K direct*, is minimal.

Table B1 Distribution of optimal merging point estimates: Tax data biased until the 90^{th} percentile

MP method	0-0.79	0.80-0.88	0.89-0.91	0.92-1	Total
BFM	0.0	0.0	86.8	13.2	100
K	0.0	0.0	85.5	14.5	100
K direct	0.0	0.0	85.6	14.4	100

Note: Table B1 reports the share of merging point estimates which fall within each range. This scenario reflects the assumption of BFM whereby there is a downward bias in the tax data up until the 90^{th} percentile.

The second setup instead assumes that the bias exists only until the 70th percentile (the bias is modeled analogously to the previous simulation). Table B2 reports the results. Once there is a larger interval on which the survey and tax data have no systematic bias, the results become in line with my benchmark specification: The *K* method identifies the merging point near the optimal value, the 90^{th} percentile, more often than the *BFM* method. The *K direct* method again performs the worst.

Table B2 Distribution of optimal merging point estimates: Tax data biased until the 70^{th} percentile

MP method	0-0.79	0.80-0.88	0.89-0.91	0.92-1	Total
BFM	16.9	41.3	37.9	3.9	100
K	6.1	42.2	45.2	6.5	100
K direct	43.3	43.7	11.9	1.1	100

Note: Table B2 reports the share of merging point estimates which fall within each range. This scenario assumes that the tax data are biased until the 70^{th} percentile and accurate afterwards.

Finally, I report the sensitivity of results to the population size. The Monte Carlo simulations in this Appendix are computed using a population of 1 million, which is much less than the 9 million population in BFM. Population size determines the gross sample size, which is 1 % of the total. The smaller population was chosen due to computational demands of my method, which relies on integrating the survey's adaptive kernel density over each percentile-based bracket. Using the benchmark setup (i.e., with unbiased tax data), I run two new simulations with population size of 1 and 2 million, respectively. Results are compared in Table B3. When the population size is increased, the performance of both methods generally tends to improve, but the *K* method remains superior. This suggests that the results are not sensitive to the population size.

Table B3 Distribution of optimal merging point estimates: Changing the population size

MP method	0-0.79	0.80-0.88	0.89-0.91	0.92-1	Total
BFM 1 mil.	59.6	26.4	13.5	0.5	100
K 1 mil.	6.0	42.5	45.1	6.4	100
K direct 1 mil.	86.7	10.1	3.2	0.0	100
BFM 2 mil.	52.1	27.1	20.6	0.2	100
K 2 mil.	6.0	39.1	51.1	3.8	100
K direct 2 mil.	88.0	9.5	2.2	0.3	100
BFM 9 mil.	32.1	38.9	29.0	0.0	100
K 9 mil.	15.4	40.9	43.5	0.2	100
K direct 9 mil.	85.9	11.2	2.9	0.0	100

Note: Table B3 reports the share of merging point estimates which fall within each range. Three scenarios are compared, differing only in the population size. All simulations assume unbiased tax data.

Appendix C Top 1 % wealth share estimates with different Pareto thresholds

Table C4 Top 1 % shares with different thresholds, Pareto MLE estimates.

	Survey	1 mil.	1.5 mil.	2 mil.
2011, original weights	23.2	37.9	32.1	30.3
	(7.3)	(22.3)	(18.5)	(16.8)
2011, calibrated weights	21.5	39.3	29.3	29.3
	(7.2)	(24.6)	(17.8)	(17.4)
2014, original weights	25.5	22.3	24.1	31.2
	(8.0)	(7.4)	(12.8)	(23.4)
2014, calibrated weights	36.6	33.5	47.6	64.4
	(14.9)	(18.8)	(29.2)	(39.1)
2017, original weights	22.8	21.9	20.8	23.1
	(5.8)	(4.1)	(5.1)	(9.2)
2017, calibrated weights	27.3	28.3	26.4	30.1
	(6.8)	(9.9)	(12.0)	(17.7)

Note: Bootstrap standard errors in parentheses.

Table C5 Top 1 % shares with different thresholds, Pareto OLS estimates.

	Survey	1 mil.	1.5 mil.	2 mil.
2011, original weights	23.2	33.3	29.2	27.8
	(7.3)	(19.8)	(16.5)	(15.2)
2011, calibrated weights	21.5	32.7	26.9	26.0
	(7.2)	(21.4)	(16.2)	(15.7)
2014, original weights	25.5	25.1	28.7	34.7
	(8.0)	(9.6)	(15.8)	(23.9)
2014, calibrated weights	36.6	36.8	53.8	66.9
	(14.9)	(20.6)	(32.6)	(41.4)
2017, original weights	22.8	23.0	23.7	25.6
	(5.8)	(5.1)	(7.1)	(10.7)
2017, calibrated weights	27.3	29.3	29.3	32.6
	(6.8)	(10.1)	(12.0)	(16.3)

Note: Bootstrap standard errors in parentheses.

Table C6 Top 1 % shares with different thresholds, OLS + rich list estimates.

	Survey	1 mil.	1.5 mil.	2 mil.
2011, original weights	23.2	41.2	41.3	40.8
	(7.3)	(1.0)	(1.3)	(1.9)
2011, calibrated weights	21.5	39.3	39.6	38.9
	(7.2)	(0.8)	(0.9)	(1.6)
2014, original weights	25.5	41.5	39.9	38.0
	(8.0)	(0.6)	(1.2)	(2.3)
2014, calibrated weights	36.6	39.5	38.4	37.6
	(14.9)	(0.7)	(1.3)	(2.2)
2017, original weights	22.8	44.8	44	41.8
	(5.8)	(0.6)	(0.9)	(1.5)
2017, calibrated weights	27.3	43.1	43	41.9
	(6.8)	(0.6)	(0.8)	(1.5)

Note: Bootstrap standard errors in parentheses.

Appendix D Differences between implicates

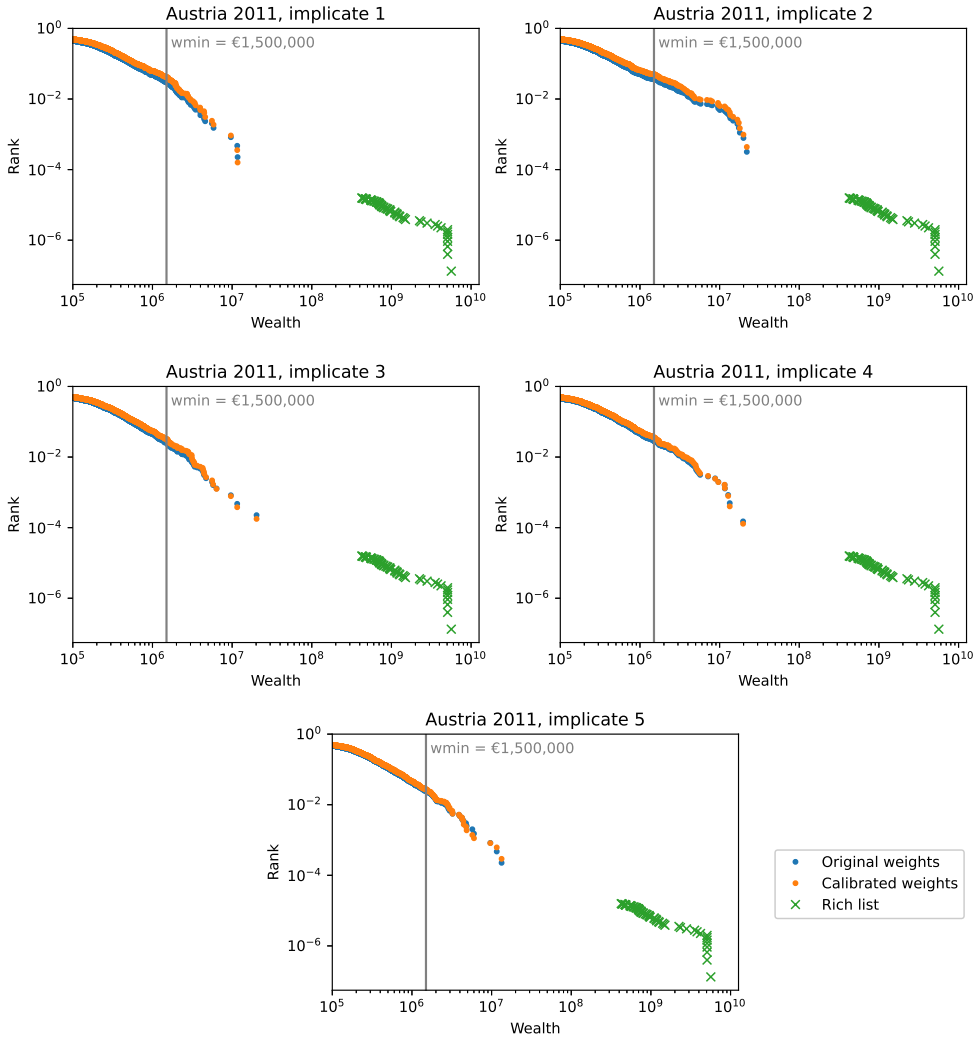

Fig. D3 Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, First wave.

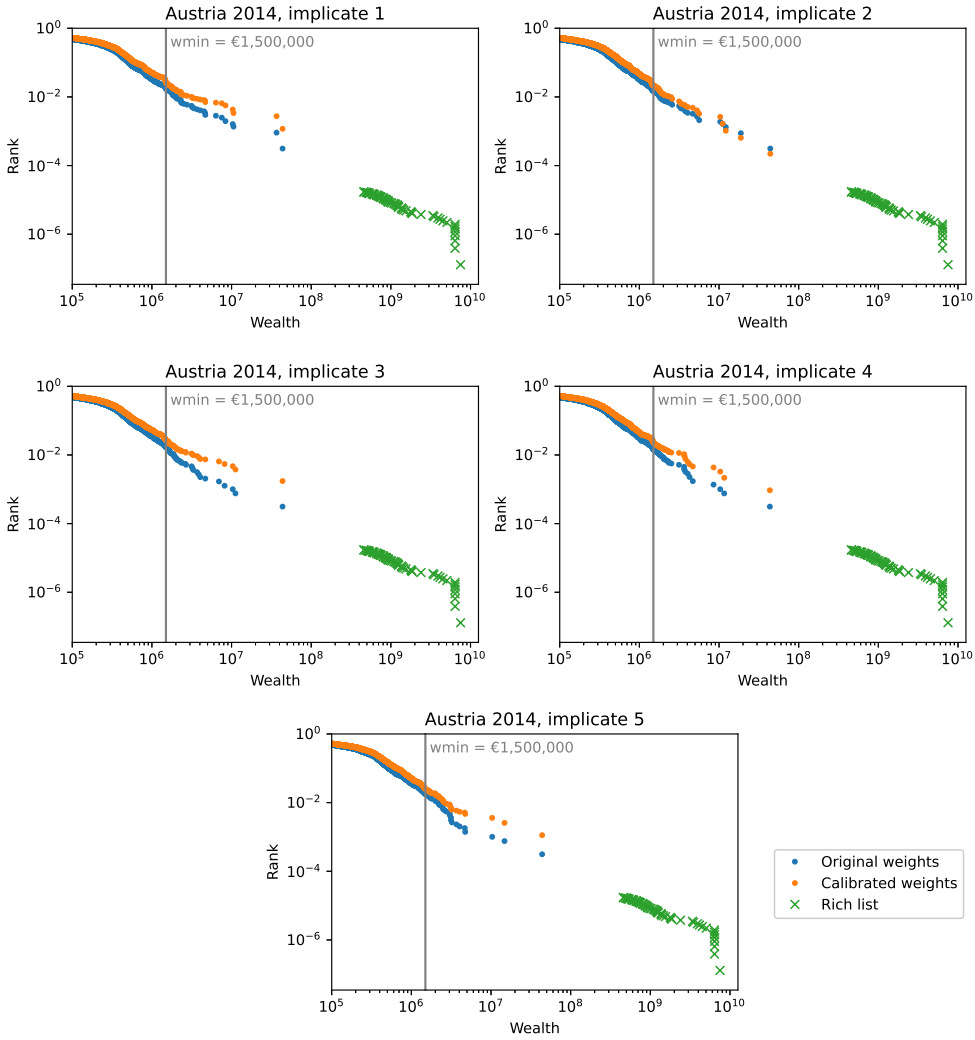

Fig. D4 Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, Second wave.

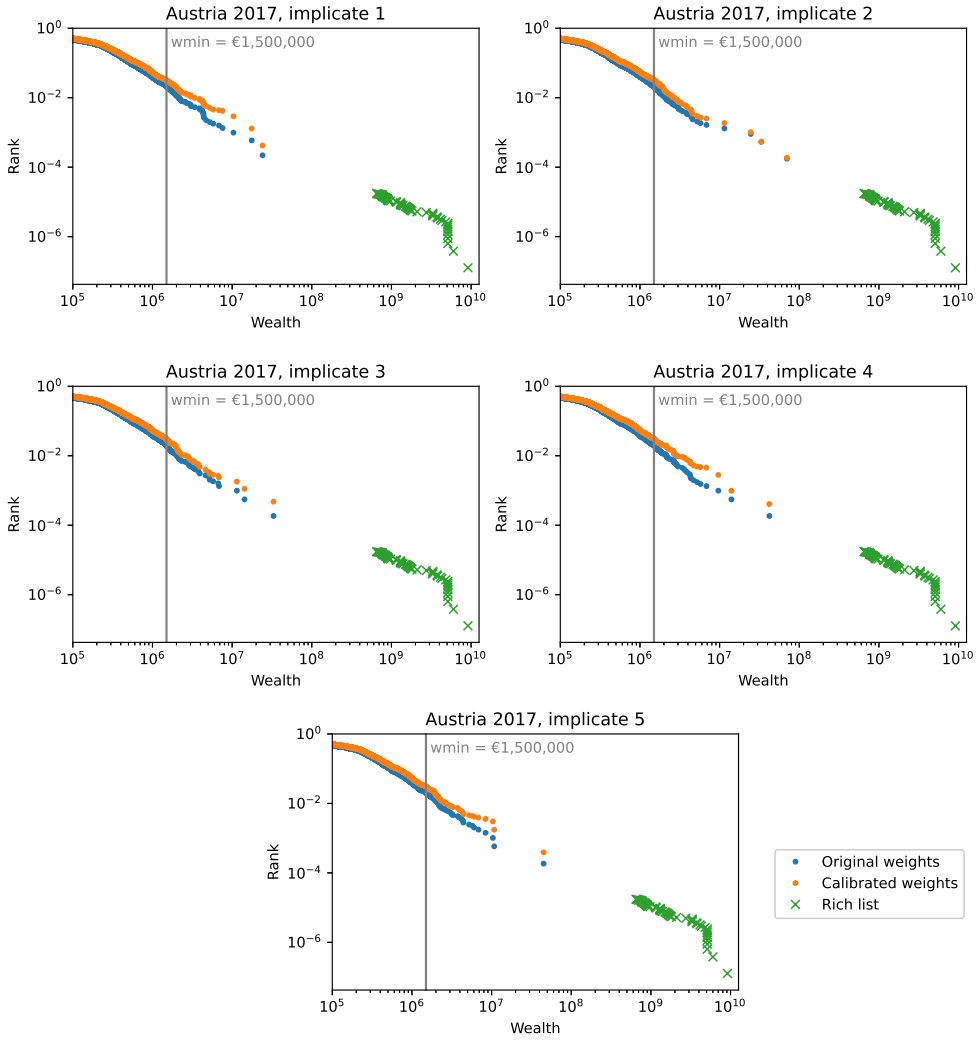

Fig. D5 Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, Third wave.