IARIW 2022

IARIW 2022

Monday 22 - Friday 26 August

Wealth Survey Calibration: Imposing Consistency with Income Tax Data

Daniel Kolář
(Charles University in Prague, Czech Republic)
daniel.kolar@fsv.cuni.cz

Paper prepared for the 37th IARIW General Conference

August 22-26, 2022

Session 6A-2, Dynamics of the Wealth Distribution around the World II

Time: Friday, August 26, 2022 [11:00-12:30 CEST]

Wealth Survey Calibration: Imposing Consistency with Income Tax Data

Daniel Kolář[*]

July 2022

Abstract

Wealth surveys tend to underestimate wealth concentration at the top due to the "missing rich" problem. We propose a new way of improving the credibility of wealth surveys: We make them consistent with tabulated income tax data. This is possible with the Household Finance and Consumption Survey (HFCS), which takes place in most European countries every three years and collects data on both income and wealth. Consistency is achieved by calibrating survey *weights* using the income part of HFCS. We apply the calibration method of Blanchet, Flores and Morgan (2022), but propose a new way to determine the merging point where the calibration starts. Calibrated weights are then used with HFCS wealth values. We test the method on Austrian data and find that calibration increases the top 1 % wealth share from 26 % to 37 % in 2014 and from 23 % to 27 % in 2017. The effect is small and negative in the 2011 HFCS wave, even though the net worth of the top 1 % increases. We also highlight a strong downward bias in the Austrian HFCS income distribution, which begins even before the 80th percentile. Following the calibration, we 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 — D31, C83

[*] Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague, Czech Republic. E-mail: daniel.kolar@fsv.cuni.cz. I express my gratitude to Petr Janský, Thomas Blanchet, Ignacio Flores, Marc Morgan and Emanuel List for their comments and discussions. Any errors are mine. This work was supported by the 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, richest households may not be adequately sampled in the survey. Moreover, even if they are, they may underreport wealth, leave some questions unanswered or refuse to participate in the survey altogether.

To mitigate the problem, existing literature proposes several types of adjustments to the survey wealth distribution. Vermeulen (2014) estimates a Pareto distribution for the top tail using survey observations combined with the Forbes World's billionaires list. The Pareto distribution may also be estimated without adding the rich list, which is the preferred approach of Eckerstorfer et al. (2016). Subsequent research extends this work by using more detailed national rich lists (Bach et al., 2019; Brzeziński et al., 2020), by scaling up different asset classes to match the National Accounts totals (Vermeulen, 2016) or by determining a non-arbitrary lower bound of the Pareto distribution (Brzeziński et al., 2020; Eckerstorfer et al., 2016). However, as we also show in this paper, the rich list plays a dominant role in the Pareto coefficient estimation and its quality is therefore crucial. On the other hand, estimating the Pareto tail with survey data alone (that is, without the rich list) does not solve the differential non-response problem, unless a detailed oversampling of the wealthy is performed (Vermeulen, 2018). Perhaps as a result, the World Inequality Database prefers the income capitalization method as a starting point for their wealth inequality estimates (Alvaredo et al., 2021).[1]

In this paper, we propose a novel way to improve the credibility of wealth surveys: We make them consistent with another external source–tabulated income tax data. This is possible with the triannual Household Finance and Consumption Survey (HFCS), which will cover 23 European countries in its last wave and which collects data on both income and wealth. Consistency is ensured by applying the calibration method of Blanchet, Flores and Morgan (2022) to the income part of the survey. As we only adjust survey *weights*, we can then use the new weights together with the wealth data to estimate wealth inequality. The flexibility of linear calibration allows us to preserve the main socio-demographic characteristics of the survey and to overcome the issue of HFCS recording income at the individual level but wealth at the household level.

In addition, we construct a new method to determine the optimal percentile from which calibration should start, called the *merging point*. While we share the aim of Blanchet, Flores and Morgan to preserve the continuity of the density function, our approach relies on a (perhaps more intuitive and informative) graphical comparison of survey and tax densities. It may also lead to more than one candidate merging point, giving researchers more flexibility. In a Monte

[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). This method assumes a constant rate of return for each asset class across wealth groups.

Carlo simulation, our method performs comparably or better, depending on the specification. Moreover, it does not rely on the assumption of monotonically decreasing survey to tax density ratio and is not sensitive to how we set the lower bound from which the tax data are considered reliable. On the other hand, our method is more computationally demanding and less suitable when the merging point should be above the 99th percentile. The latter drawback can nevertheless be mitigated.

The advantage of the correction is that it 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, the method does not make any assumptions about the correlation of income and wealth in the population, although this correlation will determine the adjustment's effect: If it is positive, then increasing the weight of top income holders in the calibration will also increase the weight of top wealth holders, leading to larger wealth inequality estimates. Finally, the 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 calibration effect will be minimal. In contrast, we can expect a significant effect in case no oversampling strategy whatsoever has been implemented.

The drawback of our method, typical of any wealth survey adjustment, is that it cannot be empirically tested against data where the true wealth distribution is known. It is also not guaranteed that "fixing" the income distribution will correct the bias in the wealth distribution in its entirety.

We test the method on Austrian data and find that calibration increases the top 1 % wealth share from 26 % to 37 % in 2014 and from 23 % to 27 % in 2017. On the other hand, the effect is small and negative in the first HFCS wave in 2011: Even though calibration increases the net worth of the top 1 %, the denominator (that is, estimated total wealth) increases even more, leading to a decrease in the top 1 % share. Our merging point method highlights a large bias in the Austrian HFCS income distribution which starts as early as before the 80th percentile. This bias is much larger than in the EU Statistics on Income and Living Conditions (EU-SILC) data, which we also show.

After calibration, we combine our wealth distribution adjustment with adjustments in the existing literature: We 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. The impact of calibration here is small because the rich list "dominates" the objective function of the Ordinary Least Squares (OLS) Pareto estimator. We recommend our adjustment especially when the rich list is not considered reliable and when it is desirable to preserve all properties of the survey dataset for subsequent analysis.

The remainder of the paper is organized as follows. Section 2 introduces the data that we use to apply our method: wealth surveys, income tax data and rich lists. We 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 we report the results, including of a Monte Carlo simulation which evaluates our 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 ever–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], 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 thus have to be carefully matched with these concepts to ensure accurate calibration. For wealth inequality measurement, we use the net wealth concept of the HFCS: It is the sum of a household's 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 Blanchet, Flores and Morgan (2022) (hereafter referenced as Blanchet et al., 2022) refer to the ”non-sampling error”. First and foremost, when approached by the interviewer, wealthier households are more likely to refuse to participate (Kennickell and Woodburn, 1997). 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.[2] Second, the rich may leave unanswered 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., 2022), 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

[2] For example, the Czech Statistical Office rewards participating households with commemorative coins with a total face value of 120 Czech crowns, i.e., approximately 5 euro.

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. Some countries may not oversample at all. 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). Consequently, it should be noted that oversampling reduces the bias of wealth inequality estimates not so much because the richest are sampled more (this mainly reduces the variance), but because the rich strata are more accurately defined and the adjustment of survey weights for non-response is more specific. In addition, while oversampling can improve wealth inequality estimation, the fact that strategies differ across countries hinders comparability.

Our 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 we 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 the year 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 HFCS income data, we compare it with the EU Statistics on Income and Living Conditions (EU-SILC) data for reference years 2009, 2013 and 2016. The data are provided by Eurostat, which is not responsible for any conclusions drawn from the data.

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 incorrect due to tax evasion, tax exemptions or withholding of some 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 published by most EU countries (Blanchet, Chancel and Gethin, 2021), typically in 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, Fournier and Piketty, 2017).

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.[3] Based on the documentation, we 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,

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

we 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, Chancel and Gethin (2021), we estimate and deduct social contributions from survey data where appropriate, based on rates recorded by the OECD.[4]

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. We 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.

3 Methodology

In the first part of our adjustment, we adjust survey weights to impose consistency of the survey's income distribution with the tax data. We follow the reweighting approach of Blanchet et al. (2022), but propose a new method to determine the merging point–the percentile from which consistency is imposed. Second, the new weights are 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 Blanchet et al. (2022), consists of several steps. First, the tax data (which represent the "true" distribution) 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 matched to a bracket sum up to the tax population size of that bracket. Blanchet et al. 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

[4] For wages, we use rates recorded in the annual OECD Taxing Wages publication, for pensions we use the OECD Pensions at a Glance publication and for self-employment income the OECD Tax Statistics. In addition, we set the limit for maximum social security contributions from self-employment (17793 euro in 2016) as the limit for the total contributions of an individual.

tax distribution, we only calibrate brackets above the predefined *merging point*. The choice of merging point is discussed in Section 3.1.2.

After matching survey observations to fractile-based brackets, we 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. Blanchet et al. (2022) choose $x = y = 5$, which we 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:

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

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, we can impose other conditions on the calibrated weights by expanding the vectors $\boldsymbol{x}_i \forall i$ and $\boldsymbol{t}$. Like Blanchet et al. (2022), we 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 (Blanchet et al., 2022, Equation 5). In addition, no new weight should be lower than 1, which we enforce with a simple algorithm.[5]

In the context of our application, one specific problem arises: While income is recorded at the individual level in surveys as well as 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,[6] 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 we 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

[5] If the linear calibration leads to some weights lower than one, we 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, we correct weights from the first iteration directly by setting them to one and adjusting (i.e., slightly decreasing) all other weights to keep the population total constant. Blanchet et al. (2022) instead use an iterative method described in Singh and Mohl (1996).

[6] If, for example, a surveyed household of three members represents 200 three-member households in the population, then each of its members also represents 200 individuals.

consistent. Instead, we 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 we begin to merge the survey and tax data. It is, in principle, possible to start the calibration as soon as the tax data are available. However, as Blanchet et al. (2022) argue, this would unnecessarily distort the survey, which should be corrected only once the bias at the top starts. Blanchet et al. 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.

We propose a new way of choosing the merging point while also aiming to preserve the continuity of the calibrated density function. Our starting point is a visual comparison of tax and survey frequencies at different percentile-based brackets, illustrated in Figure 1 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, we work with an adaptive kernel estimator of the survey density (Cowell and Flachaire, 2015). We integrate it over each bracket's income interval and multiply the integral by the survey population, obtaining a smoother frequency estimate for each bracket.

We search for the *best* rather than *optimal* merging point: We test each percentile and observe whether the new distribution can be considered continuous. Our test statistic is the absolute distance between the frequency where we begin to merge (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 total population remains the same. This adjustment is assumed to be uniform for all brackets below the merging point.

Which percentile should be the merging point? A straightforward choice 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 is assumed by Blanchet et al., 2022) 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.

We 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 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

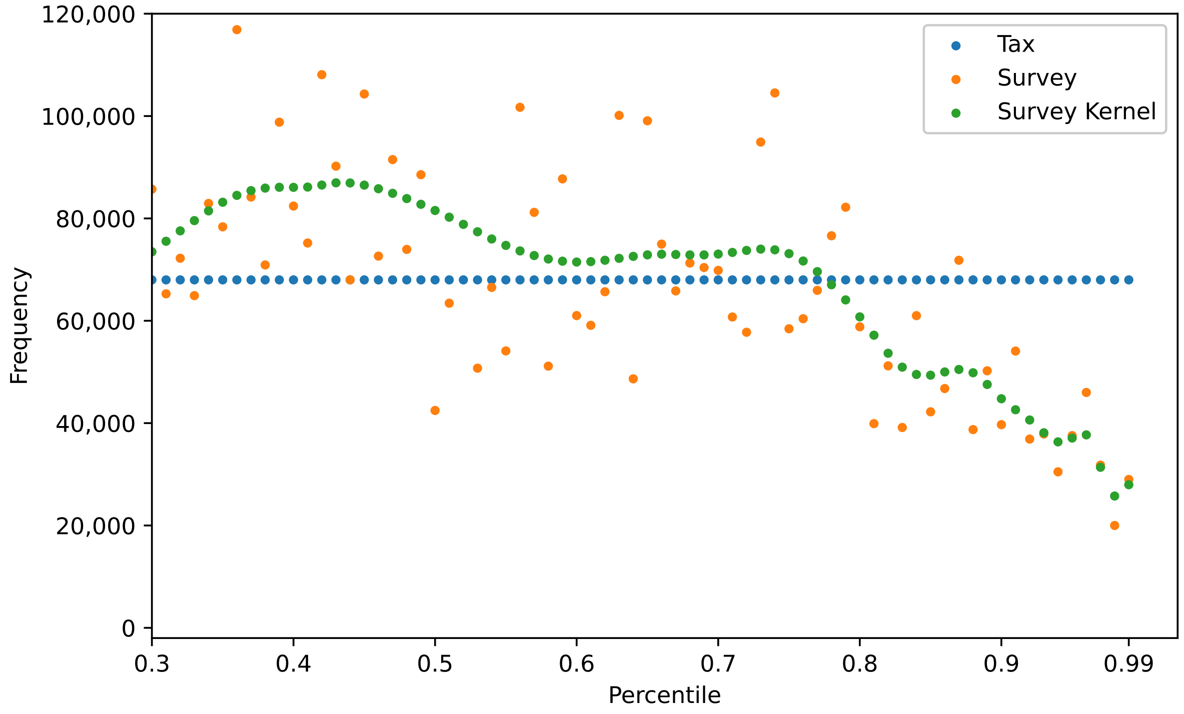


Figure 1. Tax and survey frequencies.

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.[7] Our algorithm should always be accompanied by a visual inspection of tax and survey frequencies, which is informative on its own.

We illustrate the merging point choice in Figure 2, using the same data as in Figure 1. Setting the trustable span to start at the 30th percentile, our 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 our general recommendation, which we also apply, is to choose the highest candidate merging point–in our 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.

How does our method compare to the optimal merging point algorithm of Blanchet et al.? We argue that our method should be considered due to its intuitiveness and simplicity. It avoids the assumption of 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 Blanchet et al. allows for the merging point to be above the 99th percentile. If the frequency comparison

[7] 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, Blanchet et al. (2022) extrapolate the ratio of survey to tax density for percentiles not covered.

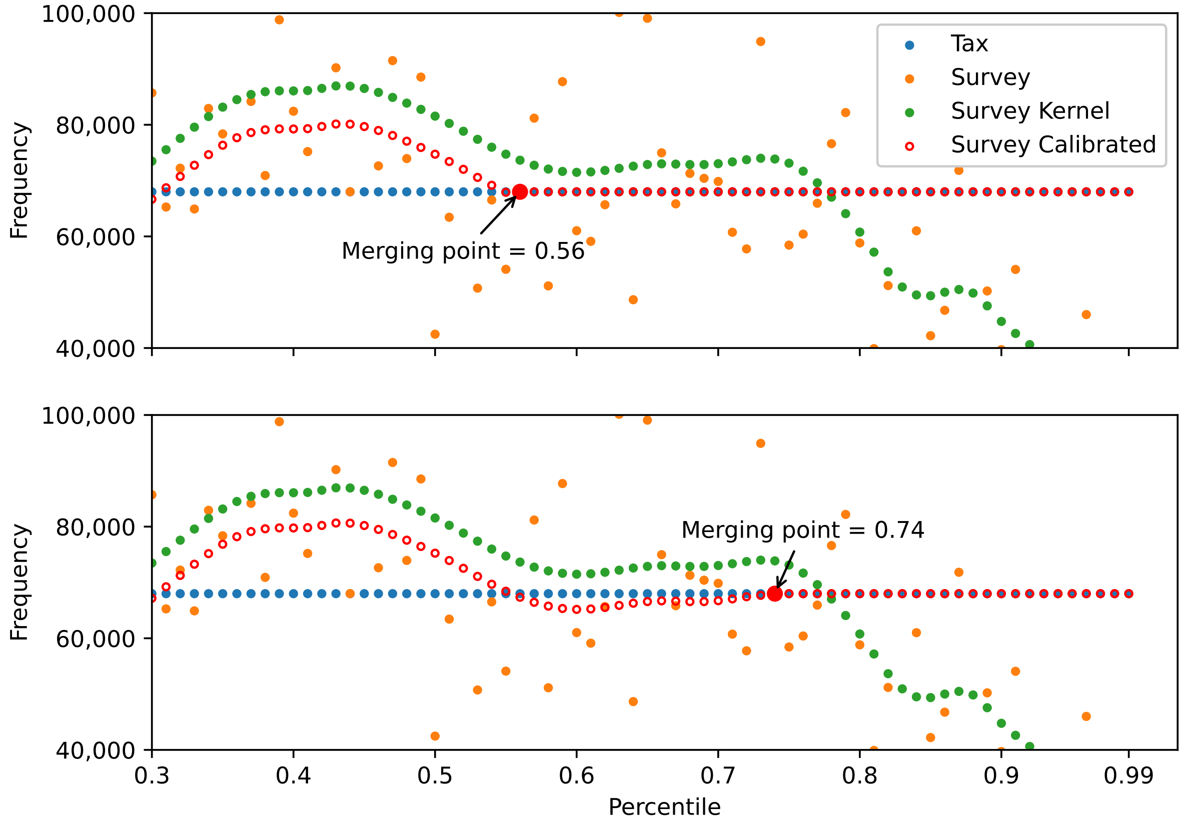


Figure 2. Choosing the optimal merging point. 0.56 and 0.74 are two candidate merging points, each leading to a visually continuous density after calibration. We choose 0.74 as the optimal one as it is the highest. Same dataset as in Figure 1.

suggests this could be the case, our method can be easily extended to more granular frequency brackets (e.g., 1/2 or 1/4 of a percent). We illustrate this extension in Section 3.1.2.

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 our method in Figures 1 and 2: Our method leads to a merging point at the 74$^{\text{th}}$ percentile, just like the method of Blanchet et al. An additional comparison based on simulated datasets is provided in Section 4.1.

Extension 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. We illustrate such extension using Austrian EU-SILC data from the 2016 wave. Figure 3 introduces the dataset. The survey kernel density follows the tax density quite closely until around the 90$^{\text{th}}$ 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 97$^{\text{th}}$-98$^{\text{th}}$ percentile, our 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.

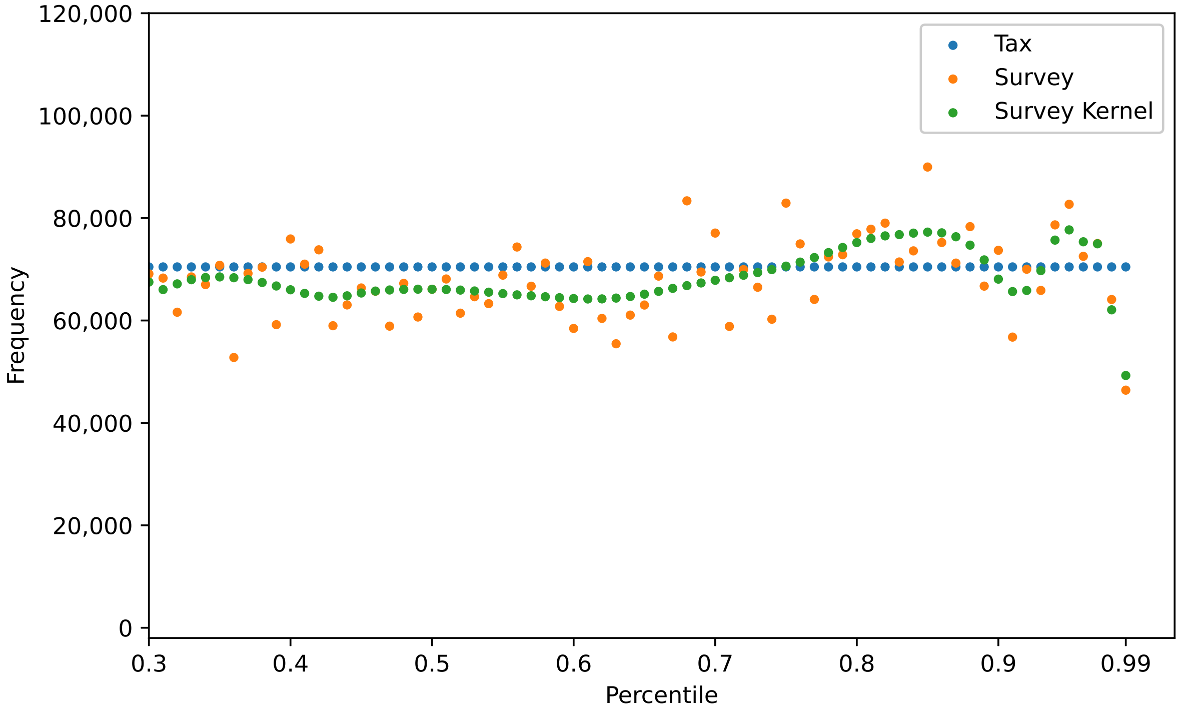


Figure 3. Tax and survey frequencies. Gaps in the survey kernel distribution suggest more granular frequency brackets are desirable.

Figure 4 shows how to make the comparison of survey and tax distributions more informative and suitable for our optimal merging point algorithm. We 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 our 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". Our 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.

We applied the approach of Blanchet et al. (2022) to this dataset and found that their optimal merging point is sensitive to the choice of the trustable span, i.e., the left-bounded interval on which the tax data are considered reliable.[8] Our 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 over the merging point algorithm of Blanchet et al.

[8] 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.

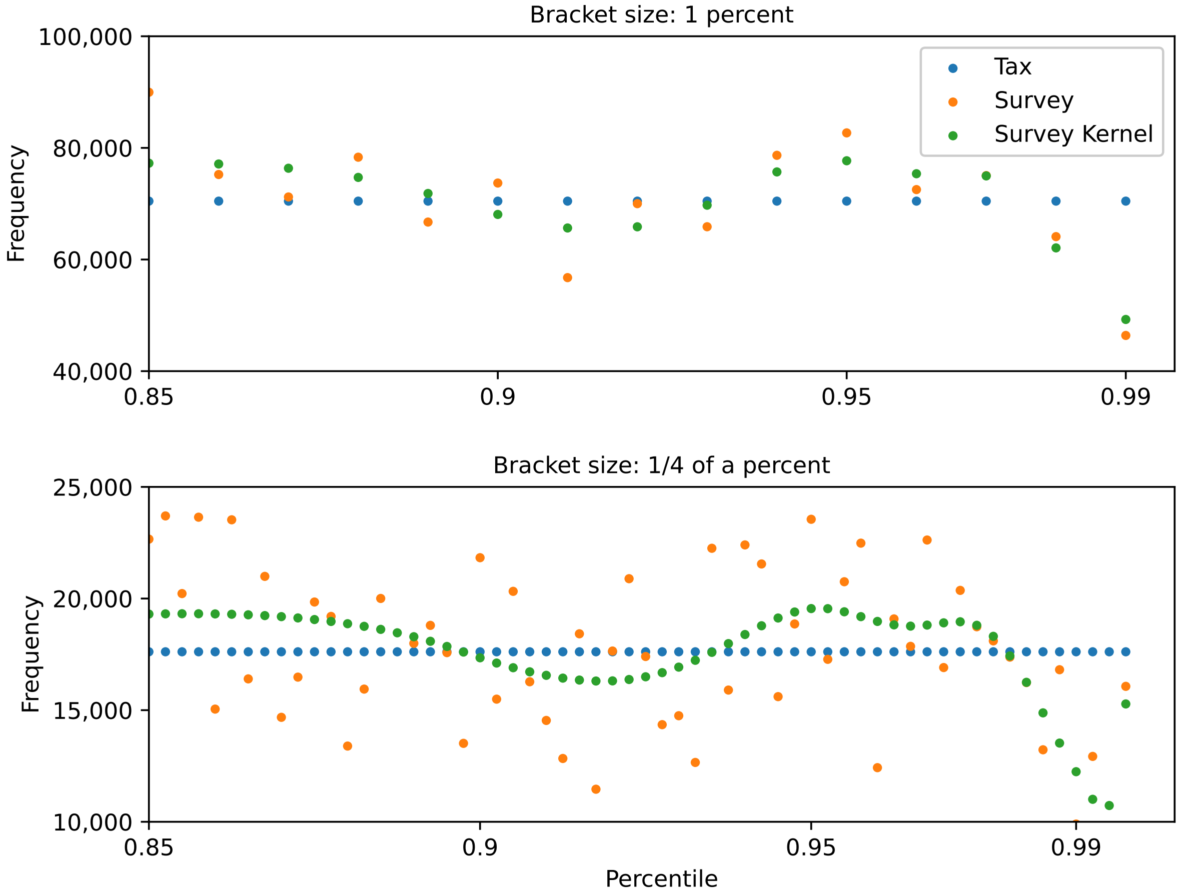


Figure 4. Comparing tax and survey distributions using more granular frequency brackets. Y-axis scale in the bottom graph is 4 times lower than in the top graph. Same dataset as in Figure 3. Small random noise added to the survey distribution in the bottom graph to ensure confidentiality.

3.1.3 HFCS-specific adjustments

As the HFCS data are provided in five dataset replicates due to the imputation of missing values, our method must be adjusted to take this into account. Rather than estimating a potentially different optimal merging point for each replicate, we estimate one merging point based on the average test statistic of each percentile.[9] 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. We 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 we study the inequality of wealth, which is reported at household level. However, if one was interested in income inequality estimates of adult individuals, ”young rich”

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

people should be removed from the adult population after calibration, as they would be if one had perfect tax microdata.

3.2 Adjusting the top of the wealth distribution

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 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, we refer the reader to 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

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

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.[10] Note that if all weights are equal to 1, then $n = N$ and we have a standard maximum likelihood estimator for a Pareto distribution. The extension thus means that observations that represent more households have a larger impact on the estimate. In other words, it is as if we

[10] Vermeulen (2018) calls this the pseudo-maximum likelihood estimator. Our formula differs slightly from his in that we use w_{min} in the denominator instead of $\min\limits_{1\dots n} w_i$. We consider our version to more accurately represent the stated assumptions but this choice has, algebraically, no impact on the α estimate.

replaced respondent i of weight N_i by N_i respondents of weight 1, all having the same wealth w_i.

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 our 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 we do as well. A similar correction is applied by Vermeulen (2018), but the difference is that our setting allows us to estimate α without the intercept. This improves the fit of the Pareto line to the data because the intercept does not enter the Pareto distribution. We 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 lower variance, especially in the presence of oversampling. When estimating the Pareto distribution from survey data only, we therefore prefer $\hat{\alpha}_{MLE}$, as in, e.g., Eckerstorfer et al. (2016). On the other hand, the weighted maximum likelihood estimator is unsuitable when survey data are combined with rich lists, where each entry has a weight of only 1. We explain this in Section 3.2.2. Pareto coefficients based on survey data combined with a rich list are therefore estimated 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 we risk dealing with observations that are not high enough to be well approximated by a Pareto distribution. By setting w_{min} too high we decrease the sample size and thus increase the 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).[11] 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.

We considered applying the KS statistic to determine w_{min} for our benchmark results but found that it may fail to identify deviations from a Pareto tail at the very top of the distribution. We illustrate this in Figure 5 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 we illustrate in Figure 5 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. We 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.

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 5 is only illustrative; the included values relate to the difference of ccdfs of the entire population.

We experimented with adjusting the setup of the KS test but it did not produce reliable results. First, we tried comparing the logarithms of ccdfs in the KS test, which corresponds more closely the graphical comparison in Figure 5. Second, we 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-

[11] 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).

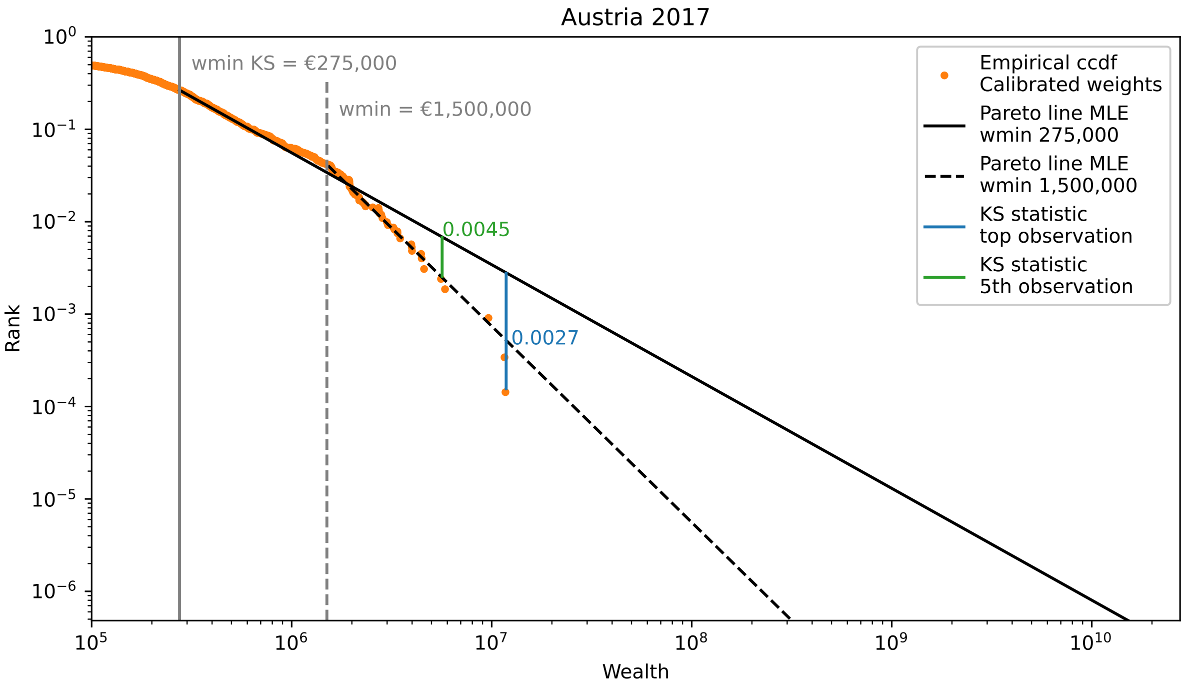


Figure 5. Estimating the Pareto distribution at different thresholds. 275,000 euro is the optimal threshold according to the Kolgomorov-Smirnov test. The blue and green lines illustrate how the KS statistic is evaluated for two observations at the top. Because of the logarithmic scale, visual deviations from the Pareto line at the top are not accounted for sufficiently in the KS test. Data: HFCS, Austria 2011, First implicate. Survey weights are calibrated based on income tax data.

wealth graph, we resort to setting w_{min} based on a visual examination of the distribution. As a result, our benchmark estimates for all three years are based on a threshold of 1.5 million euro and we 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. Since we observe a break in van der Wijk's law at around this threshold in 2014, we choose to proceed with the benchmark threshold of 1.5 million euro nevertheless.

Finally, it should be noted that judging the Pareto tail based on its fit to the data may be problematic in principle. If one believes that our 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. As we 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 5, 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 we estimate the Pareto tail based on data which we first adjusted using income calibration, we consider our 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.[12] $\hat{\alpha}_{MLE}$ 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). We illustrate this point in Figure 6 on the same Austria 2011 dataset as in Figure 5: The Pareto line estimated at the 1.5 million euro threshold is now aligned with Pareto the line estimated at $w_{min} = 275,000$, even though survey observations above 1.5 million euro are not. Another demonstration is provided in the Appendix, where we extend a Monte Carlo simulation of Vermeulen (2018).

As a consequence, 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 is no bias in the survey data.

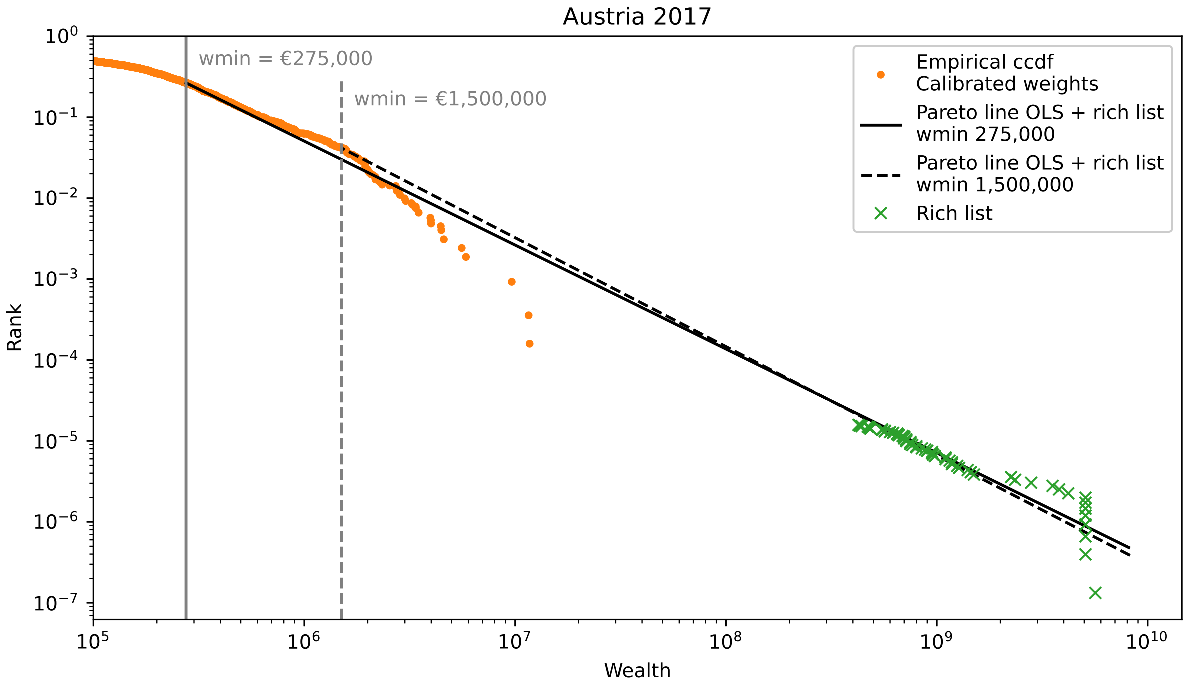


Figure 6. Estimating the Pareto distribution from survey data combined with a rich list. Once the rich list is included in the estimation, it becomes the main determinant of the slope of the Pareto line. Data: HFCS, Austria 2011, First implicate. Survey weights are calibrated based on income tax data.

[12] To preserve the original population size, we 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. First, 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. We believe that this approach is used by Vermeulen (2018) because we obtain identical results for Austrian first wave data which are also analyzed in his work.

A second option is to 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), and we apply it as well. 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, 2018). We 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. We also 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 we ignore 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 it 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 our context is that these weights are not suitable for variance estimation if we 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. We 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. We create 500 new bootstrap weights for each set of weights. For the non-calibrated weights, we can compare the standard errors of top 1 % share estimates based on the provided weights with those based on our 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, we apply the formula in European Central Bank (2020, Section 7.3).

4 Results

4.1 Simulation: Different merging point approaches

In Section 3.1.2 we developed a new method to determine the merging point, i.e., the percentile where the calibration of survey and income tax data starts. We presented a different framework than Blanchet et al. (2022) 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 we present a more systematic comparison using simulated Monte Carlo data.

The setup of the Monte Carlo simulation largely, but not entirely, follows Blanchet et al. (2022). A population of 1 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 90th percentile and then decreases linearly with rank until nearly reaching 0 %. The probability of misreporting is 20 % until the 95th 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 income distribution is recorded in the income tax data and available in an aggregated tabulated form.

In the tax data quality lies our main deviation from the setup of Blanchet et al.: The tabulated tax data in our setup are accurate for the entire distribution, while they assume a downward bias up until the 90th percentile. The second difference is the population size, which is 1 million in our setup and 9 million in the setup of Blanchet et al. As a consequence, our sample size in each iteration is approximately 9 times lower. We decreased the population size due to our method's computational demands, which result from the need to integrate the kernel density over each percentile-based bracket. We discuss the sensitivity of results to these changes later in this Section.

We perform 1,000 iterations of the setup and apply both merging point approaches to the simulated data. Our 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 we describe in Section 3.1.2 and apply in the empirical part: We determine candidate merging points, i.e., percentiles for which the test statistic is considered low, and, if there are more than one, we choose the highest candidate merging point. In the second variant, denoted $Kdirect$, we disregard candidate merging points and proceed directly with the percentile with the lowest test statistic.

Figure 7 and Table 1 report the results. Had the non-response profile of the population been publicly known, the hypothetical researcher should set the merging point at the 90th percentile.

Such merging point would correct the bias at the top in its entirety while minimizing the survey distortion below this point. Our benchmark method K generally identifies the merging point closer to the 90th percentile. In 45.1 % of simulations, the K optimal merging point was between percentiles 0.89 and 0.91, while the estimator of Blanchet et al. (2022), denoted *BFM*, was within this range in 13.5 % of simulations. On the other hand, our method located the merging point above percentile 0.91 more often, in which case it did not correct the entire bias. Nonetheless, the overestimation was not dramatic: Our method identified the merging point at the 92nd percentile in only 64 out of the 1000 iterations and never above it.

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 90th 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.

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	59.6	26.4	13.5	0.5	100
K	6.0	42.5	45.1	6.4	100
K direct	86.7	10.1	3.2	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.

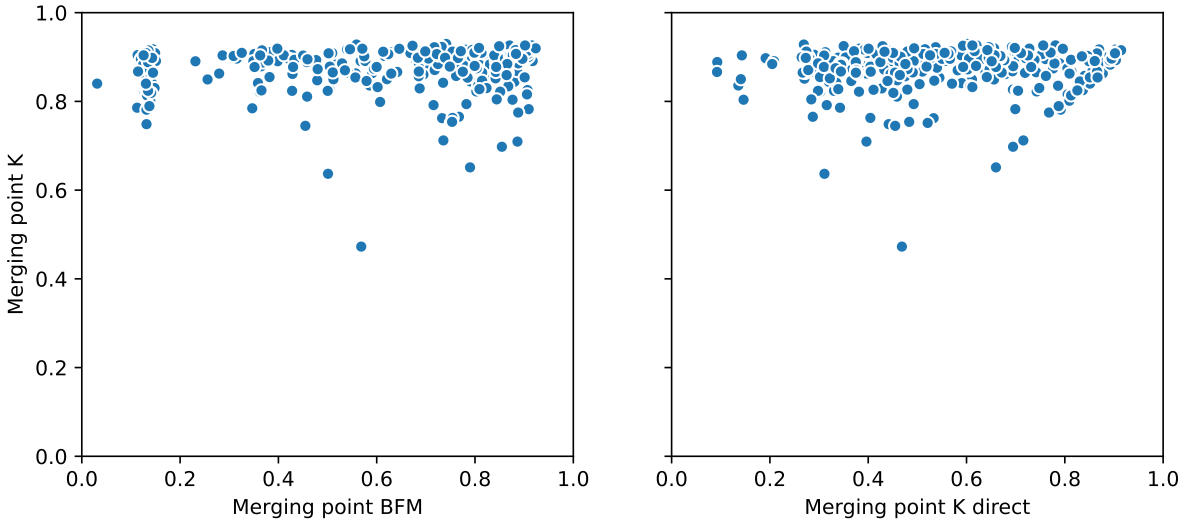


Figure 7. The correlation of optimal merging point approaches. For visualization purposes, only 300 simulations are shown and small random noise is added to prevent overlapping.

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 (0.965) 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 90th percentile. These assumptions may not hold in real life and we still consider it desirable to aim for as low a survey distortion as possible, especially in the presence of covariates. Finally, it was already established by Blanchet et al. that the estimates themselves are significantly closer to the true value and with less variance than the original survey estimates.

In our simulation, we have assumed that the tabulated tax data accurately represent the entire distribution. Blanchet et al. instead assume that they are downward biased up until the top 10 %. Under this assumption, the *K* and *BFM* methods perform comparably. However, we consider this assumption quite extreme because it implies that the unbiasedness of survey and tax data overlaps at precisely one point, the 90th percentile. When the survey and tax data are unbiased over a larger interval, say from the 70th to the 90th percentile, our method will again tend to estimate the merging point closer to the optimal value, which is the 90th percentile. As per the population size, increasing it to 9 million as in Blanchet et al. improves the performance of the *BFM* method, but not to the extent that would match the *K* method in our baseline setup. Moreover, the *K* method also tends to perform better with increased population (and therefore survey sample) size. Detailed results of Monte Carlo simulations under these alternative assumptions are provided in the Appendix.

4.2 Empirical data: Comparing tax and survey 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 8, where we also include the income distribution 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) until around the 97th-98th percentile. In contrast, the downward bias in HFCS is observable as early as before the 80th percentile.

Austria is a country with no recorded oversampling strategy in the second and third waves and with only basic oversampling of Vienna in the first wave, which may explain the poor performance. Another important difference is the sample size. The HFCS distribution is based on 4,151, 5,078 and 5,225 adult individuals in the first, second and third waves respectively (for the first implicate). The corresponding sample sizes in the Austrian EU-SILC data are more than twice as large, between 10,500 and 11,000.[13] EU-SILC also takes labor, pension and unemployment income data from public registers (Heuberger et al., 2013), which eliminates bias arising from the untruthful reporting of these variables.

The comparison of survey and tax densities, which is part of our 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

[13] As explained in Section 3.1.3, the sample size referred to in this paragraph consists of individuals aged 20 or over, plus younger individuals with income over the threshold of 10,000 euro. We also note that reference years for income and wealth differ for each HFCS wave.

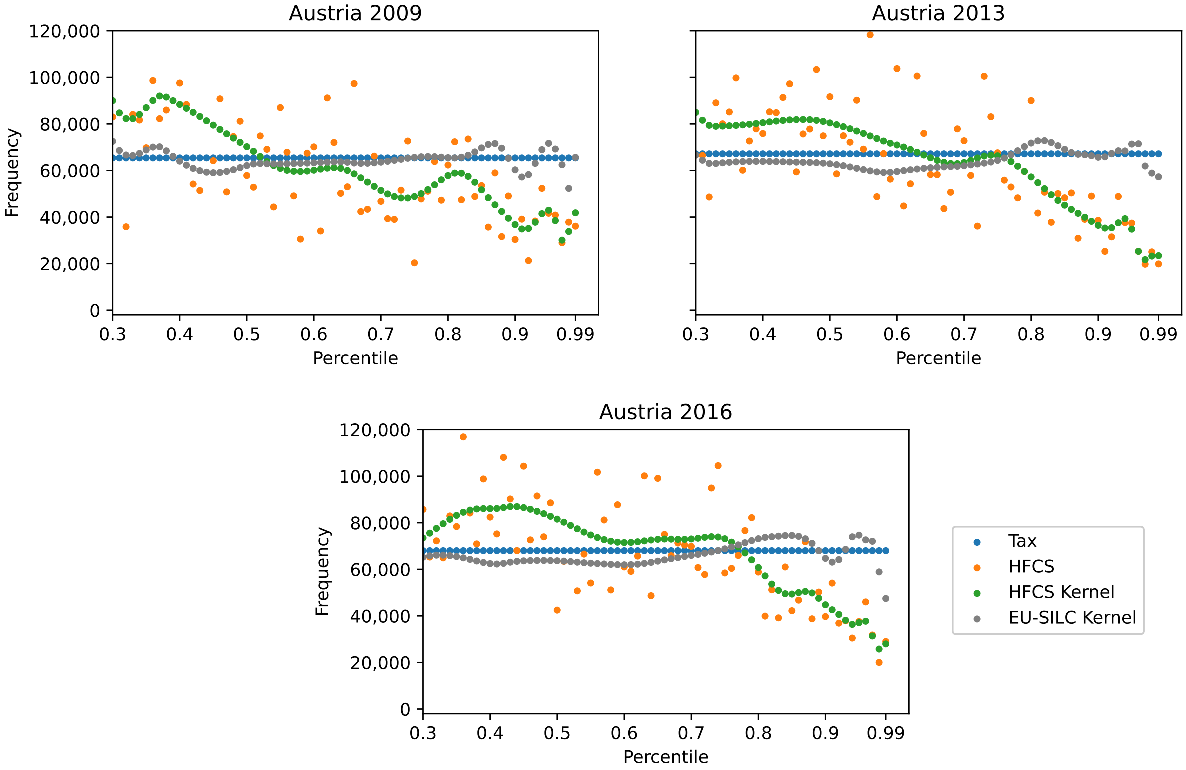


Figure 8. A comparison of survey (EU-SILC, HFCS) and tax distributions. Percentile-based brackets computed using the HFCS adult population size. The first implicate of HFCS is used.

(which we cannot check externally), then the income distribution should be as well. Figure 8 strongly suggests that this is not the case with Austrian HFCS data.

4.3 Wealth inequality with calibrated weights

How does the calibration of weights based on income tax data affect wealth distribution? As Figure 9 shows, the impact varies depending on the studied dataset. In the Austrian 2011 data, the impact is small. Correlation between top income holders in the survey (whose weights have been increased) and top wealth holders appears to be low in 2011. In contrast, income calibration visibly increases survey weights of wealthy families in 2014 and 2017. This suggests that the calibration based on income tax data can improve the credibility of wealth surveys and mitigate the non-response problem, although this claim cannot be made with certainty in the absence of an external check.

The impact of calibration on the Pareto tail estimation is highlighted in Figure 10 using Austrian 2017 data. A notably more horizontal Pareto line estimated using calibrated weights implies larger inequality. Intuitively, for each rank at the top (for example, 10^{-3}, which corresponds to a household in the top 0.1 %), estimated wealth is larger. 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 top tail that is being replaced is larger. When the list of wealthiest individuals is included, the difference in the Pareto tail slope becomes negligible.

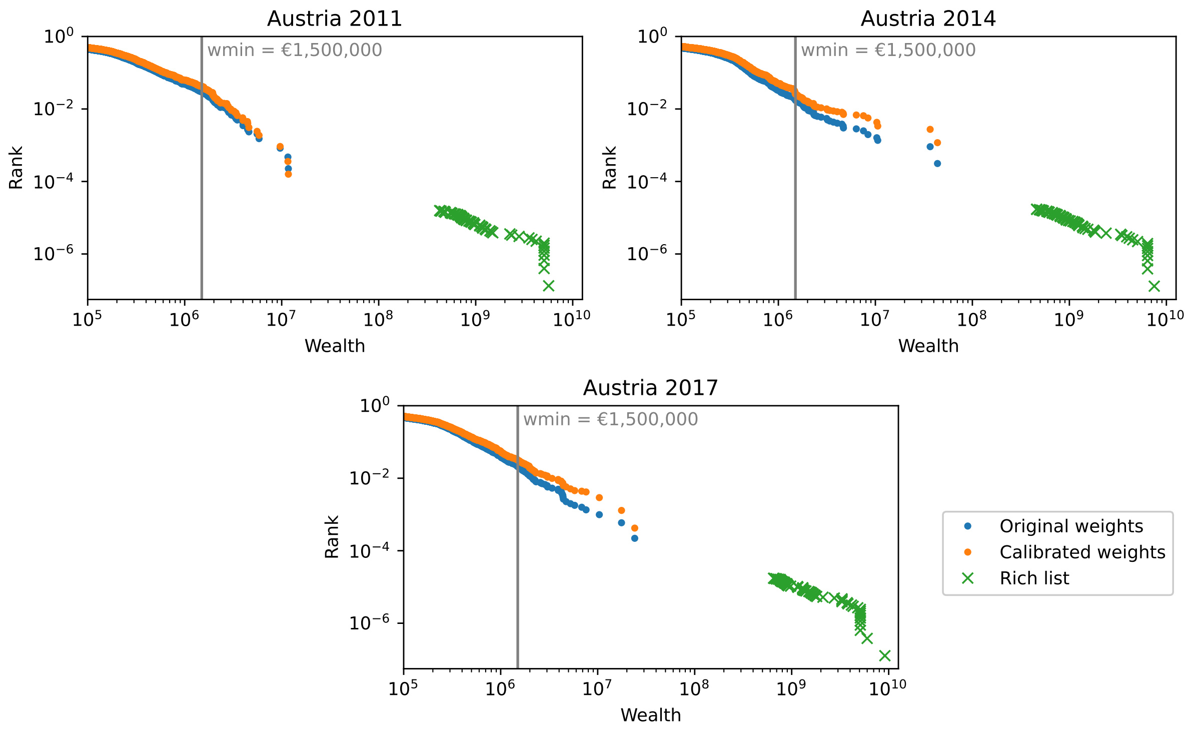


Figure 9. Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, first implicate. Figures for all five implicates are provided in the Appendix.

This is because the rich list tends to "dominate" 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, all with the benchmark w_{min} threshold of 1.5 million euro. Let us first consider the survey data alone, without any distributional adjustment. Survey weight calibration based on income tax data 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 the impact of calibration is small and the top 1 % wealth share even decreases, from 23.2 % using original weights to 21.5 % following calibration. This implies weak a correlation between top income and top wealth holders in the 2011 sample, perhaps partly attributable to a smaller sample in 2011 (2,380 households) than in 2014 and 2017 (2,997 and 3,072 households, respectively). At the same time, the impact of calibration is not that the top 1 % hold less wealth but that the denominator, estimated total wealth, increases more than the wealth of the top 1 %.

The impact of fitting a Pareto tail estimated using survey data alone, using our preferred $\hat{\alpha}_{MLE}$ estimator, 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

[14] For completeness, we list reasons why our 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, we create synthetic households for the top tail rather than working with

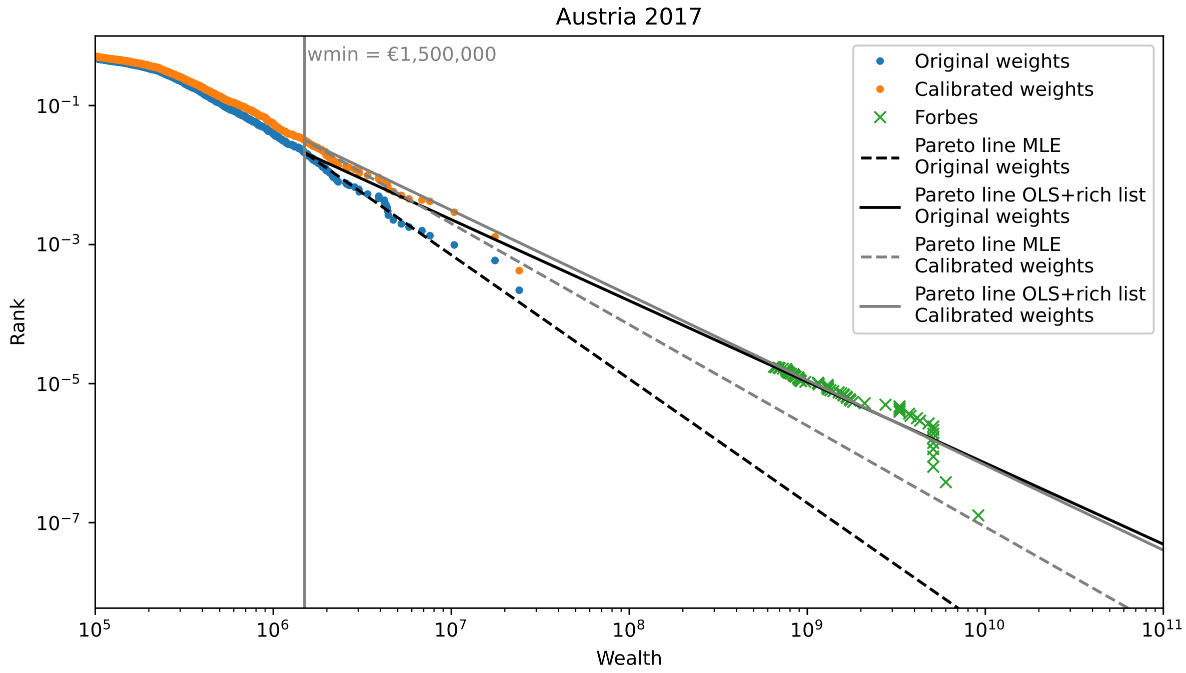


Figure 10. Pareto distribution estimation. Data: HFCS, Austria 2017, first implicate.

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. Survey weight calibration can therefore have a similar effect as adding the rich list, but with the advantage of relying on high-quality administrative tax data rather than a magazine ranking.

Table 2. Top 1 % share estimates.

	Survey	MLE	OLS + rich list
2011, original weights	23.2	31.3	41.3
	(7.3)	(18.0)	(1.3)
2011, calibrated weights	21.5	28.6	39.6
	(7.2)	(17.2)	(0.9)
2014, original weights	25.5	23.1	39.9
	(8.1)	(11.3)	(1.3)
2014, calibrated weights	36.6	44.8	38.4
	(15.0)	(28.2)	(1.3)
2017, original weights	22.8	20.3	44.0
	(5.8)	(4.7)	(0.9)
2017, calibrated weights	27.3	25.5	43.0
	(6.8)	(10.8)	(0.8)

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

(conditional) expected values–see Section 3.2.3. Second, $\hat{\alpha}_{OLS}$ is estimated using a regression *without* intercept. Third, due to calibration we must work with our own bootstrap weights rather than those provided by HFCS–see Section 3.2.4.

In Table 3 we present estimates for the top 5 % wealth share. The impact of survey calibration is again positive in 2014 and 2017 and negligible in 2011. In 2017, the 5 % wealthiest households owned 43.2 % of total wealth according to original survey data and 47.4 % when the survey weights were calibrated. The impact of Pareto fitting does not change either since it is the same Pareto tail that is being fitted. The top 5 % share is a more robust measure and the differences between methods are generally lower, as are the standard errors. In the Appendix we provide the top 1 % share estimates for thresholds 1 million euro and 2 million euro. The sensitivity of MLE estimates to threshold choice is lowest in the third, most recent wave.

Table 3. Top 5 % share estimates.

	Survey	MLE	OLS + rich list
2011, original weights	47.7	52.7	60.3
	(7.7)	(14.0)	(2.7)
2011, calibrated weights	46.7	50.9	59.5
	(8.2)	(14.1)	(2.6)
2014, original weights	43.5	41.8	54.6
	(6.3)	(8.8)	(1.7)
2014, calibrated weights	52.5	59.0	54.2
	(12.5)	(21.9)	(2.2)
2017, original weights	43.2	41.4	59.3
	(4.5)	(3.8)	(1.3)
2017, calibrated weights	47.4	46.6	60.0
	(5.7)	(8.7)	(1.6)

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

Furthermore, estimates based on calibrated weights tend to exhibit higher standard error 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 positive that standard errors tend to decrease with each HFCS wave, suggesting that imputed missing values are becoming less of a concern. In the Appendix, we provide the log-rank log-wealth graphs for all five implicates (the only difference between each implicate are 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. We propose and apply a new approach that makes wealth surveys consistent with income tax data, a high-quality external source. First, we carefully match income concepts in HFCS to those in the tax data. Then we apply the calibration method of Blanchet et al. (2022) using our own algorithm to determine the optimal merging point. Using the calibrated weights with HFCS wealth data, we find that calibration can, depending on the dataset, have a similar effect as replacing the top with a Pareto distribution and a rich list.

We believe that our method should be considered by researchers working with surveys to study the top of the wealth distribution (e.g., Garbinti et al., 2021; Palomino et al., 2021). In our optimal merging point algorithm we 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., 2020). Our paper reveals a strong bias in the Austrian HFCS income data when compared with the tax distribution and even with another survey, EU-SILC. A key prerequisite for the successful application of our method is the availability of reliable income tax data and the correct matching of this data to income concepts in the survey.

References

Alstadsæter, A., Johannesen, N. and Zucman, G. (2019). 'Tax Evasion and Inequality'. *American Economic Review*, 109(6).

Alvaredo, F., Atkinson, A. B., Bauluz, L., Fisher-Post, M., Blanchet, T., Chancel, L., Flores, I., Morgan, M., Garbinti, B., Goupille-Lebret, J., Martínez-Toledano, C., Neef, T., Piketty, T., Robilliard, A.-S., Saez, E., Yang, L. and Zucman, G. (2021). 'Distributional National Accounts Guidelines: Methods and Concepts Used in the World Inequality Database'. *WID.world.*

Bach, S., Thiemann, A. and Zucco, A. (1st Dec. 2019). 'Looking for the missing rich: tracing the top tail of the wealth distribution'. *International Tax and Public Finance*, 26(6).

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. and Gethin, A. (2021). 'Why Is Europe More Equal than the United States?' *American Economic Journal: Applied Economics* (Forthcoming).

Blanchet, T., Flores, I. and Morgan, M. (2022). 'The weight of the rich: improving surveys using tax data'. *The Journal of Economic Inequality* (Forthcoming).

Blanchet, T., Fournier, J. and Piketty, T. (2017). 'Generalized Pareto Curves : Theory and Applications'. *Working Papers, HAL.*

Brzeziński, M., Sałach, K. and 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).

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (4th Nov. 2009). 'Power-Law Distributions in Empirical Data'. *SIAM Review*, 51(4).

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

— (2011b). *Measuring Inequality.* Oxford University Press.

Cowell, F. and Flachaire, E. (2015). 'Statistical Methods for Distributional Analysis'. In: *Handbook of Income Distribution.* Ed. by A. B. Atkinson and F. Bourguignon. Elsevier.

Dalitz, C. (2018). 'Estimating Wealth Distribution: Top Tail and Inequality'. *arXiv:1807.03592 [stat].*

Eckerstorfer, P., Halak, J., Kapeller, J., Schütz, B., Springholz, F. and Wildauer, R. (2016). 'Correcting for the Missing Rich: An Application to Wealth Survey Data'. *Review of Income and Wealth*, 62(4).

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

Gabaix, X. and Ibragimov, R. (2011). 'Rank — 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents'. *Journal of Business & Economic Statistics*, 29(1). Publisher: American Statistical Association.

Garbinti, B., Goupille-Lebret, J. and Piketty, T. (1st Feb. 2021). 'Accounting for Wealth-Inequality Dynamics: Methods, Estimates, and Simulations for France'. *Journal of the European Economic Association*, 19(1).

Heuberger, R., Glaser, T. and Kafka, E. (2013). '10. The use of register data in the Austrian SILC survey'. In: *The use of registers in the context of EU–SILC: challenges and opportunities.*

Jestl, S. and List, E. (2020). 'Distributional National Accounts (DINA) for Austria, 2004-2016'. *wiiw Working Paper*, 175.

Kennickell, A. B., Lindner, P. and Schürz, M. (2022). 'A new instrument to measure wealth inequality: distributional wealth accounts'. *Monetary Policy & the Economy* (Q4/21).

Kennickell, A. B. and Woodburn, R. L. (1997). 'Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth'. *Federal Reserve Board Survey of Consumer Finances Working Papers.*

Lustig, N. (2019). 'The "Missing Rich" in Household Surveys: Causes and Correction Approaches'. *Commitment to Equity (CEQ) Working Paper Series*, 75.

Palomino, J. C., Marrero, G. A., Nolan, B. and Rodríguez, J. G. (2021). 'Wealth inequality, intergenerational transfers, and family background'. *Oxford Economic Papers.*

Singh, A. C. and Mohl, C. A. (1996). 'Understanding calibration estimators in survey sampling'. *Statistics Canada.*

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

Vermeulen, P. (2014). 'How fat is the top tail of the wealth distribution?' *European Central Bank Working Paper Series*, 1692.

— (2016). 'Estimating the Top Tail of the Wealth Distribution'. *The American Economic Review*, 106(5).

Vermeulen, P. (2018). ‘How Fat is the Top Tail of the Wealth Distribution?’ *Review of Income and Wealth*, 64(2).

Wald, A. (1945). ‘Statistical Decision Functions Which Minimize the Maximum Risk’. *Annals of Mathematics*, 46(2).

Wildauer, R. and Kapeller, J. (2019). ‘A comment on fitting Pareto tails to complex survey data’. *ICAE Working Paper Series*, 102.

Yonzan, N., Milanovic, B., Morelli, S. and Gornick, J. (2020). ‘Drawing a Line: Comparing the Estimation of Top Incomes Between Tax Data and Household Survey Data’. *Stone Center Working Paper Series.*

Appendix: Anchoring of the Pareto tail to the rich list: A simulation

With a simple extension of one of the Monte Carlo simulations performed by Vermeulen (2018), we demonstrate the extent to which the rich list determines the estimated Pareto coefficient and therefore the top tail wealth.

The non-extended setup of the Monte Carlo simulation is as follows: A population of 1 million households is drawn from a Pareto distribution with *wmin* equal to 1 million and a Pareto coefficient of 1.5. From that population, 750 households are sampled at random and their probability of response declines with wealth according to the formula $P(response) = 0.9028329 - 036594 * \ln w$. This relationship is estimated by Vermeulen (2018) based on non-response rates observed in the US Survey of Consumer Finances in 1992 (Kennickell and Woodburn, 1997). Those who respond are assigned equal weights that sum up to the total population of 1 million (the assumption is that the w_{min} threshold, as well as the size of the population above that threshold, is known to the statistical office). In addition, the wealth of all households richer than 740 million euro is publicly known (this corresponds to the rich list, which is assumed to be without error). The Pareto coefficient is estimated using different specifications, namely with and without the rich list. The simulation is repeated 10000 times.

Our extension is that we work with only *one* simulated population from which we sample in all 10000 iterations, meaning that the rich list remains the same. In contrast, the original specification simulates a new population for each iteration. Comparing these two specifications allows us to in effect separate the standard error of the Pareto estimate due to survey sampling and due to variation in the rich list.

Table 4

Specification	OLS	OLS & rich list
Original	1.61	1.50
	(0.13)	(0.03)
Extended	1.61	1.52
	(0.13)	(0.006)

Pareto coefficient estimates based on two Monte Carlo specifications. The true coefficient is equal to 1.5. Standard errors in parentheses.

The results, reported in Table 4, confirm that the rich list is the main determinant of the Pareto tail and thus of the wealth inequality estimates. In other words, survey observations that also enter the estimation process are largely unimportant. The first row replicates the Monte Carlo simulation in Vermeulen (2018, Table 6) where a new population of 1 million is drawn in each iteration of the simulation. The second row reports our extension where we sample from the same population in each iteration. Including the rich list nearly eliminates the bias and significantly reduces the variance of the Pareto estimate. This is the main point of Vermeulen (2018) and it largely holds in our specification as well, although some bias is present as we only have one specific draw of the rich list.

The main difference between the two setups are the standard errors (SE). Without including the rich list, the SE of the OLS Pareto estimate is similar in both specifications, suggesting that our extension does not limit the variation in survey samples. This is an expected result since the population size is 1 million and only 750 households are sampled each time. However, when the rich list is included in the estimation, our specification reduces the standard error to only 0.006. In the original specification where the population (and thus the rich list) varies in each iteration, the standard error is 0.03. Holding w_{min} and the top tail population size fixed, it is the rich list that determines the top wealth distribution and not the survey observations.

Appendix: Simulation of different merging point approaches: Alternative scenarios

In our benchmark simulation in Section 4.1, we have deviated from the setup of Blanchet et al. (2022) by assuming that the tabulated tax data accurately represent the entire distribution and that the population size is 1 million. Here we report results of two Monte Carlo simulations which assume that the tax data are biased and of one simulation where the population size is increased.

For the first specification, bias is introduced by multiplying each tax bracket by a coefficient that is 0 until the 50th 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 Blanchet et al. (2022) and the results are reported in Table 5. The difference between the three optimal merging point approaches, *BFM*, *K* and *K direct*, is minimal.

Table 5. Distribution of optimal merging point estimates: Tax data biased until the 90th 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 5 reports the share of merging point estimates which fall within each range. This scenario reflects the assumption of Blanchet et al. (2022) about that there is a downward bias in the tax data up until the 90th 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 6 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 our benchmark specification: The *K* method identifies the merging point near the optimal value, the 90th percentile, more often than the *BFM* method. The *K direct* methods again performs the worst.

Table 6. Distribution of optimal merging point estimates: Tax data biased until the 70th 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 6 reports the share of merging point estimates which fall within each range. This scenario assumes that the tax data are biased until the 70th percentile and accurate afterwards.

Finally, we report the sensitivity of results to the population size. Our Monte Carlo simulations are computed using a population of 1 million, which is much less than the 9 million population in Blanchet et al. (2022). Population size determines the gross sample size, which is 1 % of the total. The smaller population was chosen due to computational demands of our method, which relies on integrating the survey's adaptive kernel density over each percentile-based bracket. We run a new simulation with the population size increased to 2 million and compare it with our benchmark setup in Table 7. When the population size is increased, the performance of both the *BFM* and *K* methods improves, with the *K* method remaining superior. This suggests that our results are not sensitive to the population size.

As the *BFM* method is much less computationally demanding, Table 7 also reports its performance when a population of 9 million is utilized. The *BFM* method's performance further improves: The merging point is between the 89th and 91st percentiles in 29 % of cases, compared to 13.5 % in the baseline setup and 20.6 % when the population size is 2 million. Nonetheless, it still lags behind our method in the baseline setup, which is in the optimal interval in 45.1 % of cases.

Table 7. 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

Note: Table 7 reports the share of merging point estimates which fall within each range. Three scenarios are compared, differing only in the population size.

Appendix: Top 1 % wealth share estimates with different Pareto thresholds.

Table 8. Top 1 % shares with different thresholds, MLE estimates.

	Survey	1 mil.	1.5 mil.	2 mil.
2011, original weights	23.2	37.3	31.3	29.4
	(7.3)	(21.9)	(18.0)	(16.2)
2011, calibrated weights	21.5	38.6	28.6	28.4
	(7.2)	(24.2)	(17.2)	(16.7)
2014, original weights	25.5	21.8	23.1	28.5
	(8.1)	(7.1)	(11.3)	(20.2)
2014, calibrated weights	36.6	32.6	44.8	59.5
	(15.0)	(18.9)	(28.2)	(38.3)
2017, original weights	22.8	21.6	20.3	22.2
	(5.8)	(3.9)	(4.7)	(7.8)
2017, calibrated weights	27.3	27.8	25.5	28.5
	(6.8)	(9.2)	(10.8)	(15.4)

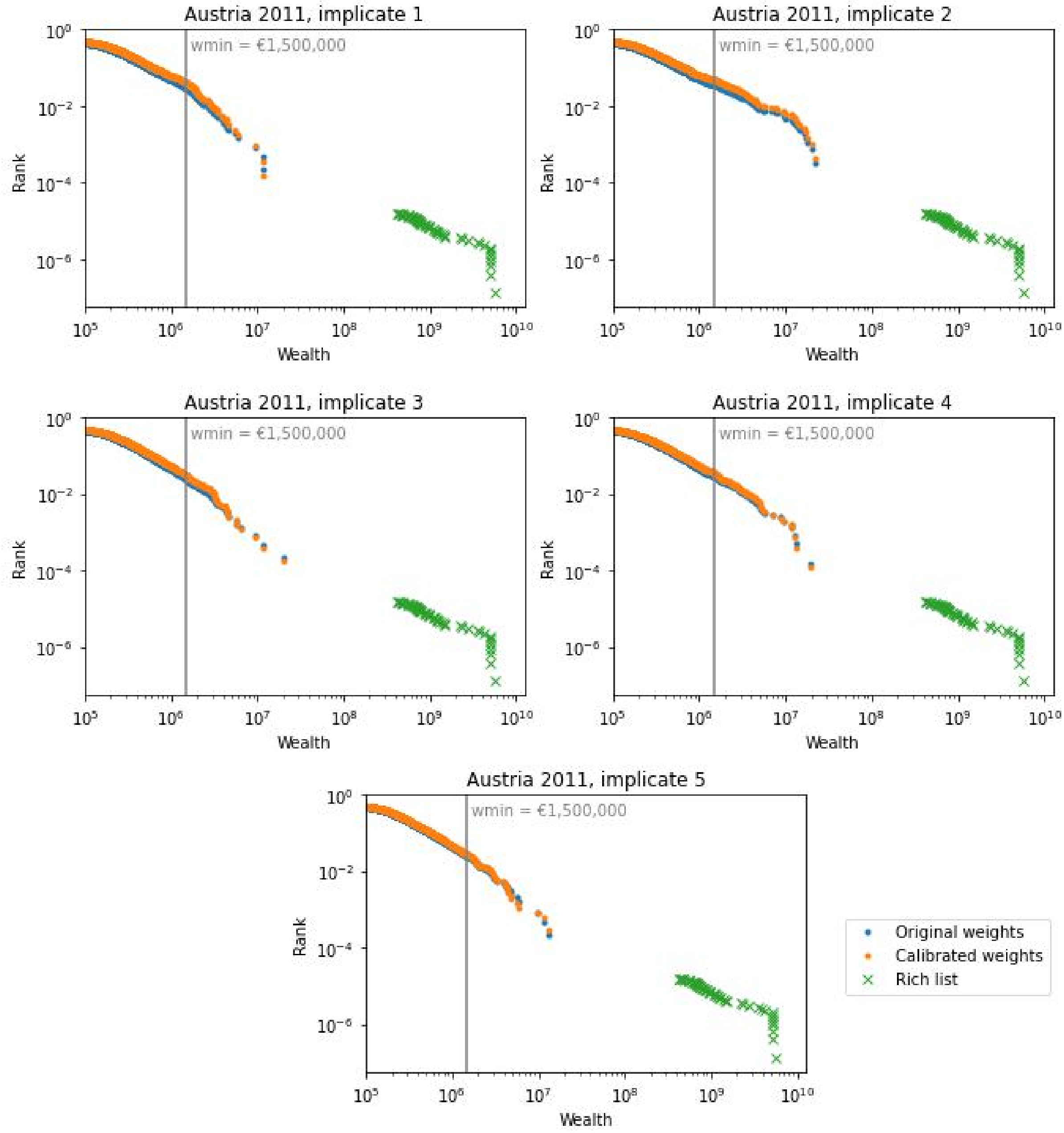
Note: Bootstrap standard errors in parentheses.

Table 9. 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
	(8.1)	(0.6)	(1.3)	(2.3)
2014, calibrated weights	36.6	39.5	38.4	37.6
	(15.0)	(0.7)	(1.3)	(2.2)
2017, original weights	22.8	44.8	44.0	41.8
	(5.8)	(0.5)	(0.9)	(1.5)
2017, calibrated weights	27.3	43.1	43.0	41.9
	(6.8)	(0.6)	(0.8)	(1.5)

Note: Bootstrap standard errors in parentheses.

Appendix: Differences between implicates



Austria 2011, implicate 1

Austria 2011, implicate 2

Austria 2011, implicate 3

Austria 2011, implicate 4

Austria 2011, implicate 5

Figure 11. Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, First wave.

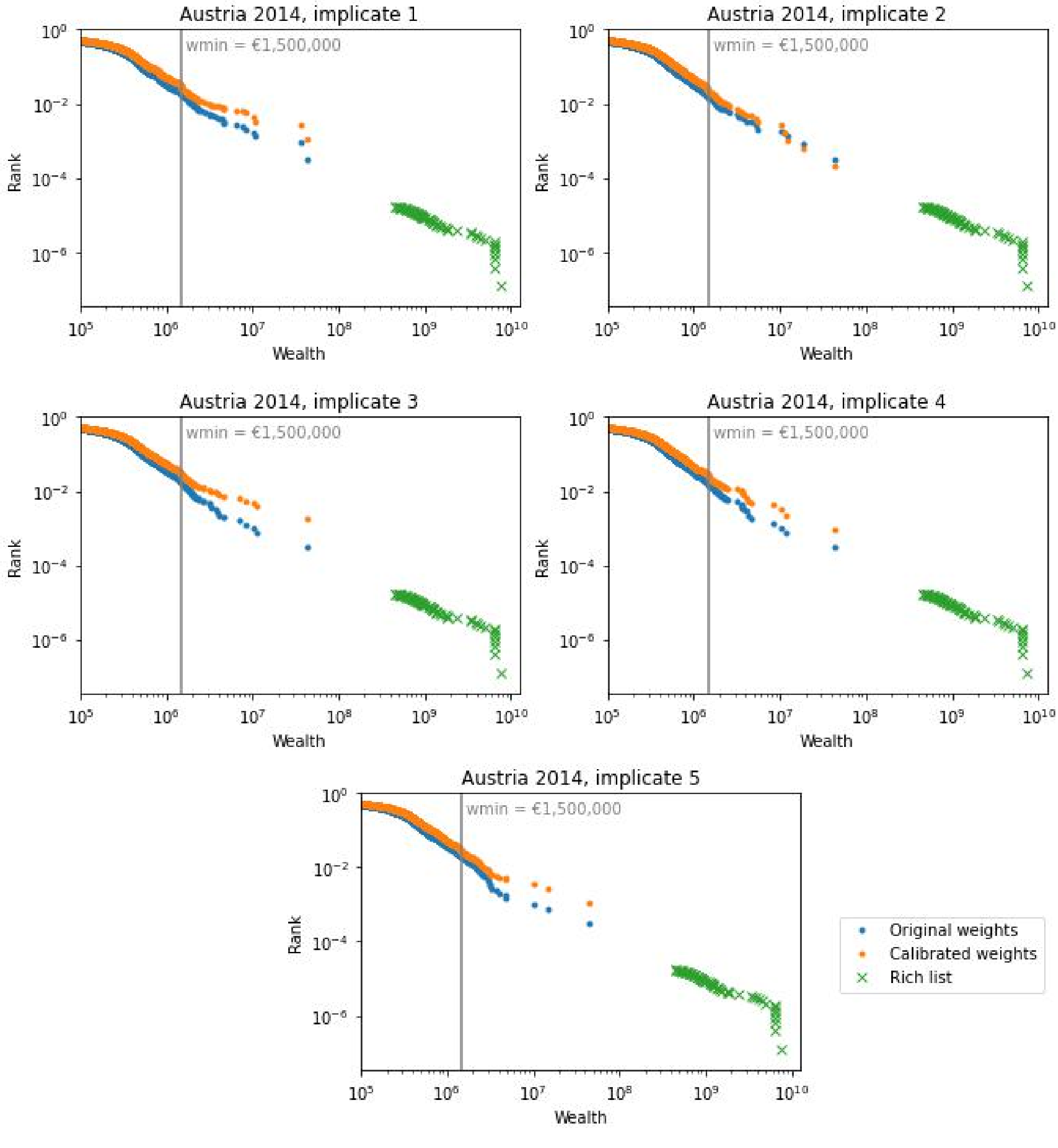


Figure 12. Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, Second wave.

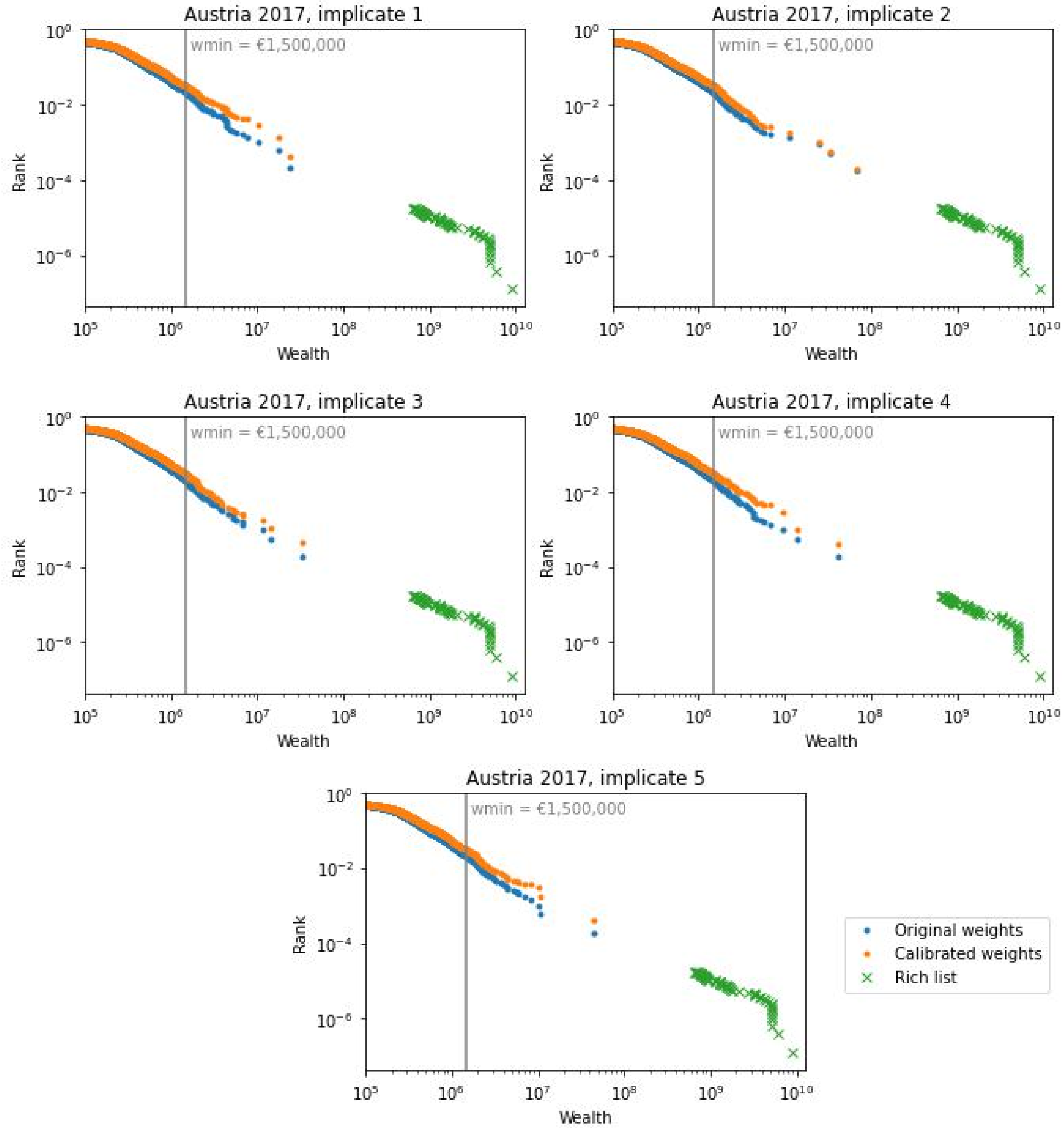


Figure 13. Wealth distribution (complementary cumulative distribution function) based on original and calibrated weights. Data: HFCS, Third wave.