Pareto Tail Estimation in the Presence of Missing Rich in Compiling Distributional National Accounts

Jorrit Zwijnenburg
(OECD)
Jorrit.ZWIJNENBURG@oecd.org

Joseph Grilli
(OECD)
Joseph.GRILLI@oecd.org

Pao Engelbrecht

Paper prepared for the 37th IARIW General Conference

August 22-26, 2022

Session 6C-1, Reducing Gaps between Micro and Macro Statistics on Household Income, Consumption, and Wealth in Compiling Distributional National Accounts I

Time: Friday, August 26, 2022 [9:00-10:30 CEST]

Organisation for Economic Co-operation and Development

DOCUMENT CODE

For Official Use **English - Or. English**

Distributional National Accounts

Pareto tail estimation in the presence of missing rich in compiling distributional national accounts

Joseph Grilli, Pao Engelbrecht and Jorrit Zwijnenburg

Recent measurements of inequality have aimed to align distributional results from micro-level data to national accounts totals. However, micro data (and in particular survey data) is often considered to underestimate the top end of the distribution. Pareto distributions are often appended to represent these missing households, making assumptions about its presence and shape or relying on external information.

Using data from the Luxembourg Income Study for Japan, the Netherlands and the United States, this paper tests for the existence of Pareto distributions for components of primary income, and whether these distributions are truncated. It then presents a novel sampling method to append the survey data.

The results find evidence of greater concentrations of primary income at the top of the income distribution than shown in the unadjusted micro data, with the top 10% of households holding 50.2% in the United States, 33.3% in the Netherlands, and 33.4% in Japan.

Keywords: Income Inequality, National Accounts, Pareto distribution, Truncated distribution
JEL Classification: C24, C46, D31

Joseph Grilli (Joseph.GRILLI@oecd.org)
Jorrit Zwijnenburg (Jorrit.ZWIJNENBURG@oecd.org))

1. Introduction

In response to recent research findings and political changes in the 21st century, there has been increasing demand for distributional results in line with national accounts' totals. The recent economic crises and the subsequent pursuit of (sometimes unconventional) monetary and fiscal policies, such as asset purchase programmes, austerity programmes, and furlough schemes, have triggered questions on how these have affected inequalities. While micro statistics are available that may provide more insight into these questions, these results usually don't align with macroeconomic statistics used in modelling and as input for policy decisions. This has resulted in a number of initiatives to link micro and macro statistics for the household sector, to arrive at distributional results in line with macroeconomic estimates.

While literature on inequality has long existed, with Kuznets (1955[1]) questioning the link between economic growth and inequality, contemporary work has largely been shaped by Piketty (2001[2]; 2001[3]; 2003[4]) and the World Inequality Database (WID) (2020[5]). The WID approach links administrative data from tax returns, augmented with survey data, to national accounts data to estimate the concentration of income and wealth in different percentiles of the economy. The work finds that inequality has been growing since the early 1980s, with over 30% of total income concentrated in the top 10% of earners by 1998. Their method of linking micro statistics to national accounts totals has become known as DINA (Distributional National Accounts).

Furthermore, a lot of work has been done over the past couple of years by international organisations. The Organisation for Economic Co-operation and Development (OECD) and Eurostat launched a joint expert group in 2011 to develop methodology for the compilation of distributional results in line with national accounts' totals, focusing on income, consumption and saving, and the European Central Bank (ECB) recently started work on the development of distributional results on wealth. These work streams heavily rely on both survey and administrative data, but focusing on somewhat different concepts than the DINA work and relying on slightly different assumptions in the compilation process. This work is known as DNA (Disparities in a National Accounts framework) and DFA (Distributional Financial Accounts).

As these projects follow a process of linking micro and macro statistics, requiring a source of the disaggregated distributional information that can be matched to concepts in the macro statistics so as to be consistent with national accounts data, one of the main challenges is to deal with any gaps between the micro and the national accounts' aggregates.[1] These will have to be allocated to relevant households, depending on the underlying cause for the gap.

Administrative data, such as tax data used in Alvaredo et al. (2020[5]) and Piketty, Saez and Zucman (2018[6]), better capture the top of the distribution and can provide detailed and accurate granular data, or aggregate totals for various groupings within the economy. However, due to exemption conditions, it may not provide information on the full population (e.g., omitting specific groups, often at the lower end of the distribution). Furthermore, the definitions of the items included in the dataset may not perfectly match the statistical requirements. Finally, some specific items (or sub-items) may be not be covered.[2] This may all lead to gaps with the macro totals.

In contrast, survey data typically targets households or individuals (henceforth, households), including tax exempt groupings, to collect a wide array of data, including components that may not be covered by administrative records, socio-demographic variables, and other variables such as spending and saving

[1] Another important challenge is to determine the distribution for items for which no direct corresponding micro data source is available. For more information on this issue, see OECD (2022, forthcoming[15]) and Zwijnenburg (2021[43]).

[2] It also needs to be borne in mind that strict confidentiality conditions may sometimes limit the use of administrative data for statistical purposes.

habits. As such, survey data usually contains more complete information on the lower regions of the distribution who might fall below the tax threshold, and on untaxed items. However, surveys normally suffer from two key problems: Misreporting and (differential) non-response.

Misreporting occurs when households report incorrect values. This may be intentional by the household, or unintentional due to imperfect recall.[4] Whereas, theoretically, the values may both exceed or fall short of the actual values, under-reporting is more common than over-reporting. Differential non-response occurs when the households' contact for the survey chooses not to respond, with evidence suggesting that this is more common at the top of the income and wealth distributions (Kennickell and Woodburn, 1999[7]; Sabelhaus et al., 2013[8]). Despite oversampling methods to try and account for this, estimates of wealth in the top share of the distribution are still almost always underestimated when survey data is not augmented with external data sources (Vermeulen, 2016[9]; 2018[10]). It is assumed that this is also the case for the income distribution and that this may explain part of the gap between survey aggregates and national accounts' totals (Zwijnenburg, 2022[11]). Whereas administrative data may also suffer from misreporting and (in case of tax exempt groups) from differential non-response, survey data usually show larger gaps towards the macroeconomic totals than administrative data (Angel, Heuberger and Lamei, 2018[12]; Burkhauser et al., 2018[13]; Burkhauser et al., 2018[14]).

This paper aims to address the differential non-response problem for the top-tail when using survey or administrative data as input for distributional national accounts, by estimating Pareto forms the top-tail of item distributions on the basis of a range of approaches established in the literature. The objective is to account for the differential non-response by rich or high income households in survey data, using methods that rely as much as possible on the available information and that are easily reproducible. This contributes to the so-called centralized approach as described in OECD (2022, forthcoming[15]) to compile inequality measures for countries that are not compiling these results themselves.

The paper is structured as follows. Section 2 outlines the literature on differential non-response for the top of the distribution and how this may be approximated by a Pareto distribution. Section 3 presents the methodology used in this paper to estimate forms of the top-tail of the distribution in DNAs to find the best fitting estimation, and to explore possible options to adjust the underlying data. Section 4 applies the method to micro data available from the Luxembourg Income Study (LIS) to compare measures of inequality and the impact of the gap between macro and micro statistics, focusing on estimates of primary income. Section 5 reports the key findings and concludes the paper.

[4] Survey data may also include measurement differences or errors, or timing differences when compared to national accounts' measurement. While these may contribute to a gap between the national accounts and the survey aggregates that should be minimised when constructing distributional national accounts, this is not regarded as misreporting by the households. This is because survey data is not designed to be a complement national accounts statistics, serving their own distinct purpose unrelated to the national accounts.

2. Treating Differential Non-Response at the top

While distributional national accounts are a relatively recent development in the inequality literature, the issue of missing rich affecting the top of the distributions pre-dates this. Davis and Shorrocks (2000[16]) discuss how household surveys have to be augmented with independent estimates of the richest households, and Vermeulen (2014[17]) summarises the evidence of positive skews in the wealth and income distributions and why downward bias may exist in surveys. When constructing distributional national accounts on the basis of survey data, it has been noted that known rich individuals were not represented in the results, impacting the totals captured and the inequality measures calculated (Chakraborty and Waltl, 2018[18]; Vermeulen, 2016[9]; 2018[10]). Despite methods to capture rich households in the samples, these households were not reflected in the results due to their lower propensity to respond to surveys (Kennickell and Woodburn, 1999[7]).

Davis and Shorrocks (2000[16]) describe how the distribution of wealth in the top-tail of the wealth distribution can be well approximated by a Pareto distribution. This has been used to approximate the tail that may be poorly covered in surveys, providing greater values with higher densities than seen in survey data. The approximation means that wealth above a threshold γ, is proposed to follow a Pareto distribution, with the Pareto Type I given,

$$P(X > x) = \left(\frac{\gamma}{x}\right)^{\alpha}, x \geq \gamma$$

where $x \in [\gamma, \infty)$ and $\alpha > 0$. α is the shape parameter, determining the "fatness" of the top-tail, with smaller values of α corresponding to greater density at higher levels of wealth. This approximation by the Pareto distribution approximately captures a power law in wealth, identified as early as work by Pareto (1896[19]; 1897[20]) and Kuznets (1955[1]), where higher values of wealth are held by a small number of individuals. While recent work has deviated from the Type I specification of the Pareto distribution, such as Atkinson (2017[21]) and Blanchet, Fournier and Piketty (2017[22]), it is widely accepted that the family of Pareto distributions well represents the density of wealth of the population in the top-tail of the wealth distribution.

Building on the existing literature of power laws, Vermeulen (2016[9]) shows how high quality survey data can be approximated by a Pareto distribution and that publicly available sources on rich individuals who have not been captured in the survey tend to follow this distribution, using the United States' Survey of Consumer Finance (SCF) and the Forbes 400 list to demonstrate this. He uses this external data, combined with survey data above the top-tail threshold, to estimate Pareto distributions for top-tails of wealth distributions to correct for downward bias in the distribution parameters seen when estimating on survey data alone elsewhere. Similarly, Burkhauser et al (2018[13]) and Piketty and Saez (2003[23]) use external data from administrative data sources to correct the top-tail of the survey using power laws.

Törmälehto (2019[24]) demonstrates that by using an external source for the share of income held by individuals above a threshold γ, the parameters of a Pareto Type 1 distribution can be easily recovered. If administrative data provides the share of income earned by individuals with values above threshold γ, i.e., $\sum_{i=1}^{N} \mathbb{I}(x_i \geq \gamma) x_i = s_{x_i \geq \gamma} \sum_{i=1}^{n} x_i$, then the shape parameter of a Pareto distribution for this total income item can be estimated via the following formula if $\alpha > 1$,

$$s_{x_i \geq \gamma} \sum_{i=1}^{n} x_i = \sum_{i=1}^{N} (\mathbb{I}(x_i \geq \gamma) \times x_i) = N_0 \times E(x | x_i \geq \gamma) = N_0 \times \frac{\alpha}{\alpha - 1} \gamma$$

where x_i is the earning of an income item by individual i, N_0 is the total number of individuals above threshold γ, and α is the shape parameter of the Pareto Type 1 distribution.[5] This approach uses the administrative data for the total value of the tail and the assumption of a Pareto Type I form to calculate the distribution of values above γ, meaning that the concentration of individuals above γ can be calculated from this distribution, approximating the missing households. While this offers a solution to the differential non-response at the top, Atkinson (2017[21]) warns against this simplistic form of the Pareto distribution. He demonstrates how estimates of α can vary significantly depending on the estimation method and on the lower threshold value γ. He concludes moving beyond the Pareto Type I form is necessary to resolve these deviations.

Blanchet, Fournier and Piketty (2017[22]) move beyond the standard Pareto distribution to the generalized Pareto distribution. This introduces a scale parameter σ, to the distribution, with the function becoming,

$$P(X > x) = \left(1 + \alpha\left(\frac{x_i - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}, x_i \geq \gamma$$

To estimate the function, they use splines to interpolate tabulations of income and wealth data from tax authorities, extrapolating the last bracket with the generalized Pareto, and to estimate the parameters.[6] They conclude that the generalized Pareto distribution provides additional richness to move beyond the Pareto type 1 distribution and explain the deviations from the standard approaches seen in tabulated tax data.

Although, as shown above, literature shows different forms of application, it is clear that the top-tail of income and wealth distributions can be modelled by a Pareto distribution of sufficient richness and that fitting this to survey data can capture the rich households missing due to differential non-response. Blanchet, Chancel and Gethin (2019[25]) and Jenkins (2017[26]) use this approach, i.e., estimating the rich households based on tax return data and a generalized Pareto form, and 'non-rich' households based on survey data, to demonstrate how this corrects for the bias seen in pure survey data when matched to macro statistics. While these approaches offer sophisticated methods to estimate inequality, they require external administrative data on tax returns or lists of omitted rich households. From this external data, they project the shares of income held by households within this distribution based on the density function of this top-tail rather than the construction of a dataset that can be merged or appended with the existing survey data. In practice, access to this external data can be difficult to obtain, and may also have the risk of only representing a subset of the population and of items reported in national accounts. Furthermore, it may suffer from conceptual or technical differences when comparing to national accounts.

In the absence of external data, Lakner and Milanovic (2015[27]) use assumptions on the share of the gap between micro and macro statistics held by the richest households (which can be informed from external data or based on an ad-hoc assumption) and calculate the shares of total income within this group using the Pareto distribution. However, this approach assumes that a Pareto distribution exists in the top-tail that

[5] $\mathbb{I}(x)$ denotes the identity function where $\mathbb{I}(x) = 1$ for all $x \in X$. In the Pareto literature, where the function is only defined for $x \geq \gamma$, we use the identity function to select observations above the Pareto minimum threshold. This is especially applicable here as we consider multiple threshold values.

$E(x)$ denotes the expected value, here specially for the Pareto Type 1 distribution. This is equivalent to the mean of the Pareto Type 1 distribution, which here is $E(x|x_i \geq \gamma) = \frac{\alpha}{\alpha-1}\gamma$.

[6] Splines are a wide class of functions used for interpolation between fixed points. Blanchet, Fournier and Piketty (2017[22]) use quantic splines (i.e., polynomials of degree 5) to create functions that meet the restrictions of equalling administrative data totals and to be twice differentiable. This provides the basis to estimate the distribution, which is then constrained to estimate the generalized Pareto distribution. Blanchet et al. (2018[42]) demonstrate the effect of this on calculations of the top 10% share.

is both related to the size of the micro-macro gap and is uncaptured. Furthermore, the estimation does not utilise the available micro data in estimating or testing the goodness of fit.

For the approaches discussed so far, results do not provide a micro dataset consistent with national accounts' totals as they calculate the total value of sections of the tail, therewith limiting the possibility to publish socio-demographic characteristics for all household groups across the distribution and to arrive at very granular results.[7]

To provide micro statistics that are consistent with national accounts, this paper draws upon the literature on the top-tail adjustments and investigates methods to estimate the top-tail using assumptions on survey data (rather than external data sources), and sampling from this survey data to construct a granular dataset from which micro statistics in line with national accounts' totals can be constructed. Household data from the Luxembourg Income Study (LIS) is linked to national accounts, following OECD (2022, forthcoming[15]). In Section 3, the process to estimate top-tail distributions using this data and how to draw samples from the estimation to recover a granular dataset for the centralized approach is described.

[7] Where administrative tax data is available and consistent with national accounts, one could use approaches such as Blanchet, Fournier and Piketty (2017[22]) instead of those laid out in Section 3.1 to estimate the top-tail of the distribution, then apply the adjustment method in Section 3.3 to the survey data. However, here we consider approaches calculated on survey data alone and compare these results with those using administrative data to see whether estimates can be made to fill the gaps when administrative data is not available.

3. Methodology

When treating differential non-response by applying Pareto adjustments to the top-tail of survey data, the top-tail of the survey first needs to be identified to decide at which point households are considered 'rich'. From there, the form of the tail needs to be estimated, as well as how to reflect the distribution in the data. Explicitly, we begin with the following assumptions:

> *Assumption I: Some instruments of income and consumption have a Pareto distribution above a threshold γ.*

> *Assumption II: The probability of individuals above γ entering the sample is lower than their density in the true population they are drawn from, as they are less likely to respond to surveys, even those including oversampling, and are therefore less represented by those who do respond.*

> *Assumption III: In survey data, the maximum observation is not always representative of the maximum values in the population, truncating the distribution.*

Assumption I is necessary to define the lower threshold parameter of the Pareto distribution as well as the point where 'rich' households begin. A number of approaches are taken in the literature to define γ, but the most common approach in the income inequality literature is to define γ by the threshold value for the top 10% of households, ordered by income (Blanchet, Chancel and Gethin, 2019[25]). Occasionally, this is adjusted to the top 5% or top 1% when the proportion of households holding an instrument is low, with Törmälehto (2019[24]) using the top 5% for self-employed income and interest and dividend income.[8] Sabelhaus et al. (2013[8]) and Blanchet, Chancel and Gethin (2019[25]) discuss this threshold, motivating its use by non-response rates that are stable over most of the distribution, but increase linearly for the top 10%. While literature elsewhere has considered using fixed values for thresholds or by determining the best fit of the Pareto, this motivation is more consistent with the differential non-response motivation.

Assumption II represents the lower response rate, stating that observations above γ are less represented in the sample than in the population. This assumption means that we consider the reporting households as being as truthful as those anywhere else in the distribution, but that they do not represent the underlying population as closely as households elsewhere in the distribution. An alternative hypothesis to explain the missing density would be that the households in the 'rich' part of the distribution do report, but under-report to a greater extent, resulting in lower densities at higher values. However, there is no evidence to support under-reporting being relatively greater for higher levels of income. We therefore want to retain the observation data as much as possible to provide richer granular data for analysis while also correcting for the missing households, rather than trimming the sample and replacing the top-tail.

Assumption III states that surveys do not capture the richest households in the economy. This is the basic assumption for the method used by Vermeulen (2018[10]), adding external data to correct for this omission.

[8] Some sources of income and wealth are owned by a small number of households, typically richer households who are not restricted by barriers to entry. Here, we may consider increasing the threshold to the top 5% or top 1% to remove households with zero or small holdings who could still be captured in the top 10% but would not follow a Pareto distribution. However, this has to be balanced against the number of observations used to estimate the tail, as increasing the threshold reduces the number of households and reduces the degrees of freedom in the estimation.

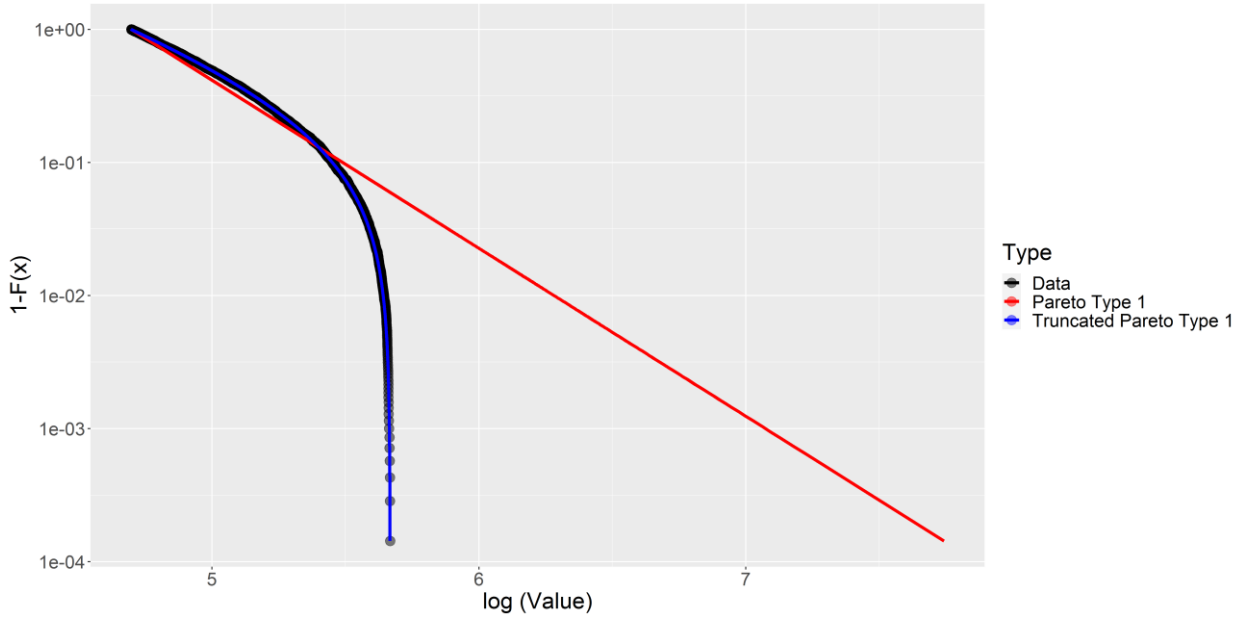
Missing rich households from the survey can be by design, as the survey has to maintain anonymity amongst respondents, but it can also be related to the fact that the very rich may not be captured in the survey design.

In cases where items are heavily concentrated, this can cause large gaps and is likely to significantly bias the parameter estimates when minimising least squares methods as the model will treat the top micro data observation as the richest observation when fitting the sample, and does not account for the missing observations in the population that are ranked higher, reducing the estimated number of households at the top of the distribution. In Figure 1, the data points are simulated from a Pareto Type 1 distribution, then truncated by removing the top observations. If this truncation is not accounted for, the Pareto Type 1 estimation would appear as a straight line on the Zipf plot, attempting to minimise the deviations of the data points from the line. Without correcting for the truncation, the rank of the observations would be biased upwards as missing observations from the top of the distribution would result in households with lower incomes being ranked higher (for example, if the top 100 households were removed, the 101st richest household would be treated as the richest household, with no density above their reported value). Without correction, this would cause the model to estimate incorrect parameter values, producing a steeper line and underestimating the density for higher income values

Vermeulen (2018[10]) demonstrates how estimates with low density at the top of the distribution overestimate α in the Pareto Type 1 distribution, giving a lower inequality estimate. This is typically resolved via the use of external data, as obtained from administrative sources or rich lists. In this paper, however, we consider an alternative approach.

Aban, Meerschaert and Panorska (2006[28]) observe that while the underlying (true) distribution of data may be Pareto distributed, cases may exist where the tail in the available micro data behaves like a truncated Pareto. They highlight how this often occurs due to practical limitations, pointing to sampling practices such as taking average measurements and the application of upper bounds in data collection from distributions that generate extreme values. In these cases, truncated distributions fit the data better than untruncated ones, even if the true underlying distribution is likely to be untruncated, such as is expected for the distributions of the very rich. When truncated, they describe a characteristic downward curve produced in a Zipf plot, i.e., a log-log graph of the ordered rank of observations against the values. In contrast, the standard Zipf plot of an untruncated Pareto distribution normally shows a linear log-log relationship between the value of the observation and the rank of the frequency. Figure 1 shows how for the truncated Pareto distribution the data initially suggests a linear relationship, before sharply declining at the truncation point. This is, because compared to the untruncated case, the lower value observations have a higher rank, as the top observations are missing. If a standard Pareto is estimated on truncated data, α increases to reduce the density at greater values, resulting in a biased estimate from the misspecification.

Figure 1. Estimation of Simulated Truncated Pareto Data using Untruncated (red) and Truncated (blue) Pareto Type 1



In this paper, the top-tail adjustment is applied on an item-by-item basis focusing on household primary income from the linking process in OECD (2022, forthcoming[15]), using the top 10%, 5%, and 1% of the survey data for each item to estimate the top-tail of the distribution. As Blanchet, Fournier and Piketty (2017[22]) demonstrate, the standard Pareto Type 1 distributions, while easier to work with, can lack the richness needed to explain the data, being outperformed by the generalized Pareto distribution. We therefore estimate both the Pareto Type 1 and generalized Pareto top-tail distributions using the survey data.

In this section, we discuss how the method is applied. We first consider the estimation method, deriving the estimator for the selected forms so that the parameter estimates can be recovered for the candidate distributions for the top-tail. We then present a goodness-of-fit test that can be applied to each of the estimated distributions. These tests show whether the data used to estimate the distributions support the Pareto form, as this allows us to identify whether the data provides evidence of a Pareto form for specific items, and, if so, which form of the Pareto tail is most suitable and whether part of the Pareto tail may already be captured in the data. Finally, we examine a range of methods to adjust the top-tail of the data having found evidence of a Pareto distribution. This identifies the proportion of the tail omitted and how this can be captured while retaining features of the survey data. We conclude the methodology with a final step to match national accounts' totals, as micro-macro gaps still often exist in cases where there is no evidence of missing rich households or after these households have been adjusted for.

3.1. Estimation of Model

The first stage of the methodology is to construct the estimators for the top-tail of the distribution. Having first defined these untruncated specifications of the Pareto distribution, the truncated versions of each form are then derived.

For each of the candidate distributions, the probability density function (pdf) is used to assess the likelihood of observing the survey values seen above the top-tail threshold conditional on the distributions form and the parameter values, giving $p(x_i|f(x,\theta),\theta)$ for each observation x_i, where θ gives the parameter values and $f(x,\theta)$ gives the distribution density function. The maximum likelihood estimator then determines the parameter values most likely to create the observations for the forms considered. This estimate is then used to examine whether the top-tail of items is likely to follow one of the candidate distributions, and to sample missing rich households from in case there is evidence that these are missing. As we use complex survey data where observations have weights, the pseudo-maximum likelihood estimator as defined in Vermeulen (2018[10]) is used. This incorporates the weights for each observation in the maximum likelihood estimator.

Considering first the Pareto Type 1, the simplicity of the form means that α can be solved for with a log transformation. Defining the function,

$$F(x) = 1 - \left(\frac{\gamma}{x_i}\right)^{\alpha}, x_i \geq \gamma$$

the (log) likelihood function can be derived to calculate the pseudo-maximum likelihood estimate of α,

$$L = \prod_{i=1}^{n} \left(\frac{\alpha}{x_i}\left(\frac{\gamma}{x_i}\right)^{\alpha}\right)^{w_i}$$

$$\log(L) = LL = \sum_{i=1}^{n} \log\left(\left(\frac{\alpha}{x_i}\left(\frac{\gamma}{x_i}\right)^{\alpha}\right)^{w_i}\right) = \sum_{i=1}^{n} w_i \log\left(\frac{\alpha}{x_i}\left(\frac{\gamma}{x_i}\right)^{\alpha}\right)$$

$$LL = \sum_{i=1}^{n} w_i\big(\log(\alpha) - \log(x_i) + \alpha(\log(\gamma) - \log(x_i))\big)$$

$$LL = N\log(\alpha) + N\,\alpha\log(\gamma) - (1+\alpha)\sum_{i=1}^{n}\log(x_i)$$

$$\frac{\partial LL}{\partial \alpha} = \frac{N}{\alpha} + N\log(\gamma) - \sum_{i=1}^{n}\log(x_i) = 0$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i \log\left(\frac{x_i}{\gamma}\right)}$$

where $N = \sum_{i=1}^{n} w_i$, with w_i being the weight of household i and x_i being the value reported by household i. Using the generalized Pareto distribution from Blanchet, Fournier and Piketty (2017[22]), the function is defined,[9]

$$F(x) = 1 - \left(1 + \alpha\left(\frac{x_i - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}, x_i \geq \gamma$$

As with the standard Pareto, the log likelihood function is then derived. However, the shape parameters for the generalized Pareto, $\{\alpha, \sigma\}$, cannot be separated here, so that numerical methods have to be used to find (pseudo-) maximum likelihood estimates,

[9] In Blanchet, Fournier and Piketty (2017[22]), the function is actually defined, $F(x) = 1 - \left(1 + \xi\left(\frac{x_i-\gamma}{\sigma}\right)\right)^{-1/\xi}, x_i \geq \gamma$, with ξ characterising the shape parameter. This is equivalent to $\alpha = \frac{1}{\xi}$, where α is the Pareto shape parameter. However, here we use α to denote all shape parameters for the purposes of displaying estimation output.

$$L = \prod_{i=1}^{n}\left(\frac{1}{\sigma}\left(1+\alpha\left(\frac{x_i-\gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}-1}\right)^{w_i}$$

$$\log(L) = LL = \sum_{i=1}^{n} w_i \log\left(\frac{1}{\sigma}\left(1+\alpha\left(\frac{x_i-\gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}-1}\right) = \sum_{i=1}^{n} w_i\left(\log\left(\frac{1}{\sigma}\right) - \left(\frac{1+\alpha}{\alpha}\right)\log\left(\frac{\sigma+\alpha(x_i-\gamma)}{\sigma}\right)\right)$$

$$LL = -\sum_{i=1}^{n} w_i\left(\log(\sigma) + \left(\frac{1+\alpha}{\alpha}\right)\log\left(\frac{\sigma+\alpha(x_i-\gamma)}{\sigma}\right)\right)$$

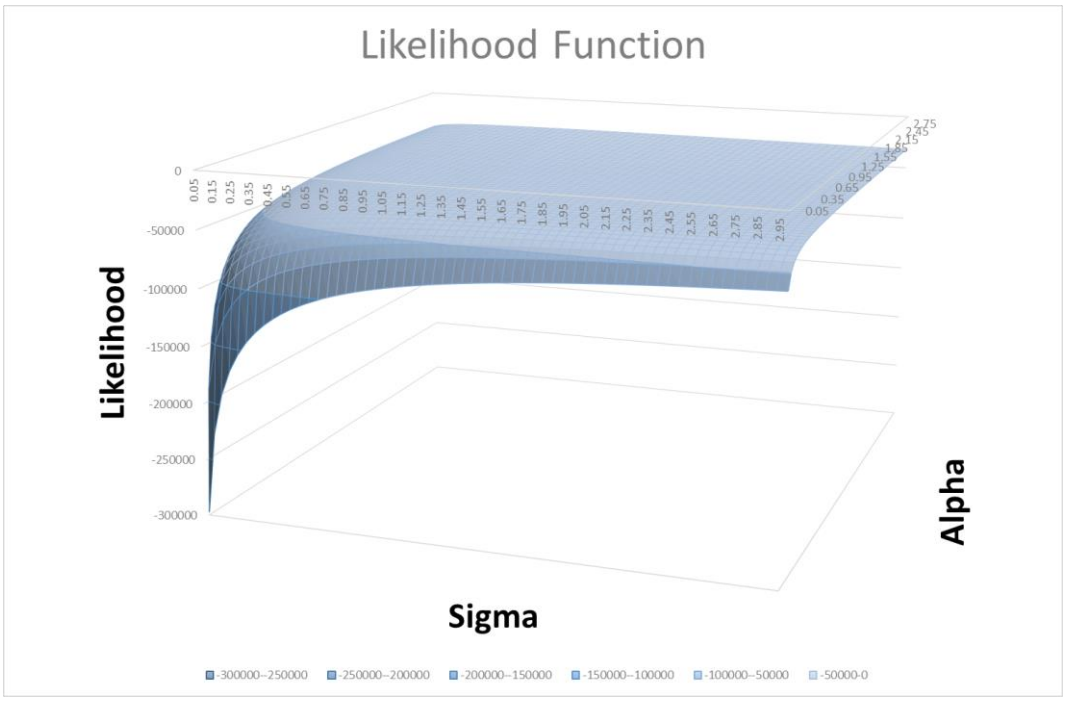


Figure 2. Likelihood function for generalized Pareto distribution (Simulated data using α=1.8, σ=1.2, and γ=100)

In using pseudo-maximum likelihood estimation, Figure 2 shows that the likelihood function is very flat. This poses a challenge for the pseudo-maximum likelihood estimator as smooth functions can be difficult to optimise since multiple sets of parameter values can produce similar likelihoods. This affects optimisation search algorithms and makes identifying the best performing specification difficult as turning points may be missed and there is little indication of where candidate maximums might be. As such, optimisation algorithms may be affected by their starting values, number of iterations, tolerance or other parameters so that the maximum likelihood parameter estimates identified are determined by these parameters rather than those of the underlying distribution of interest.

One could consider introducing Bayesian Estimation Methods, using prior distributions to weight parameter estimate values to give additional curvature over flat areas of the function. However, there is little evidence for what the distribution of such priors should be. We therefore maintain an agnostic approach and use

pseudo-maximum likelihood estimation, although applied approaches may wish to consider priors to ensure solutions with defined moments for the distribution, as finite moments of Pareto distributions are depending on the parameterisation.

Using the Pareto Type I and generalized Pareto forms improves the representation of missing rich households at the top of the distribution, but does not address Assumption III. If the sample data has a significantly lower maximum than the population, the functional form needs to be adjusted to account for the unobserved households not covered by the sample. The forms so far considered have untruncated specifications To estimate the truncated forms of the Pareto distributions, and to create tests checking whether the micro data is significantly truncated, the truncated distribution from Aban, Meerschaert and Panorska (2006[28]) is considered. This introduces an upper truncation point which, following assumption III, is defined: $\nu = \max(x)$. This includes all the survey data but reflects that the distribution in the survey data may be incomplete, biasing the parameter estimates. From assumption I, we believe that the true population distribution is an untruncated distribution and so we want to capture the effect of any possible truncation to remove the bias when estimating α. The truncation point affects the relative frequencies expected from the distribution, most obviously that the density for $x > \nu$ is 0, explaining the shape in Figure 1. For the Pareto Type 1 distribution, the truncated form is given,

$$F(x) = \frac{1 - \left(\frac{\gamma}{x_i}\right)^{\alpha}}{1 - \left(\frac{\gamma}{\nu}\right)^{\alpha}}, \nu \geq x_i \geq \gamma$$

Using this form, Aban, Meerschaert and Panorska (2006[28]) give an estimate of both the shape parameter, α, and also the threshold, γ. This specification then jointly optimises the parameters,

$$L = \prod_{i=1}^{n} \left(\frac{\frac{\alpha}{x_i}\left(\frac{\gamma}{x_i}\right)^{\alpha}}{1 - \left(\frac{\gamma}{\nu}\right)^{\alpha}} \right)^{w_i}$$

$$\log(L) = LL = \sum_{i=1}^{n} \log\left(\left(\frac{\frac{\alpha}{x_i}\left(\frac{\gamma}{x_i}\right)^{\alpha}}{1 - \left(\frac{\gamma}{\nu}\right)^{\alpha}} \right)^{w_i} \right) = \sum_{i=1}^{n} w_i \log\left(\frac{\frac{\alpha}{x_i}\left(\frac{\gamma}{x_i}\right)^{\alpha}}{1 - \left(\frac{\gamma}{\nu}\right)^{\alpha}} \right)$$

$$LL = \sum_{i=1}^{n} w_i \left(\log(\alpha) - \log(x_i) + \alpha(\log(\gamma) - \log(x_i)) - \log\left(1 - \left(\frac{\gamma}{\nu}\right)^{\alpha}\right) \right)$$

$$LL = N\log(\alpha) + N\,\alpha\log(\gamma) - N\log\left(1 - \left(\frac{\gamma}{\nu}\right)^{\alpha}\right) - (1+\alpha)\sum_{i=1}^{n}\log(x_i)$$

While previously the threshold has been determined so that $\gamma = \min(x)$ of the households in the top-tail (i.e., the threshold of the top 10% of households) to estimate the lower threshold, the parameters must be estimated on $r < N$ households, with $x_{r+1} < x_r$. This then gives the parameters,

$$\hat{\gamma} = r^{\frac{1}{\hat{\alpha}}}(x_{r+1})\left(n - (n-r)\,\left(\frac{x_{r+1}}{\nu}\right)^{\hat{\alpha}} \right)^{-\frac{1}{\hat{\alpha}}}$$

$$0 = \frac{r}{\hat{\alpha}} + \frac{r\,\left(\frac{x_{r+1}}{\nu}\right)^{\hat{\alpha}} \ln\left(\frac{x_{r+1}}{\nu}\right)}{1 - \left(\frac{x_{r+1}}{\nu}\right)^{\hat{\alpha}}} - \sum_{i=1}^{r}(\ln(x_i) - \ln(x_{r+1}))$$

This approach is then applied for the truncated generalized Pareto distribution. Using the same approach, the form is defined,

$$F(x) = \frac{1 - \left(1 + \alpha\left(\frac{x_i - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}}{1 - \left(1 + \alpha\left(\frac{\nu - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}}, \nu \geq x_i \geq \gamma$$

As before, the estimation of the generalized function requires numerical methods to get the (pseudo-) maximum likelihood estimates of the parameters,

$$L = \prod_{i=1}^{n} \left(\frac{\frac{1}{\sigma}\left(1 + \alpha\left(\frac{x_i - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}-1}}{1 - \left(1 + \alpha\left(\frac{\nu - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}} \right)^{w_i}$$

$$\log(L) = LL = \sum_{i=1}^{n} w_i \log\left(\frac{\frac{1}{\sigma}\left(1 + \alpha\left(\frac{x_i - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}-1}}{1 - \left(1 + \alpha\left(\frac{\nu - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}} \right)$$

$$LL = \sum_{i=1}^{n} w_i \left(\log\left(\frac{1}{\sigma}\right) - \left(\frac{1+\alpha}{\alpha}\right)\log\left(\frac{\sigma + \alpha(x_i - \gamma)}{\sigma}\right) - \log\left(1 - \left(1 + \alpha\left(\frac{\nu - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}\right)\right)$$

$$LL = -\sum_{i=1}^{n} w_i \left(\log(\sigma) + \left(\frac{1+\alpha}{\alpha}\right)\log\left(\frac{\sigma + \alpha(x_i - \gamma)}{\sigma}\right) + \log\left(1 - \left(1 + \alpha\left(\frac{\nu - \gamma}{\sigma}\right)\right)^{-\frac{1}{\alpha}}\right)\right)$$

For the estimation of the parameters, the Hill (1975[29]) estimator is applied. This is discussed in Aban, Meerschaert and Panorska (2006[28]) and is a common approach to address bias in pseudo-maximum likelihood estimation. For the Hill estimator, we define r, where $0 \leq r < n$ and $x_r > x_{r+1}$, so that estimates of the parameters use a subset of the data. To select r, a simplistic approach is used that maximises the r-squared between the cumulative density function (CDF) and the empirical CDF (ECDF) for the untruncated Pareto Type 1 using 5% to 95% of the ordered data. This r is used over all estimates, as the estimates tend to vary little for small changes in r.[10,11]

Other approaches to select r were considered, including the mean excess function (Langousis et al., 2016[30]) or by minimising the Kolomogorov-Smirnov (KS) test statistics (Clauset, Shalizi and Newman, 2009[31]). However, computational restrictions emerged due to the size of the dataset. Weighted maximum likelihood estimation would provide an alternative approach, but implementation to correct the bias is problematic in the presence of complex survey weights, as it requires reweighting. It was also found that the correlation approach performed better in survey data, although the value of the Hill estimator over the pseudo-maximum likelihood estimation in the more complicated forms of the model is questionable.

3.2. Goodness of Fit

Having estimated possible forms of the top-tail distribution, the fit of each model to the observations from the survey needs to be examined. Firstly, estimated distributions need to be compared to survey data to test whether the distributions are statistically similar, in other words assessing whether the actual observations indeed confirm the existence of any of the proposed top-tail distribution. Secondly, if multiple

[10] If the sample size exceeds 40,000, we use $r = 0.5n$ to reduce computational intensity.

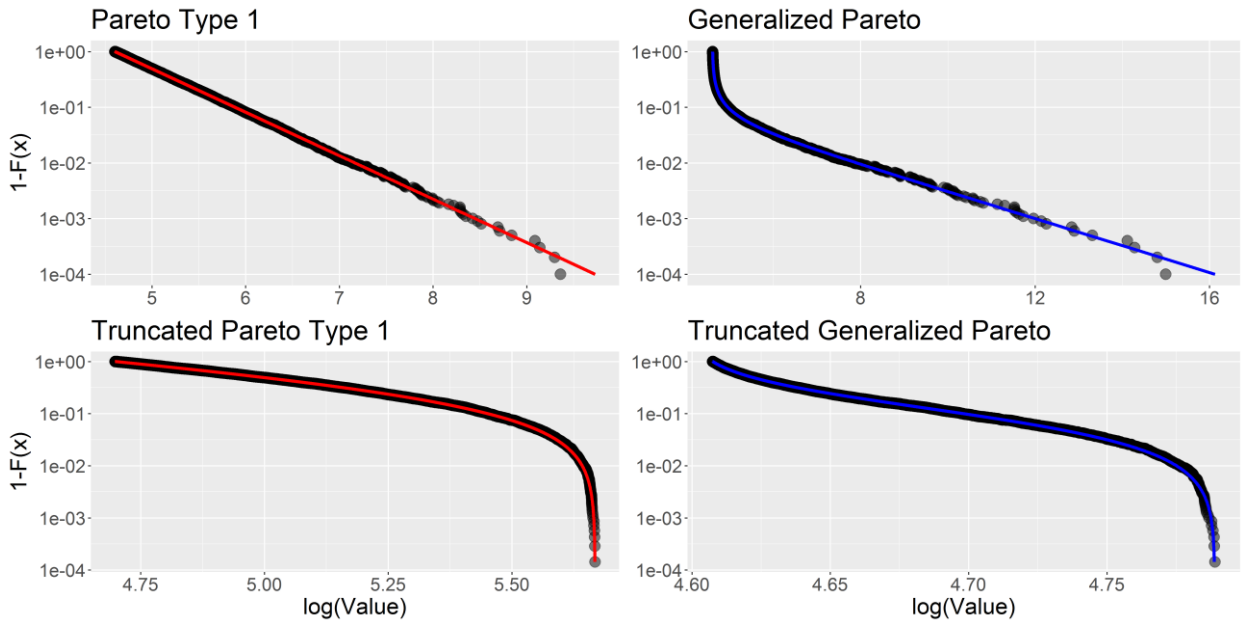
[11] This is not applied to the generalized Pareto Function, as the Hill (1975[29]) estimator was found to be outperformed by the pseudo-maximum likelihood estimator in simulations.

forms are retained in the goodness-of-fit tests, then a preferred model needs to be selected from the retained forms.

From a visual inspection, one could hypothesise which form best fits the data. From the simulated plots in Figure 3, the data on Zipf plots demonstrate varying characteristics across the specifications, with Pareto Type I plots in red, and generalized Pareto plots in blue. The Pareto Type I is characterised by a linear relationship, given by the power law. The generalized Pareto, with the additional parameters, is not necessarily linear, with a limit case that it equals the Pareto Type I. This is seen in the top row of Figure 3, which plots the untruncated forms of the functions. The bottom row of Figure 3 plots the truncated forms, which are characterised in both forms by the characteristics hook shape described in Aban, Meerschaert and Panorska (2006[28]) at the top end of the distribution. The truncation introduces this by removing the highest ranked observations, which has a much larger effect at the top of the distribution (where a lower maximum value is found compared to the untruncated form) than at the bottom of the distribution (where the same minimum exists). Accounting for this truncation is important when estimating the correct parameters for the distribution as it can be used to explain missing density at the top of the distribution due to missing observations that would otherwise bias the estimation of the shape parameters.

Since data can share characteristics of multiple specifications, a hypothesis-based-testing approach is used to identify which form should be used. Whereas in most literature it is assumed that a single specification explains the data, we consider a selection of possible forms and the possibility that data does not follow a power law. When considering an item-by-item approach for income, it is possible that some items may not follow Pareto distributions, with the patterns commonly identified for aggregate income being driven by specific income components.

Figure 3. Simulated data and trends for Pareto specifications considered (data simulated with α=1.8 and, where applicable, σ=1.2, and truncated at 15% and 85%).



To test the goodness-of-fit for each estimated model, a KS test is used. This draws from the findings of Krieger and Pfeffermann (1997[32]) that the KS test performs sufficiently well under complex survey design in large samples, and from Clauset, Shalizi and Newman (2009[31]) stating that the use of KS is important to ensure the power law is present and that the use of bootstrapping is relevant to calculate confidence intervals. Here, we use a standard expression of the KS test, where the null hypothesis is that the data and the estimated distribution have the same CDF, with the test statistic,

$$D = \max_{x \geq \gamma} |S(x) - P(x)|$$

where $S(x)$ is the ECDF of the data for observations above the minimum threshold, and $P(x)$ is the CDF for the power law evaluated at $x \geq \gamma$. Clauset, Shalizi and Newman (2009[31]) propose a reweighting adjustment to make the KS test more uniformly sensitive, but as they find the results to be similar to those found with the standard test statistic, it is not implemented here. The KS test is also proposed as a method for selecting the value for γ, finding the value with the best fit to the distribution. Given the motivation of this paper to adjust for missing households at the top of the distribution, as opposed to identifying a threshold value across the entire distribution that fits the best Pareto distribution to the data, we instead use the threshold values determined by top shares of the survey data and leave this for further research.

While the KS test gives a goodness-of-fit to a proposed distribution, it gives no measure of preference. This is of particular note since we not only have two forms for the top-tail, i.e., each form also has a truncated form. Unlike other work, we also do not assume that the top-tail must have a Pareto distribution form, as explained above.

Firstly, a test for truncation is run. This is derived in Aban, Meerschaert and Panorska (2006[28]) to test the null hypothesis that the truncation at the top of the distribution is infinite (i.e., untruncated). As they do not implement a generalized Pareto, we derive the equivalent p-value for the generalized Pareto,[12]

$$p = e^{-r\left(\frac{\sigma+\alpha(\nu-\gamma)}{\sigma}\right)^{-\frac{1}{\alpha}}}$$

We then follow a parsimony principle of using the most simple model form that is retained by the KS and truncation tests. One could consider using a likelihood ratio test of the retained models to select the preferred form. However, due to the computational complexity and greater degree of interpretation, we prefer simplicity over complexity, and rely on the goodness-of-fit tests to exclude poor fits. The Pareto Type 1 is preferred to the generalized Pareto and the truncated form is only used if the untruncated form is rejected by the test, with Figure 8 in Annex A graphically displaying the order of these decisions.

In Table 9 in Annex B, we see from simulated data that using this method, the model is able to correctly identify the distribution it has been sampled from. Results presented show the mean and variance of parameter estimates for data simulated from distributions. The results show that estimates return the parameters used to simulate the distribution and that the goodness-of-fit test can be used to identify the most likely form. The results suggest that generalized Pareto forms can represent Pareto Type I forms but not vice versa, and that truncated models can capture untruncated models but again not vice versa. Therefore, a generalized Pareto form should only be used when the Pareto Type I is not retained, and the truncated form should only be used when the untruncated is not retained. This is because more complexity adds superfluous parameters to the estimation (σ and/or ν) which are unnecessary. By combining the goodness-of-fit tests and the truncation tests, models can be identified that capture characteristics of the data.[13]

---

[12] This constructs a value where as $\nu \to \gamma$, $plim(p) = e^{-r_{\cdot}}$ As $r > 0$, this goes towards 0, rejecting the null that the top truncation is infinite. As $\nu \to \infty$, $plim(p) = e^{-r(\infty)^{-\frac{1}{\alpha}}} = e^{-\infty^{\frac{1}{\alpha}}} = e^{0} = 1$, retaining the null of an infinite top truncation.

[13] 1Annex B also contains tables showing how robust the adjusted truncation estimator is to variation in the upper truncation limit. The results show that the Pareto Type 1 estimator performs better at identifying the true parameter than the generalized Pareto estimator. However, the KS Test and Truncation Test p-values are able to identify the existence of truncation points and the adjusted truncation estimator outperforms the unadjusted estimator in the presence of truncation. As discussed in Vermeulen (2018[10]), the MLE shows some signs of bias. However, as the method described in that paper cannot be applied here, and the Hill Estimator is found to be unreliable, the MLE is applied and the question of the bias is left for future research.

Comparing the performance to Vermeulen (2018[10]), we see that it performs similarly to the (pseudo-) maximum likelihood estimates with oversampling. Given the lack of rich lists and the non-separability of parameters, the preferred Ordinary Least Squares options are not possible. The importance of the goodness-of-fit tests are especially evident when looking at the truncated distributions, where the difference between truncated and untruncated shape parameter estimates are significant. This demonstrates the impact of this method that, to our knowledge, has not been explored elsewhere in the literature.

3.3. Adjusting the top-tail

Having estimated the shape of the top-tail of distribution, the underlying data has to be corrected to reflect these results. While other literature previously has applied the distribution to the total value of the top-tail share to capture the proportions within the tail (see section 2), the centralized approach as developed by the OECD (2022, forthcoming[15]) aims to recover micro statistics from the adjustment so that breakdowns by gender, labour status, and other socio-demographic measures are still possible.

There are three approaches considered in this paper: (i) adjust weights, (ii) impute synthetic households, and (iii) adjust values. Ultimately, this paper combines the imputation of synthetic households and the adjust values approach, creating a fourth method. This latter method imputes additional households using the synthetic household method and attributes them to existing observations using the adjust values approach. This is chosen as it retains information from the original sample while also exploring the missing region of the distribution to gain a micro statistics dataset that can be used to calculate inequality measures.

3.3.1. Adjust Weights Approach

From the estimated top-tail, we see that survey data do not exactly match the densities one would expect from this distribution. In the case of a truncated top-tail, this is especially evident, as individuals are missing at the top end of the distribution. Cantarella, Neri and Ranalli (2021[33]) follow methods from Deville and Särndal (1992[34]) and Särndal (2007[35]), using a calibration approach to adjust the weights of observations to be consistent with the estimated Pareto distribution top-tail while also retaining aggregate demographic shares of survey respondents within the tail.

The adjust weights approach is implemented using an iterative method. At each observation, the weight of the household is scaled by the ratio of the estimated CDF for the reported value to the ECDF from the data, given by the sum of weights reporting this value, correcting the original weight w_i to a corrected weight w_i',

$$w_i' = w_i \frac{F(x_i)}{\hat{F}(x_i)}$$

with $F(x_i)$ being the estimated CDF for the top-tail, and $\hat{F}(x_i)$ being the ECDF from the data, with $\hat{F}(x_i) = \frac{\sum_{j=1}^{n} w_j (x_j \leq x_i)}{\sum_{j=1}^{n} w_j}$. This approach fits the observation to the estimated top-tail. However, key aggregates are not necessarily retained, such as the gender ratio or unemployment rate. The weights are therefore adjusted again, using an optimisation function to minimise deviations subject to the constraint that totals for variables in y are retained (so that $\sum_{i=1}^{n} w_i y_i = t(y)$),

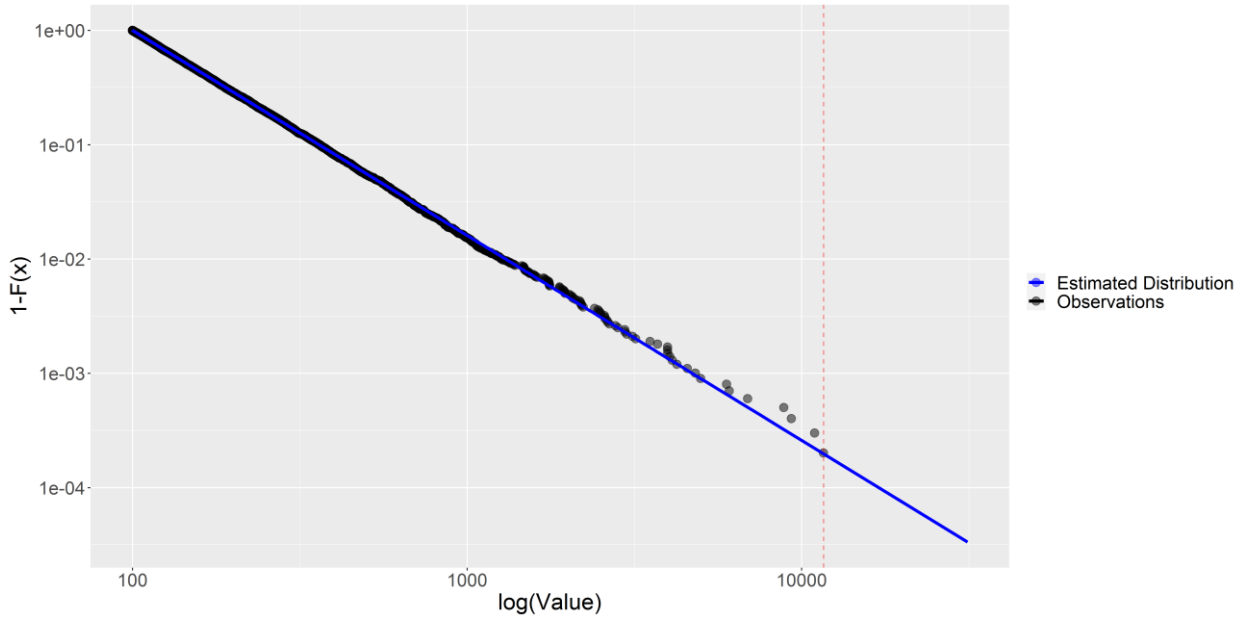
$$\min_{w_i''} \sum_{i=1}^{n} \frac{(w_i'' - w_i')^2}{w_i'} \quad s.t. \quad \sum_{i=1}^{n} w_i'' y_i = t(y)$$

where w_i'' is the post-adjustment weight and $y_i = (y_{i,1}, \dots, y_{i,k})$ is a vector of k demographic variables for household i. This is iterated over until the weights converge, matching the estimated tail and maintaining the demographic totals.

This approach retains a granular dataset, adjusting weights to match the top-tail, and can be seen as calibrating the complex survey weights to an additional dimension. However, it has a number of limiting factors despite its widespread use. In the absence of external data or a truncation point, adjustments are very minor as the pseudo-maximum likelihood estimator has already minimised the residuals in the sample. The approach also cannot be easily extended to use for multiple items as it would have to find weights that satisfy multiple top-tail specifications across a range of items for sets of households that partially overlap while also retaining demographic totals.

However, the main drawback is that it ultimately does not address differential non-response at the very top of the distribution. The approach relies on adjusting weights of survey respondents to appear as though they are from the estimated distribution. However, as seen in Vermeulen (2018[10]), the maximum value for items in survey data often fall far below values found in other sources. This is shown in Figure 4.

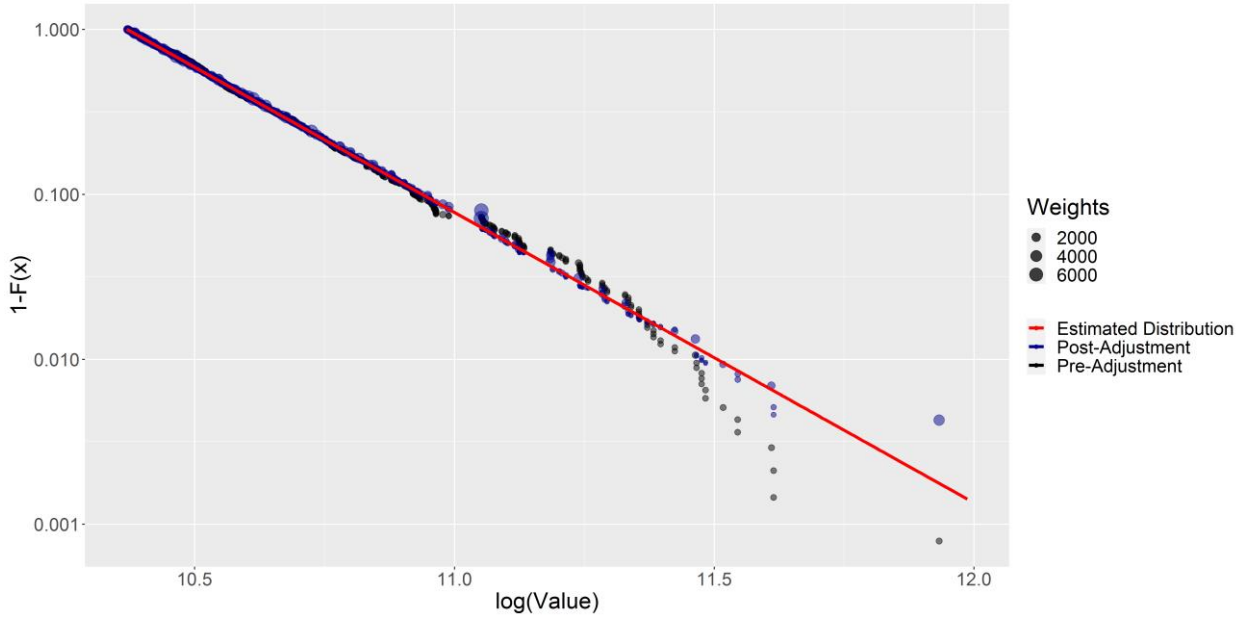
Figure 4. Simulated top-tail data from Pareto Type 1 distribution



Considering the plot of the top-tail in Figure 4, the distribution can be split into two sections either side of the maximum value in the observations, v: observed, $x \leq v$, and unobserved, $x > v$. This follows from Assumption III that the distribution maximum exceeds the data maximum. The adjust weights method operates exclusively in the observed section of the distribution, as by definition there are no observations with $x > v$ to be adjusted.

Figure 5 shows the effect of the adjust weights method in correcting the top-tail for synthetic data from Alfons and Templ (2013[36]). This approach, while matching the top-tail, only reweights observations in the observed sector of the distribution and cannot affect the 'missing' rich, despite the estimated distribution suggesting the existence of missing rich households. As a result, despite changing the aggregate totals of the top-tail, the maximum value remains the same, as does the inequality between observations. This also suggests perfect equality amongst $w_{\max(x_i)}$ households at the top of the distribution.

Figure 5. Adjust weights method to correct top-tail on Alfons and Templ (2013[36]) synthetic data



Furthermore, the households also retain the same structure in reported income (or wealth) items, resulting in a counter-intuitive case where aggregates may show that dividend earnings are highly concentrated in rich households, as suggested by external data, but that this cannot necessarily be reflected in the underlying data, as the values cannot be adjusted. Instead, households with large dividend incomes are given higher weights, which increases the total income but does not affect the ordering of observations by value. This can be seen as problematic as while it can be used to match the distribution, it does so by introducing more households, rather than richer households. Therefore, the relative importance of items reported by each household does not change, but the adjustment in weights will increase the importance of households with desired income sources in the sample.

This can be potentially problematic. For example, as property income received usually has low coverage, the adjust weights approach will increase the weight of households with high proportions of their income received from property income, as it can be used to change the density of property income while not affecting other income items. However, this does not guarantee these households have large incomes, and so may be ranked lower in the income distribution. While this would result in a Pareto top-tail being fitted and an increase in the coverage, it does not necessarily result in the missing rich being represented by similarly rich households in the survey, especially for low coverage items where the number of households available to be adjusted may be limited. There is no guarantee the algorithm will select rich households, but rather those which allow it to make the smallest changes to the weights as possible.

While the adjust weights approach retains a granular dataset, it is not the favoured approach here as it does not propose a solution to capture unobserved households. As the household has a single weight, the adjustment would increase the weight of the households in all components of income. Applying this adjustment over multiple items is likely to complicate the solution, as all the optimisation problems would have to be jointly solved.

3.3.2. Synthetic Households Approach

Having estimated the top-tail, an alternative approach to correct for the rich households not captured in the survey results would be to sample them from the estimated top-tail. This approach relies more heavily

on the complex design weights by assuming the observed distribution for $\gamma \leq x \leq \nu$ is correct, and adjusts the top-tail by adding simulated households beyond this point.

Engel et al. (2022, forthcoming[37]) estimate the top-tail of the wealth distribution using the methodology of Vermeulen (2018[10]), including external data from Forbes rich lists to correct estimates. With this estimation, they can estimate a corrected number of households in the top-tail, including the missing households. Firstly, they assume that $F(x_{max})$ is known from survey data. This means that the distribution below x_{max} is captured, and the total number of households in $F(x_{max})$ can be calculated by the sum of weights in the top-tail , i.e., the sum of weights for households with $x \in [\gamma, x_{max}]$. Here, we assume that no additional households have to be captured from the observed proportion of the distribution, and that the observations are independently identically distributed (i.i.d), although the weights are not.[14] This acknowledges that the weights are designed to capture these households, but that the households in the unobserved section are not represented. The total population of the top-tail, $\hat{m}$, consists of the observed proportion, $\hat{m}_1$, and the unobserved proportion, $\hat{m}_2$, which is estimated from the observed proportion.

$$\hat{m} = \hat{m}_1 + \hat{m}_2$$

$$I_1 = [\gamma, \nu]$$
$$I_2 =]\nu, \infty)$$

$$\hat{m}_1 = \mathbb{E}\left[\sum_{i=1}^{\hat{m}} \mathbb{I}_{x_i \in I_1}\right]$$

$$\hat{m}_1 = \hat{m}\mathbb{E}\left[\mathbb{I}_{x_i \in I_1}\right]$$

$$\hat{m}_1 = \hat{m}P(x_i \in [\gamma, \nu])$$

$$\hat{m}_1 = \hat{m}P(x_i \leq \nu)$$

Given that $\hat{m}_1 = \sum_{i=1}^{\hat{m}} w_i\left(\mathbb{I}_{x_i \in I_1}\right)$, $\hat{m}_1$ is the sum of weights of households identified as being in the top-tail. This means that $\hat{m}$ can be calculated, giving $\hat{m}_2$,

$$\hat{m} = \hat{m}_1[P(x_i \leq \nu)]^{-1}$$

$$\hat{m}_2 = \hat{m} - \hat{m}_1$$

This gives the number of households that have to be drawn from the distribution above ν. However, while the aggregate population is now known, the values of specific synthetic households need to be determined. Since for the untruncated Pareto distribution, $F(\infty) - F(\gamma) = 1$, this can be used to calculate the populations in each range of the distribution, using the results above,

$$\hat{m}[F(\infty) - F(\gamma)] = \hat{m}$$
$$\hat{m}[F(\nu) - F(\gamma)] = \hat{m}_1$$
$$\hat{m}[F(\infty) - F(\nu)] = \hat{m}_2$$

Therefore, to find the value for the first synthetic household, we need to find x_b such that,

$$\hat{m}[F(x_b) - F(\nu)] = 1$$

This is dependent on the form identified for the top-tail. Solving for x_b,

$$x_b = F^{-1}\left(\frac{1}{\hat{m}} + F(\nu)\right)$$

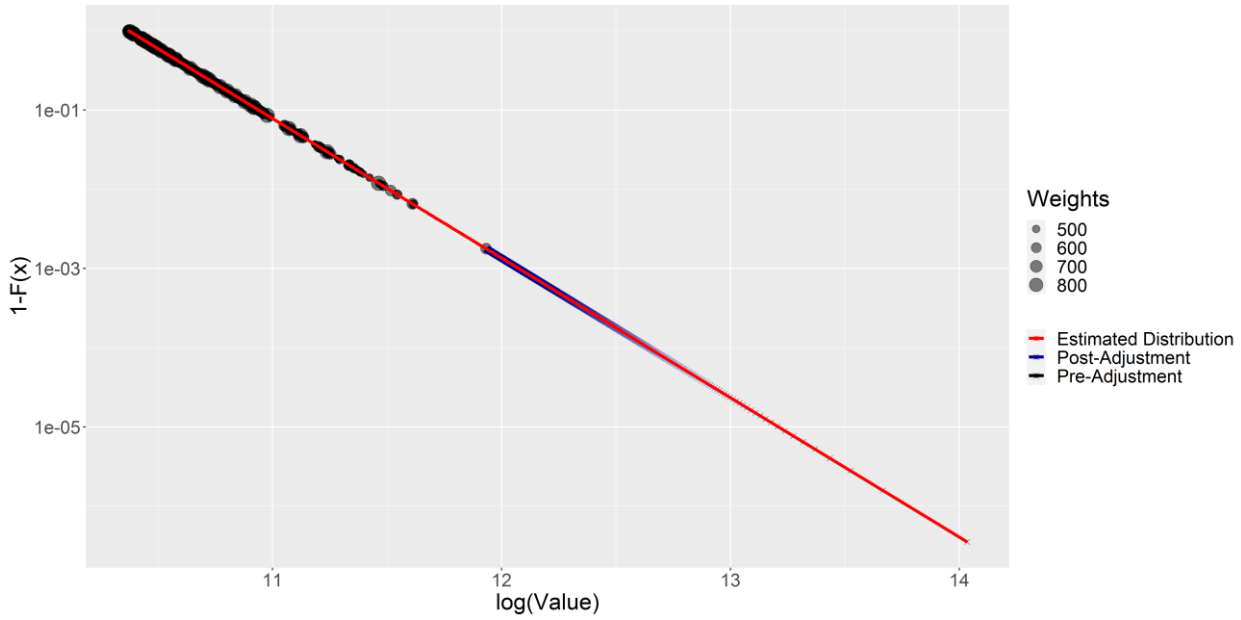
This gives x_b, which then becomes the new maximum, and can be used to find x_{b+1},

[14] The impact of complex survey design is left to future research, as per Engel et al. (2022, forthcoming[37]).

$$x_{b+1} = F^{-1}\left(\frac{1}{\hat{m}} + F(x_b)\right)$$

This is continued $x_b, x_{b+1}, \dots, x_{\hat{m}_2-1}$ to get values from the unobserved portion of the top-tail. This approach has the clear advantage that it addresses the missing rich, exploring the entirety of the proposed tail. Figure 6 shows how the original data, covering the observed area, remains untouched, while the unobserved area is now populated with draws from the estimated distribution. This generates richer households and does not rely on existing surveyed households to represent the missing rich households.

Figure 6. Sythnetic household method to correct top-tail on Alfons and Templ (2013[36]) synthetic data



However, the approach requires additional information to complete. As households are only drawn with a value, Engel et al. (2022, forthcoming[37]) have to impose portfolios and demographics on their synthetic households or leave them undefined. This therefore requires external research to complete the missing information or it results in incomplete granular data. Furthermore, this poses an additional problem when applied item-by-item, as single households would be drawn for each item. Depending on the proportion of households owning these items, the estimated population would differ and could not be linked, since appearing in a top-tail for one item does not necessitate appearing in another. Therefore, while better in line with the motivation than the adjust weight approach, the synthetic household approach does not provide a granular dataset without the availability of external data or explicit assumptions, or can only be conducted at an aggregated level.

3.3.3. Adjust Value Approach

The adjust value approach aims to resolve the problem of providing a granular dataset and exploring the unobserved area of the Pareto distribution, while also being implementable on an item-by-item basis. Törmälehto (2019[24]) estimates the top-tail Pareto distribution and draws $\sum_{i=1}^{n} w_i \mathbb{I}(x_i \geq \gamma)$ values from it, representing the total number of households in the top-tail of the survey. 250 sets are drawn and ordered,

with the value at each rank being averaged across the draws to give a new value sampled from the estimated distribution,

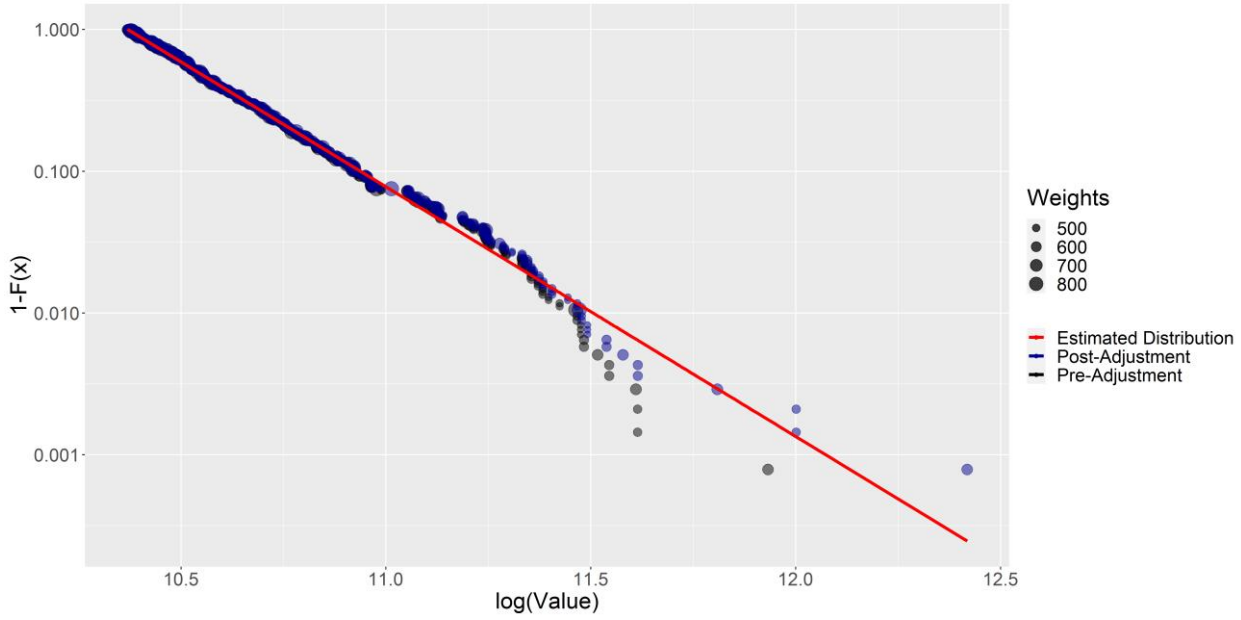
$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,250} \\ \vdots & \ddots & \vdots \\ x_{\sum_{i=1}^{n} w_i \mathbb{I}(x_i \geq \gamma),1} & \cdots & x_{\sum_{i=1}^{n} w_i \mathbb{I}(x_i \geq \gamma),250} \end{bmatrix} \Rightarrow \begin{bmatrix} \frac{\sum_{k=1}^{250} x_{1,k}}{250} \\ \vdots \\ \frac{\sum_{k=1}^{250} x_{\sum_{i=1}^{n} w_i \mathbb{I}(x_i \geq \gamma),k}}{250} \end{bmatrix} = A$$

The new values are then allocated to households in the top-tail to give them corrected values. To do this, the households are ordered by their reported value and have w_i draws allocated to them, based on their rank. If $x_i > x_{i+1}$, $w_0 = 0$ and if the vector of new draws is A, the new value is given,

$$x_i' = \sum_{l=\sum_j^{i-1} w_j}^{\sum_j^{i} w_j} a_l$$

This approach samples items' values from the estimated distribution to correct the top-tail. In doing so, it moves density from the observed region to the unobserved area, as shown in Figure 7. This provides observations with demographic information, retaining the granular dataset. As the items are independent of each other, this can be done on an item-by-item basis, with a distribution estimated for each item.

Figure 7. Adjust value method to correct top-tail on Alfons and Templ (2013[36]) synthetic data



The main drawback of this approach is that it replaces reported data in the survey with values simulated from the distribution, removing actual information collected from the survey. As with the adjust weights approach, this means that the richest households in the survey are used to represent missing rich households, assuming that the demographic profiles are similar. However, unlike the adjust weights approach, it does not necessitate the ranking being maintained at an aggregate level, only at item level, as the top-tails for each item are likely to vary. Moreover, it leans more heavily on the complex survey methods as the weights are unaffected.

3.3.4. Modified Survey Values Approach

In this paper, a new approach is used, using the sampling from the synthetic household method and allocating the values according to the adjust value approach. This approach was selected as it utilises the strengths of each method. The synthetic household approach does not replace the survey data and adds additional observations in the unobserved region. The adjust value provides a granular dataset that doesn't affect the complex survey weights and can therefore be applied on an item-by-item basis.

The synthetic household approach produces additional observations, so that the matrix A can be said to be composed of two parts: the original data values, A_1, and the synthetic values, A_2. Each of the tail observations is then allocated a share of these values based on $s_i(x_i) = \frac{w_i(x_i)}{\sum_{i=1}^{n} w_i \mathbb{I}(x_i \geq \gamma)}$ where $s_0(x_0) = 0$, with k rows in matrix A and $x_i > x_{i+1}$,

$$A = \begin{bmatrix} A_2 \\ A_1 \end{bmatrix}, \quad q_i = k\left(\sum_{j=1}^{i} s_j - s_{j-1}\right)$$

$$x_i' = \sum_{l=q_{i-1}}^{q_i} a_l$$

For example, consider a simplified top-tail consisting of two households: Household A with weight $w_A = 2$ and item value x_A, and Household B with weight $w_B = 2$ and item value $x_B < x_A$. Therefore, the top-tail has a total weight of 4, with each household holding 50% share of the top-tail, with Household A ranked top of the distribution. Accounting for the weights, the top-tail is composed of values $\{x_A, x_A, x_B, x_B\}$. Following the Pareto adjustment, additional synthetic values x_1 and x_2 are drawn, with $x_1 > x_2 > x_A > x_B$, so that the top-tail is now comprised of values $\{x_1, x_2, x_A, x_A, x_B, x_B\}$. The household values are recalculated using the average of the values in their share of the distribution. Household A, which represents the top 50% of values, retains the weight $w_A = 2$, but has its value adjusted to using the top 50% of values in the adjusted tail: $x_A' = \frac{x_1+x_2+x_A}{3} > \frac{x_A+x_A}{2} = x_A$. Household B has its value adjusted, $x_B' = \frac{x_A+x_B+x_B}{3} > \frac{x_B+x_B}{2} = x_B$.

The approach is therefore very similar to the adjust value approach, with the main exception that it does not draw simulated values across the entire tail, only drawing sample values from the unobserved region. These are then allocated to the existing observations based on their ordering, using the same averaging approach seen in Törmälehto (2019[24]).This approach therefore has the implicit assumption that the ranked order of the households demographic information in the top-tail at item level is representative of the population distribution, but that the value reported is not. It could be argued that if demographic measures are correlated to income that the missing rich households may have non-random demographic differences to those households that respond to the data. However, the degree to which this may affect the data is unknown. Furthermore, as the complex survey weights are designed to be representative of the population, making such demographic adjustments would require adjustments elsewhere to retain the calibrated totals used in the design. Therefore, no adjustment is made to the demographics of the top-tail households.

3.4. Proportional Allocation

While the adjustment of the top-tail reduces the gap between micro statistics and national accounts, these are still unlikely to match as there may still be other causes for differences (e.g., measurement error). In this paper, the remaining gaps are closed using a proportional allocation approach. This scales the post-Pareto adjustment values across the distribution by the ratio of national accounts to the (adjusted) survey total:

$$c = \frac{NA_z}{\sum_{i=1}^{n} w_i x_{i,z}}$$

where $z \in Z$ is the item that has been linked, and $x_{i,z}$ is the value reported for item z by household i, who has weight w_i. The proportional allocation method applied a constant scaler to the value of item z across all households, as there is no information on how to allocate the gap (e.g., where measurement errors may more likely occur).

The proportional allocation is inversely related to the post-Pareto adjustment coverage ratio of item z, so that items with lower coverage ratios are scaled more heavily. This means that the proportional allocation does not affect the inequality of the distribution of item z across households, but may change the inequality of aggregated items. This is because the item z will most likely have a different share in the survey results than in the national accounts,

$$\frac{NA_z}{NA_{\sum_{z \in Z} z}} \not\equiv \frac{\sum_{i=1}^{n} w_i x_{i,z}}{\sum_{z \in Z} \sum_{i=1}^{n} w_i x_{i,z}}$$

Therefore, when the proportional allocation is applied to each item, the share of the aggregate held in that item is likely to change. If items with higher inequality have lower (higher) coverage ratios, the proportional allocation will increase (decrease) inequality in the aggregate measure.

4. Results

Using the methodology outlined above, the top-tail adjustment can be applied to micro data sets to correct for possibly missing rich households in constructing DNAs. For the results presented in this section, we follow the linkage methods used in the OECD centralized approach (see OECD (2022, forthcoming[15])). Micro data as obtained from the Luxembourg Income Study (LIS) database are linked to national accounts totals from the OECD databases to construct measures of primary income as defined by the 2008 SNA (i.e., the sum of labour and capital income as received by households, before income and wealth taxation and other forms of redistribution), for Japan (2013), the Netherlands (2013), and the United States (2018).[15]

Using this income measure, shares for the bottom 50%, middle 40%, top 10%, and top 1% are constructed, following the breakdowns as used in Alvaredo et al. (2020[5]). These are compared to results obtained (i) when only applying proportional allocation at the level of aggregated primary income, (ii) when applying proportional allocation at the level of the individual items, to (iii) results where the full micro-macro statistics gap is allocated to households in the top-tail, and to (iv) comparable estimates from other sources.[16]

As discussed in the methodology, the approach selected can be applied on an item-by-item basis to the components of primary income (B5 in the 2008 SNA). This is the sum of operating surplus (B2), mixed income (B3), compensation of employees (D1R), and net property income (D4N) as received by households, with Table 1 showing the complete sub-elements of these items as used in the DNA work. Where these are available, micro statistics are linked to national accounts in accordance with the OECD centralized approach (OECD, 2022, forthcoming[15]).

Table 1. Primary Income Components

B5	Balance of primary incomes
B2R_B3R	Operating surplus and Mixed income
B2R	Operating surplus
B3R	Mixed income
D1R	**Compensation of employees**
D4N	Net property income received / Net property income
D4R	**Property income received**
D4P	**Property income paid**

As discussed in the methodology section, it is common in the literature to determine the threshold of the Pareto distribution γ, by replacing observations in the top 10% of the survey, as this is where the response rates of households tends to fall. However, here we consider thresholds for the top 10%, top 5%, and top 1% for each item. This is because, as shown by Törmälehto (2019[24]), the rate of response can impact the fit.

[15] These represent the most recent years published on LIS for each of the countries. We use Japan as an example of a country where the centralized approach is currently being implemented by the OECD. The United States and the Netherlands have both conducted independent work on constructing distributional national accounts and are presented so that results can be compared.

[16] While primary income of the household sector is used here, other organisations use different measure of income. WID focuses on pre-tax national income which comprises of net primary income of household and NPISH, non-financial corporations, financial corporations, and general government, while Povcal focuses on net post-tax disposable income breakdowns.

Here, we use the goodness-of-fit test described in Section 3.2 to determine whether a Pareto distribution is appropriate, assuming that a) the top-tail is likely to be represented by a Pareto distribution, and b) the form is likely to be truncated. We therefore use different levels of significance for each test, rejecting the null hypothesis of the KS test that the sample is drawn from the candidate distribution at a 1% level of significance, and rejecting the null hypothesis of the truncation test that the form is untruncated at a 10% level of significance. The lower level of significance is used for the KS test to make retaining the null hypothesis more likely, as the top-tail cannot be adjusted if a distribution is not retained. As Davis and Shorrocks (2000[16]) find that the top-tail distribution is well approximated by a Pareto distribution, we choose the level of significance to favour the candidate distributions proposed, but implement a test that has the power to reject all the forms if strongly rejected by the data.

Likewise, a higher level of significance is used for the truncation test as evidence suggests that differential non-response means that the richest households are likely to be missing from the survey. By using the 10% level of significance, the null hypothesis is more likely to be rejected and the alternative truncated form used to correct the parameter estimate, but does not impose a truncation point unless supported by the data.

Table 2. Estimated Pareto distribution parameters, USA 2018

Items	Distribution	α	γ	σ	ν	KS p-value	Truncation p-value
Top-Tail Threshold: 10%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	1.430	3,012.750			0.000	
	Generalized Pareto	0.484	3,012.750	21,703.260		0.000	
	Truncated Pareto Type 1	1.367	16,951.120		1,593,997.000	0.000	0.000
	Truncated generalized Pareto	19.641	3,012.750	478.374	1,593,997.000	0.000	0.998
D1R Compensation of employees							
	Pareto Type 1	2.509	155,500.000			0.000	
	Generalized Pareto	0.394	155,500.000	67,276.000		0.000	
	Truncated Pareto Type 1	2.471	159,514.200		2,247,998.000	0.000	0.000
	Truncated generalized Pareto	10.503	155,500.000	4.850	2,247,998.000	0.000	1.000
D4R Property income received							
	Pareto Type 1	1.182	8,700.000			0.000	
	Generalized Pareto	0.552	8,700.000	13,993.240		0.008	
	Truncated Pareto Type 1	1.111	10,734.320		814,999.000	0.000	0.000
	Truncated generalized Pareto	25.472	8,700.000	112.572	814,999.000	0.000	0.998
D4P Property income paid							
	Pareto Type 1	1E+17	8,121.090			0.000	
	Generalized Pareto	0.100	8,121.090	0.100		0.000	
	Truncated Pareto Type 1	1.750	8,121.090		8,121.090	0.000	0.000
	Truncated generalized Pareto	3.000	8,121.090	3.000	8,121.090	0.000	0.000
Top-Tail Threshold: 5%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	1.430	21,000.000			0.000	
	Generalized Pareto	0.394	21,000.000	33,753.390		0.000	
	Truncated Pareto Type 1	1.367	28,108.960		1,593,997.000	0.000	0.000

	Truncated generalized Pareto	19.480	21,000.000	472.910	1,593,997.000	0.000	1.000
D1R Compensation of employees							
	Pareto Type 1	2.466	210,000.000			0.000	
	Generalized Pareto	0.486	210,000.000	78,125.060		0.000	
	Truncated Pareto Type 1	2.418	209,729.600		2,247,998.000	0.000	0.000
	Truncated generalized Pareto	9.974	210,000.000	4.661	2,247,998.000	0.000	1.000
D4R Property income received							
	Pareto Type 1	1.236	20,401.000			0.000	
	Generalized Pareto	***0.449***	***20,401.000***	***23,133.180***		***0.015***	
	Truncated Pareto Type 1	1.159	20,349.770		814,999.000	0.001	0.000
	Truncated generalized Pareto	22.721	20,401.000	168.360	814,999.000	0.000	1.000
D4P Property income paid							
	Pareto Type 1	1E+17	8,121.090			0.000	
	Generalized Pareto	0.100	8,121.090	0.100		0.000	
	Truncated Pareto Type 1	1.750	8,121.090		8,121.090	0.000	0.000
	Truncated generalized Pareto	3.000	8,121.090	3.000	8,121.090	0.000	0.000
Top-Tail Threshold: 1%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	1.755	95,000.000			0.000	
	Generalized Pareto	0.498	95,000.000	60,026.800		0.000	
	Truncated Pareto Type 1	1.599	92,017.090		1,593,997.000	0.000	0.006
	Truncated generalized Pareto	23.853	95,000.000	14.341	1,593,997.000	0.000	1.000
D1R Compensation of employees							
	Pareto Type 1	2.037	400,000.000			0.000	
	Generalized Pareto	16.083	400,000.000	0.100		0.000	
	Truncated Pareto Type 1	***1.703***	***381,332.600***		***2,247,998.000***	***0.076***	***0.000***
	Truncated generalized Pareto	40.162	400,000.000	469.540	2,247,998.000	0.000	0.000
D4R Property income received							
	Pareto Type 1	***2.090***	***83,198.000***			***0.013***	
	Generalized Pareto	*0.373*	*83,198.000*	*42,152.980*		*0.031*	
	Truncated Pareto Type 1	1.924	78,319.890		814,999.000	0.000	0.007
	Truncated generalized Pareto	29.480	83,198.000	211.240	814,999.000	0.000	1.000
D4P Property income paid							
	Pareto Type 1	1E+17	8,121.090			0.000	
	Generalized Pareto	0.100	8,121.090	0.100		0.000	
	Truncated Pareto Type 1	1.750	8,121.090		8,121.090	0.000	0.000
	Truncated generalized Pareto	3.000	8,121.090	3.000	8,121.090	0.000	0.000

Table 2 shows how the parameter estimates and preferred forms of the top-tail distribution vary depending on the threshold selected, focusing on results for the United States. For each item, the Pareto distribution forms outlined in Section 3.1 are estimated for three thresholds: top 10%, top 5% and top 1%. The estimated parameters for each specification are reported, presenting the shape and threshold parameters as well as the scale and upper truncation parameters where applicable. Furthermore, the p-values for the KS test to show if the null hypothesis that the data fit the estimated form and for the truncation test to show if the null hypothesis that the form is untruncated are retained. The use of the goodness-of-fit test means

that while the true distribution may not be one of the selected forms, it can be tested whether the forms may still provide a statistically good fit to the data conditional on the thresholds. Table 3 shows the forms selected, where the form with the lowest threshold that is not rejected in the KS test is retained.

Table 3. Selected Pareto distribution parameters, USA 2018

Items	Distribution	α	γ	σ	ν	KS p-value	Truncation p-value
B2G_B3GR Operating surplus and Mixed income	NA	NA	NA	NA	NA		
D1R Compensation of employees	Truncated Pareto Type 1	1.703	381,332.600		2,247,998.000	0.076	0.000
D4R Property income received	Generalized Pareto	0.449	20,401.000	23,133.180		0.015	
D4P Property income paid	NA	NA	NA	NA	NA		

The estimated results find that no item of primary income retains a Pareto distribution at the top 10% threshold, as the KS test rejects the null hypothesis in each case. Using the 5% threshold, a generalized Pareto form is retained for property income received. Using the top 1% threshold, forms of the Pareto distributions are retained for both compensation of employees and property income received, with compensation of employees retaining a truncated Pareto Type I, and property income received retaining both the untruncated Pareto Type I and the generalized Pareto. Pareto distributions for operating surplus and mixed income and for property income paid are rejected by the KS test at all thresholds. For property income received, a generalized Pareto is retained for the top 5% share, while both an untruncated Pareto Type I and generalized Pareto are retained for the top 1% share. As shown in Figure 3, the untruncated Pareto Type I is characterised by a constant gradient in the Zipf plot, while the generalized Pareto is characterised by changes in the gradient initially, creating curvature, before becoming linear in the Zipf plot. Examining the characteristics of the distribution in Figure 3, these results suggest that linearity is seen in the top 1% share, but not the next top 4%, with the curvature having less impact at the top of the distribution.

Introducing the ability to select and reject the Pareto distribution in the top-tail creates a problem for selecting the threshold for data selection. Not only is the true form of the tail unknown, but as the proposed forms are approximations for this unknown form, they are dependent on the threshold selected. One could therefore either assume that the top-tail follows a specific form, such as a Pareto Type I, and find the threshold that best fits this form, or assume that the top-tail begins at a specific threshold, and find the form that best fits the data. In this paper, the lowest threshold that supports a proposed functional form is used as this estimates the model using the most possible data while not imposing forms on the data.

In Table 4, we report the shares using the three candidate thresholds for the OECD method, including the top-tail Pareto adjustment, as opposed to other approaches in the literature. Other approaches include allocating assumed shares of the gap to the top 10% of rich households, presented in columns 4 to 6, and a standard item-by-item proportional allocation without adjustment for the top-tail, presented in columns 7 and 8.

Table 4. Inequality statistics for primary income for the United States, 2018

USA	OECD[6]	Assumed Top-Tail Shares[1]	Proportional Allocation

2018								
Shares	Top 10%	Top 5%	Top 1%	50% gap[5]	75% gap[5]	100% gap[5]	Component[6]	Aggregate[5]
Bottom 50%	8.38	8.35	7.93	7.74	6.07	4.39	8.38	8.36
Middle 40%	43.49	43.32	41.87	47.12	36.92	26.72	43.49	50.89
Top 10%	48.13	48.33	50.20	45.14[2]	57.02[2]	68.89	48.13	40.74
Top 1%	*15.13*	*15.55*	*17.69*	*12.52[2]*	*15.80[2]*	*19.09[2]*	*15.13*	*10.89*
Gini	0.6768	0.6782	0.6906				0.6768	

[1] Törmähleto (2017[38]) and Lakner and Milanovic (2015[27]) calculate top-tail adjustments assuming a Pareto Type 1 distribution, utilising $N_{0.1} = NE[X|X_0] = N\frac{\alpha}{\alpha-1}X_0$ to calculate the sum value of the tail and the proportions within the tail. This is therefore calculated assuming that the total share of the top 10% is calculated as the sum of survey data and the proportion of the gap, with proportional allocation applied to the rest of the distribution.

[2] Top-tail values are calculated from the moments of the distribution, using identities from the Pareto Type 1 Distribution, and combined with survey data for the rest of the distribution. To ensure that the shares sum to 100%, the Top 10% share is calculated as the remainder after the Bottom 90% is removed. This number may differ from the theoretical share given using the formula above, which is used to calculate the Top 1% share.

[3] Results from http://iresearch.worldbank.org/PovcalNet/home.aspx, using Post-Tax Disposable Income for closest survey year.

[4] Results from https://wid.world/, using Tax Unit Fiscal Income.

[5] Results shown for Pre-Tax equivalised primary income (B5).

[6] Results shown for Pre-Tax equivalised primary income from sum of linked components.

[7] Returns for Adjusted Gross Income (AGI) share for closest year reported by IRS.

[8] Returns for 2018 Taxable Income (43) reported by IRS, table 3.6.

Comparing the results, we see that the Pareto top-tail adjustments increase inequality when compared to the standard proportional allocation adjustments. The results vary dependent on the threshold selection, with adjustments to the top 10% share producing no change as the Pareto top-tails are rejected for all items, and adjustments to the top 1% share producing the largest changes, as this threshold adjusts both compensation of employees (D1R) and property income received (D4R). Please note that applying proportional allocation to each component individually creates larger inequality than applying it to aggregated primary income, which is due to the variation in the coverage ratios of the components, where larger gaps tend to exist for items that show larger inequality.

The assumed top-tail shares approach can be viewed as a generalized version of proportional allocation shown in Section 3.4, where the proportion of the gap allocated to the top-tail can vary from the share in survey data. This method, seen in Lakner and Milanovic (2015[27]), means that researchers can allocate a greater share of the micro-macro gap to the missing rich than would be suggested from the survey data. However, with a lack of external data, motivating changes in how the gap is allocated can be difficult. The assumed top-tail shares approach implemented results in very high concentrations of wealth in the top 10%, as the majority of the existing gap is allocated to these households. In the presence of low coverage rates, this means that the majority of wealth is given to the top 10% by construction. However, the implementation of the Type 1 Pareto form within the tail means that the share of the tail in the top 1% of the distribution remains constant, around 27.7% of all income in the top-tail. In contrast, the Pareto top-tail adjustment provides additional density at the top of the distribution compared to the proportional allocation while only using the micro statistics provided. This means that similar top 1% shares can be achieved as when assuming the top-tail shares, but it does not rely on also increasing the top 10%. However, how thresholds are selected needs to be considered.

In Table 5, the effects of changing the threshold on the income components can be seen. As Pareto distributions are rejected for all items using the top 10% threshold, the shares remain constant. However, by changing the threshold, we see changes in the goodness-of-fit as the data matches a Pareto form estimated using the different threshold value. We see that the share of property income received (D4R) held by the top 10% increases using the 5% threshold, changing from 66.81% to 67.63%, and then to 67.75% at the 1% threshold. This is because the goodness-of-fit test starts to retain a Pareto form,

increasing the density. As the lower threshold value is preferred as it gives more data to the model, and property income received (D4R) retains a Pareto distribution at the 10% threshold, this may suggest a mixed approach is best, using the lowest threshold that retains a Pareto distribution for each component. This suggests that missing rich households are more prominent in the distribution of some items than others, with more households being included in the top-tail Pareto distribution at lower thresholds.

Table 5. Primary Income Component Inequality, United States, 2018

Country: USA Year: 2018		Income Components				
Method	Share[1]	B2G_B3GR Operating surplus and Mixed income	D1R Compensation of employees	D4R Property income received	D4P Property income paid	B5 Primary Income
Unadjusted	Bottom 50%	1.08	15.88	5.27	51.97	8.38
	Middle 40%	19.83	55.22	27.92	38.00	43.49
	Top 10%	79.08	28.90	66.81	10.03	48.13
	Coverage Ratio	15.58%	78.98%	18.34%	100%	52.50%
Pareto (10%)	Bottom 50%	1.08	15.88	5.27	51.97	8.38
	Middle 40%	19.83	55.22	27.92	38.00	43.49
	Top 10%	79.08	28.90	66.81	10.03	48.13
	Coverage Ratio	15.58%	78.98%	18.34%	100%	52.50%
Pareto (5%)	Bottom 50%	1.09	15.85	5.22	51.99	8.35
	Middle 40%	19.74	55.22	27.15	38.00	43.33
	Top 10%	79.17	28.93	67.63	10.01	48.32
	Coverage Ratio	15.58%	78.98%	18.96%	100%	52.62%
Pareto (1%)	Bottom 50%	0.99	15.28	5.10	51.98	7.94
	Middle 40%	19.15	53.20	27.15	38.01	41.86
	Top 10%	79.85	31.53	67.75	10.01	50.20
	Coverage Ratio	15.58%	82.29%	18.90%	100%	54.79%

[1]Shares are calculated with households ordered by Primary Income (B5)

In Table 6, we calculate the shares using this mixed approach. For comparison, a number of other sources are presented, although concepts of income and the unit level may vary across these (see Zwijnenburg (2019[39]) for an overview of how different concepts and assumptions may affect distributional results). While these concepts would be recoverable from full DNAs, for the purpose of this paper, we only included results for primary income. Although this may not lead to fully comparable results across studies, these results have still been included for completeness, as these streams of literature have motivated much of the work and the indicators still come close to our measures.

With these Pareto top-tail adjustments, we find that the share of primary income held by the top 1% increases by 2.59 percentage points compared to a standard proportional allocation at the level of components, while the top 10% share increases by 2.09 percentage points. This results in the Gini coefficient increasing by 0.0169.[17] This shows that the Pareto top-tail adjustment allocates a greater share of income to the top 1%, where the missing rich are presumed to be, with 35.3% of the income earned by the top 10% being concentrated in the top 1%, as opposed to 31.4% if a proportional allocation is used.

Table 6. Inequality Statistics Mixed Approach, USA 2018

USA	OECD[6]	Assumed Top-Tail Shares[1]	Proportional Allocation	Other Literature

[17] If the adjustment is run on Primary Income reported in LIS (B5), the top 1% share is 10.89%, 7.17 percentage points lower, and the top 10% share is 40.74%, 9.48 percentage points lower.

2018										
Shares		50% gap[5]	75% gap[5]	100% gap[5]	Aggregates[5]	Components[6]	WID[4]	PovcalNet[3]	AGI[7]	Taxable Income[8]
Bottom 50%	7.93	7.74	6.07	4.39	8.36	8.38	13.48	22.48	11.25	6.63
Middle 40%	41.86	47.12	36.92	26.72	50.89	43.49	41.01	47.10	41.01	38.80
Top 10%	50.22	45.14[2]	57.02[2]	68.89	40.74	48.13	45.51	30.42	47.74	54.57
Top 1%	17.72	*12.52*[2]	*15.80*[2]	*19.09*[2]	*10.89*	*15.13*	18.92	*N/A*	*21.04*	*25.22*
Gini	0.6907				0.6552	0.6768	0.58			

[1] Törmähleto (2017[38]) and Lakner and Milanovic (2015[27]) calculate top-tail adjustments assuming a Pareto Type 1 distribution, utilising $N_{0.1} = NE[X|X_0] = N\frac{\alpha}{\alpha-1}X_0$ to calculate the sum value of the tail and the proportions within the tail. This is therefore calculated assuming that the total share of the top 10% is calculated as the sum of survey data and the proportion of the gap, with proportional allocation applied to the rest of the distribution.

[2] Top-tail values are calculated from the moments of the distribution, using identities from the Pareto Type 1 Distribution, and combined with survey data for the rest of the distribution. To ensure that the shares sum to 100%, the Top 10% share is calculated as the remainder after the Bottom 90% is removed. This number may differ from the theoretical share given using the formula above, which is used to calculate the Top 1% share.

[3] Results from http://iresearch.worldbank.org/PovcalNet/home.aspx, using Post-Tax Disposable Income for closest survey year.

[4] Results from https://wid.world/, using Pre-Tax National Income.

[5] Results shown for Pre-Tax equivalised primary income (B5).

[6] Results shown for Pre-Tax equivalised primary income from sum of linked components.

[7] Returns for Adjusted Gross Income (AGI) share for closest year reported by IRS.

[8] Returns for 2018 Taxable Income (43) reported by IRS, table 3.6

Having demonstrated the effects of the Pareto top-tail adjustment on the basis of US data, we now consider the Netherlands, for which official DNA results are compiled by the statistical office, and Japan, who have not compiled recent distributional results and for that reason have been targeted as part of the centralized approach. In both cases, a mixed approach is utilised, with Annex C containing full results for the Netherlands, and Annex D for Japan.

Törmälehto (2019[24]) outlines DNA methods that are applied to disposable income for European countries and that utilise Pareto top-tail adjustments. In addition to the difference concept of income, the approach differs from the method presented here as it only fits a Pareto Type I and estimates the parameter by allocating a proportion of the gap to the top of the distribution, as shown in Section 2. Households are then sampled as per Section 3.3.3. The results find a significant effect of the component proportional allocation on inequality compared to the raw survey data, and that Pareto adjustments to the top-tail further increase inequality.

Using the approach of this paper, Table 7 presents DNA results for the Netherlands. We find that the Gini coefficient is affected very little by the Pareto top-tail adjustment, increasing only 0.0039 compared to the 0.017 change found in Törmälehto (2019[24]).[18] This comparably small change is likely due to differences in the income concept, but may also be influenced by the form of the Pareto top-tail, with the data favouring a generalized Pareto over a Pareto Type I. The Pareto adjustment increases the top 10% share of income by 0.44 percentage points and the top 1% share by 0.59 percentage points compared to the proportional allocation on the components of primary income. These changes result in 28% of income in the top 10% share being held by the top 1%, compared to only 26.6% under proportional allocation.

Compared to the US results shown previously, applying the proportional allocation by component rather than using the primary income from the survey data has a similar impact on the top 1% share for the Netherlands. However, the Pareto adjustment has a much greater effect on inequality when applied to the US data than when applied to the Dutch data. This would suggest that Dutch survey data has better

[18] Törmälehto (2019[24]) reports a pre-adjustment gini coefficient of 26.7 and a post-adjustment gini coefficient of 32.7. Results presented here show that the Pareto adjustment does not cause large changes in the

coverage of the top-tail than the United States, either due to fewer missing rich households in the Dutch data or that the missing rich households in the Netherlands are more similar to those included in the survey than the missing rich in the United States. Comparing the shape parameter estimates, US items tend to have lower values, creating longer tails and thereby containing more missing rich households.

Comparing Table 2 to Table 10 (see Appendix C), we find that fewer income items retain top-tail Pareto distributions in the United States than in the Netherlands. However, comparing the effect on the top 10% shares in Table 5 and Table 12, we see that the effect of the Pareto adjustment on compensation of employees (D1R) for the United States is much larger than effects seen elsewhere. As this item has a much larger value in the National Accounts than property income received (D4R), this effect is likely to be key in the overall size of the Pareto adjustment.

Table 7. Inequality Statistics Mixed Approach, Netherlands 2013

Netherlands 2013		Assumed Top-Tail Shares[1]			Proportional Allocation		Other Literature		
Shares	OECD[6]	50% gap[5]	75% gap[5]	100% gap[5]	Aggregates[5]	Components[6]	WID[4]	PovcalNet[3]	Statistics Netherlands[7]
Bottom 50%	14.69	12.60	11.05	9.51	13.79	14.89	18.00	31.86	6.70
Middle 40%	52.05	50.60	44.38	38.17	55.35	52.30	50.70	47.03	58.50
Top 10%	33.26	36.80	44.56	52.32	30.86	32.82	31.34	21.11	34.80
Top 1%	*9.32*	*7.29*	*8.83*	*10.37*	*6.35*	*8.73*	*6.52*	*4.50*	*8.70*
Gini	0.5761				0.5758	0.5722		0.2810	

[1] Törmähleto (2017[38]) and Lakner and Milanovic (2015[27]) calculate top-tail adjustments assuming a Pareto Type 1 distribution, utilising $N_{0.1} = NE[X|X_0] = N\frac{\alpha}{\alpha-1}X_0$ to calculate the sum value of the tail and the proportions within the tail. This is therefore calculated assuming that the total share of the top 10% is calculated as the sum of survey data and the proportion of the gap, with proportional allocation applied to the rest of the distribution.

[2] Top-tail values are calculated from the moments of the distribution, using identities from the Pareto Type 1 Distribution, and combined with survey data for the rest of the distribution. To ensure that the shares sum to 100%, the Top 10% share is calculated as the remainder after the Bottom 90% is removed. This number may differ from the theoretical share given using the formula above, which is used to calculate the Top 1% share.

[3] Results from http://iresearch.worldbank.org/PovcalNet/home.aspx, using Post-Tax Disposable Income for closest survey year.

[4] Results from https://wid.world/, using Fiscal Income.

[5] Results shown for Pre-Tax equivalised primary income (B5).

[6] Results shown for Pre-Tax equivalised primary income from sum of linked components.

[7] Primary Income (B5) shares, with micro-macro gap unresolved calculated by Statistics Netherlands.

Japan is one of the countries for which no recent DNA results are available and for which results are currently being derived on the basis of a centralized approach, using a method that can be applied on the basis of available micro data and national accounts totals. For Japan, micro data have been obtained from the Luxembourg Income Study, combined with national accounts' data available from the OECD National Accounts database. The results for Japan demonstrate strong effects of the Pareto top-tail adjustment. Applying the proportional allocation to components of primary income increases the top 10% share 0.48 percentage points, and the top 1% share by 0.39 percentage points. In contrast, the Pareto top-tail adjustment increases the top 10% share by a further 1.81 percentage points and the top 1% share by 1.78 percentage points, resulting in a 0.0134 increase in the Gini.

In contrast to the Dutch data, the component proportional allocation only increases the top 1% share 0.39 percentage points for Japan, while the Pareto adjustment increases the value by 1.88. The Pareto top-tail adjustment provides an adjustment for Japan proportionally similar to the United States, whereas the component proportional allocation has a smaller effect on the top 1% share. Table 14 shows that the generalized Pareto is often preferred compared to the Pareto Type I in Japanese data, demonstrating the importance of considering richer distributions when modelling data.

Table 8. Inequality Statistics Mixed Approach, Japan 2013

Japan 2013		Assumed Top-Tail Shares[1]			Proportional Allocation		Other Literature			
Shares	OECD[6]	50% gap[5]	75% gap[5]	100% gap[5]	Aggregates[5]	Components[6]	WID[4]	PovcalNet[3]	ESRI[7]	Taxable Income[8]
Bottom 50%	15.64	14.98	13.22	11.45	16.32	16.12	19.52	28.34	26.44	14.06
Middle 40%	50.92	48.22	42.54	36.85	52.54	52.26	38.13	45.25	48.54	35.79
Top 10%	33.43	36.70	44.16	51.63	31.14	31.62	43.35	26.41	25.01	50.15
Top 1%	*8.57*	*7.68*	*9.24*	*10.80*	*6.40*	*6.79*	*12.46*			*20.35*
Gini	0.5427				0.5253	0.5293	0.5100			

[1] Törmähleto (2017[38]) and Lakner and Milanovic (2015[27]) calculate top-tail adjustments assuming a Pareto Type 1 distribution, utilising $N_{0.1} = NE[X|X_0] = N\frac{\alpha}{\alpha-1}X_0$ to calculate the sum value of the tail and the proportions within the tail. This is therefore calculated assuming that the total share of the top 10% is calculated as the sum of survey data and the proportion of the gap, with proportional allocation applied to the rest of the distribution.

[2] Top-tail values are calculated from the moments of the distribution, using identities from the Pareto Type 1 Distribution, and combined with survey data for the rest of the distribution. To ensure that the shares sum to 100%, the Top 10% share is calculated as the remainder after the Bottom 90% is removed. This number may differ from the theoretical share given using the formula above, which is used to calculate the Top 1% share.

[3] Results from http://iresearch.worldbank.org/PovcalNet/home.aspx, using Post-Tax Disposable Income for closest survey year.

[4] Results from https://wid.world/, using Pre-Tax National Income.

[5] Results shown for Pre-Tax equivalised primary income (B5).

[6]Results shown for Pre-Tax equivalised primary income from sum of linked components.

[7]Retults for 2014 Primary Income (B5) shares reported in (Yamazaki and Sakamaki, 2018[39])

[8]Returns for 2013 Taxable Income (43) reported by National Tax Agency, table 2-4 Self-Assessment Income Tax by Income Type https://www.nta.go.jp/publication/statistics/kokuzeicho/h25/h25.pdf

Comparing the results across these economies, the importance of adjusting components of income individually and applying Pareto top-tail adjustments can be seen, addressing downward bias identified in inequality statistics calculated using survey data. The relative sizes of the effects are not constant across the economies presented in this paper, but show that the methodology presented here is widely applicable and contributes to a better representation of inequality statistics, helping to close part of the micro-macro gap, while also returning a micro statistic dataset that can be further explored.

5.Conclusion

When using Pareto distributions to correct for missing rich households in survey data, this paper demonstrates methods to expand on the widely used Pareto Type I and generalized Pareto forms to adjust for missing households, and how statistical tests can be used in model selection. The results shown in this paper find that when estimating the upper tail of income, consumption and wealth distributions using survey data, missing observations for the top caused by differential non-response can greatly impact parameter estimations of commonly used models. It also shows that while models are often assumed to fit the data, the lack of statistical tests being implemented means that Pareto forms may be imposed even in cases where evidence of such a form is lacking.

In the paper, adjusted forms of common estimators are provided to account for missing observations at the top of the distribution, with tests for truncation points and statistical fit also shown. Having identified Pareto distributions that are estimated on survey data and found to be supported by the observations, an array of sampling methods from the literature are discussed, showing how they can be applied on household-level survey data and the impact of the adjusted forms when compared to the unadjusted models.

These models are then applied to household survey data to construct estimates for household distributional results in line with national accounts totals (DNA) for Japan, the Netherlands and the United States. In the results, we see that proposed estimates that account for truncation in the data are often preferred to the untruncated forms. Moreover, we also see that when treating income items on a component basis, that Pareto tails are not supported for all items, and that the preferred models often have thresholds above the top 10% threshold. The results support an item-by-item approach to applying Pareto tails, rather than applying Pareto tails to aggregate measures of income, such as primary income.

The country estimates of income shares find that the impact of the Pareto adjustment varies over countries relative to the role of applying proportional allocation at item level, when compared to shares calculated from survey data on aggregate income measures. In all countries, the item-level proportional allocation and Pareto adjustment increase the top 1% share compared to that of the survey data. However, in the United States and the Netherlands, the component proportional allocation accounts for a larger share of the total change than the Pareto adjustment, whereas in Japan, the Pareto-adjustment shows the larger effect.

Comparing the proposed method to the survey data, the largest change in the Gini is found in the United States, which changes 3.5 percent, while the Gini for the Netherlands changes very little. When compared to allocating 75% or 100% of the micro-macro gap to the top-tail, the approach proposed often produces a similar top 1% share but a much lower top 10% share, reflecting the ability of the more complex model to vary the shares, while also supporting these models using statistical tests. Future research on the topic should build on the results presented here, refining the estimation method and considering alternative forms such as Pareto Type IV or Log-Normal distributions.

The findings presented provide evidence for the value of using truncation adjustments to correct estimates of parameters in the Pareto top-tail. The methods presented here can also be used to calculate more complete DNAs in the future to allow comparisons with the wider array of concepts for income, wealth and consumption used in the inequality literature, and improve inequality estimates using these alternative forms of the Pareto tail.

6.References

Aban, I., M. Meerschaert and A. Panorska (2006), "Parameter estimation for the truncated Pareto distribution", *Journal of the American Statistical Association*, Vol. 101/473. [28]

Alfons, A. and M. Templ (2013), "Estimation of social exclusion indicators from complex surveys: The R package laeken", *Journal of Statistical Software*, Vol. 54/15. [36]

Alvaredo, F. et al. (2020), "Distributional National Accounts Guidelines: Methods and Concepts Used in the World Inequality Database", *WID.world*, pp. 1-186. [5]

Angel, S., R. Heuberger and N. Lamei (2018), "Differences Between Household Income from Surveys and Registers and How These Affect the Poverty Headcount: Evidence from the Austrian SILC", *Social Indicators Research*, Vol. 138/2. [12]

Atkinson, A. (2017), "Pareto and the Upper Tail of the Income Distribution in the UK: 1799 to the Present", *Economica*, Vol. 84/334. [21]

Blanchet, T., L. Chancel and A. Gethin (2019), "How Unequal Is Europe? Evidence from Distributional National Accounts, 1980–2017", *WID.world Working Paper Series* April. [25]

Blanchet, T., J. Fournier and T. Piketty (2017), "Generalized Pareto Curves: Theory and Applications", *WID.world Working Paper Series* March. [22]

Blanchet, T. et al. (2018), "Applying Generalized Pareto Curves to Inequality Analysis", *AEA Papers and Proceedings*, Vol. 108. [42]

Burkhauser, R. et al. (2018), "Survey Under-Coverage of Top Incomes and Estimation of Inequality: What is the Role of the UK's SPI Adjustment?", *Fiscal Studies*, Vol. 39/2. [13]

---------- (2018), "Top incomes and inequality in the UK: Reconciling estimates from household survey and tax return data", *Oxford Economic Papers*, Vol. 70/2. [14]

Cantarella, M., A. Neri and M. Ranalli (2021), "Mind the wealth gap: a new allocation method to match micro and macro statistics for household wealth", *Banca D'Italia Occasional Papers*. [33]

Chakraborty, R. and S. Waltl (2018), "Missing the wealthy in the HFCS: Micro problems with macro implications", *ECB Working Paper* 2163. [18]

Clauset, A., C. Shalizi and M. Newman (2009), "Power-law distributions in empirical data", *SIAM Review*, Vol. 51/4. [31]

Davis, J. and A. Shorrocks (2000), *The Distribution of Wealth*, Elsevier. [16]

Deville, J. and C. Sarndal (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, Vol. 87/418. [34]

Engel, J. et al. (2022, forthcoming), "Developing Reconciled Quarterly Distributional Financial Accounts - Insight into Inequality and Wealth Structures", *ECB Working Paper*. [37]

Hill, B. (1975), "A Simple General Approach to Inference About the Tail of a Distribution", *The* [29]

Annals of Statistics, Vol. 3/5.

Jenkins, S. (2017), "Pareto Models, Top Incomes and Recent Trends in UK Income Inequality", *Economica*, Vol. 84, pp. 261–289. [26]

Kennickell, A. and R. Woodburn (1999), "Consistent weight design for the 1989, 1992 and 1995 SCFs, and the distribution of wealth", *Review of Income and Wealth*, Vol. 45/2. [7]

Krieger, A. and D. Pfeffermann (1997), "Testing of Distribution Functions from Complex Sample Surveys", *Journal of Official Statistics*, Vol. 13/2, pp. 123-142. [32]

Kuznets, S. (1955), "Economic Growth and Income Inequality", *The American Economic Review*, Vol. 45/1, pp. 1-28. [1]

Lakner, C. and B. Milanovic (2015), "Global income distribution from the fall of the berlin wall to the great recession", *Revista de Economia Institucional*, Vol. 17/32. [27]

Langousis, A. et al. (2016), "Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database", *Water Resources Research*, Vol. 52/4. [30]

Luxembourg Income Study (LIS) Database (2021), *http://www.lisdatacenter.org*, [40]

OECD (2022, forthcoming), *Distributional Information on Household Income, Consumption and Saving in line with National Accounts Guidelines*. [15]

Pareto, V. (1896), *Cours d'Économie Politique*, Lausanne: F. Rouge. [19]

---------- (1897), "The New Theories of Economics", *Journal of Political Economy*, Vol. 5/4. [20]

Piketty, T. (2001), "Income Inequality in France 1901-98", *CEPR Discussion Papers*, Vol. 2876. [3]

---------- (2001), *Les hauts revenus en France au XXe siècle: Ine´galite´s et redistributions, 1901–1998*, Bernard Grasset. [2]

Piketty, T. (2003), "Income inequality in France, 1901-1998", *Journal of Political Economy*, Vol. 111/5. [4]

Piketty, T. and E. Saez (2003), "Income inequality in the United States, 1913-1998", *Quarterly Journal of Economics*, Vol. 118/1. [23]

Piketty, T., E. Saez and G. Zucman (2018), "Distributional national accounts: Methods and estimates for the United States", *Quarterly Journal of Economics*, Vol. 2, p. 133. [6]

Sabelhaus, J. et al. (2013), "Is the Consumer Expenditure Survey Representative by Income?", *NBER Working Paper*, Vol. 19589. [8]

Särndal, C. (2007), "The calibration approach in survey theory and practice", *Survey Methodology*, Vol. 33/2. [35]

Törmälehto, V. (2017), "High income and affluence: Evidence from the European Union statistics on income and living conditions (EU-SILC)", *Eurostat Statistical Working Papers* February. [38]

---------- (2019), “Reconciliation of EU statistics on income and living conditions (EU-SILC) data with national accounts”, *Eurostat Statistical Working Paper* June. [24]

Vermeulen, P. (2014), “How Fat is the Top Tail of the Wealth Distribution?”, *ECB Working Paper Series* 1692. [17]

---------- (2016), “Estimating the top tail of the wealth distribution”, *American Economic Review*, Vol. 106/5. [9]

---------- (2018), “How Fat is the Top Tail of the Wealth Distribution?”, *Review of Income and Wealth*, Vol. 64/2. [10]

Yamazaki, T. and T. Sakamaki (2018), “Re-estimation of detailed household accounts within the SNA framework”, *ESRI Research Note* 42. [39]

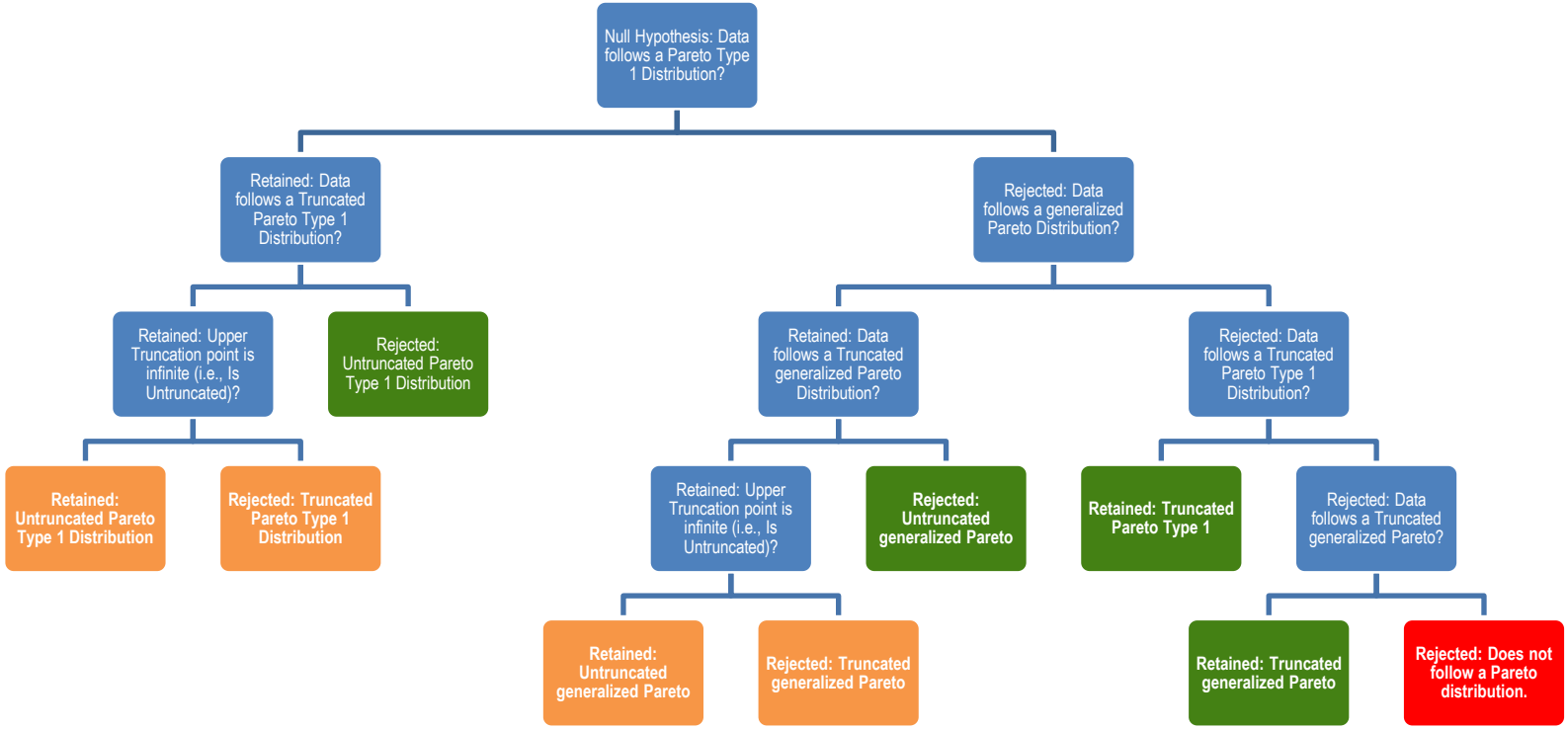
Zwijnenburg, J. (2019), “Unequal Distributions: EG DNA versus DINA Approach”, *AEA Papers and Proceedings*, Vol. 109. [41]

---------- (2021), “Distribution of Household Income, Consumption and Ssavings in line with National Accounts – Results from a 2020 Exercise”, OECD Statistical Working Papers 2021/01. [43]

---------- (2022), “The Use of Distributional National Accounts in Better Capturing the Top Tail of the Distribution”, The Journal of Economic Inequality, Vol. 20/1, pp. 245-254. [11]

Annex A. Model Selection Process

Figure 8. Flow Diagram of selection method for Pareto function



Annex B. Simulated Data Analysis

100 samples of 10,000 observations are drawn from four simulated distributions and used as input values, each given a weight of 1. The distribution is estimated for each sample, with the mean and variance. For truncated forms, the top 15% of observations is removed from the input data.

KS Test and Truncation test columns show the proportion of the 100 samples that rejected the null hypothesis. The null hypothesis for the KS test is that the data follows the distribution being tested for, while rejecting the null hypothesis in favour of the alternative hypothesis means that it is statistically likely to be drawn from a different distribution. The null hypothesis for the Truncation test is that the upper truncation parameter is infinite, and so is not truncated. The alternative hypothesis is that the upper truncation parameter is statistically different from infinite, and that the distribution is truncation. The proportion of tests where the null hypothesis is rejected is shown when using levels of significance of 10%, 5%, and 1%.

Table 9. Estimation of Simulated Distributions

	Form	Parameters				Proportion of KS Test rejecting Null Hypothesis			Proportion of Truncation Test rejecting Null Hypothesis		
		α	γ	σ	ν	10%	5%	1%	10%	5%	1%
Simulated Distribution	Pareto Type 1	1.8000	100.00	--	--	--	--	--	--	--	--
Estimation	*Pareto Type 1*	*1.7984 (0.0005)*	*100.0059 (0.0000)*			*0.02*	*0.01*	*0.00*			
	Generalized Pareto	0.5576 (0.0002)	100.0059 (0.0000)	55.5673 (0.9621)		0.00	0.00	0.00			
	Truncated Pareto Type 1	1.7967 (0.0005)	100.0290 (0.2606)		40,911.9216 (4,485,014,277)	0.05	0.04	0.03	0.05	0.02	0.01
	Truncated Generalized Pareto	2,708.568 (720,556.5)	100.0059 (0.0000)	6,895.5390 (5,353,558.7)	40,911.9216 (4,485,014,277)	1.00	1.00	1.00	0.97	0.97	0.97
Simulated Distribution	Generalized Pareto	1.8000	100.00	1.2000	--	--	--	--	--	--	--
Estimation	Pareto Type 1	0.6818 (0.0016)	100.0001 (0.0000)			1.00	1.00	1.00			
	Generalized Pareto	*1.7989 (0.0007)*	*100.0001 (0.0000)*	*1.1977 (0.0009)*		*0.00*	*0.00*	*0.00*			
	Truncated Pareto Type 1	0.6786 (0.0017)	3.0563 (0.6395)		5.96E9 (1.02E21)	1.00	1.00	1.00	0.00	0.00	0.00
	Truncated Generalized Pareto	1.8021 (0.0007)	100.0001 (0.0000)	1.1966 (0.0009)	5.96E9 (1.02E21)	0.00	0.00	0.00	0.05	0.02	0.00

Simulated Distribution	Truncated Pareto Type 1	1.8000	100.00	--	286.7416	--	--	--	--	--	--
Estimation	Pareto Type 1	2.7502 (0.0009)	100.0088 (0.0000)			1.00	1.00	1.00			
	Generalized Pareto	0.1000 (0.0000)	100.0088 (0.0000)	46.9869 (0.3859)		1.00	1.00	1.00			
	Truncated Pareto Type 1	*1.8005 (0.0015)*	*99.9598 (0.0128)*		*286.7416 (8.8106)*	*0.00*	*0.00*	*0.00*	*1.00*	*1.00*	*1.00*
	Truncated Generalized Pareto	720.215 (5,804.864)	100.0088 (0.0000)	9,672.613 (91,310.96)	286.7416 (8.8106)	1.00	1.00	1.00	1.00	1.00	1.00
Simulated Distribution	Truncated Generalized Pareto	1.8000	100.00	1.2000	119.5802	--	--	--	--	--	--
Estimation	Pareto Type 1	22.5835 (0.5942)	100.0002 (0.0000)			1.00	1.00	1.00			
	Generalized Pareto	0.6246 (0.0004)	100.0002 (0.0000)	1.3585 (0.0010)		1.00	1.00	1.00			
	Truncated Pareto Type 1	18.7183 (0.5714)	96.5852 (0.0705)		119.5802 (0.4565)	1.00	1.00	1.00	1.00	1.00	1.00
	Truncated Generalized Pareto	*1.8038 (0.0103)*	*100.0002 (0.0000)*	*1.1970 (0.0007)*	*119.5802 (0.4565)*	*0.00*	*0.00*	*0.00*	*1.00*	*1.00*	*1.00*

Note: In Proportion of KS Test Null Rejected, retaining the null means the form fits the data (i.e., the closer to 0, the more often the form is supported). In Proportion of Truncation Test Null Rejected, retaining the null means the form is untruncated (i.e., the closer to 1, the more often an upper truncation limited is found).

Data is simulated from an untruncated form of the Pareto distribution, then has a range of upper limits applied to the data set. The truncated and untruncated parameter estimates for the form of the simulated data are then compared to show how i) the adjusted and unadjusted forms of the estimator differ, ii) the ability of the truncated estimator to recall the simulated data parameters.

Table A B.1. Simulated Pareto Type 1 Upper Limits

Estimated parameter values for simulated Pareto Type 1 data with upper truncation limit applied.

	Alpha		Gamma		Nu	KS Test p-values		Truncation Test
Limits	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted	Adjusted	Unadjusted	p-value
Simulated	1.800		100.000		NA	NA		NA
(0,1.000)	1.794	1.797	100.126	100.001	11617.149	0.992	0.991	0.143
(0,0.975)	1.780	1.989	100.049	100.001	758.851	0.996	0.002	0.000
(0,0.950)	1.789	2.143	100.059	100.001	530.985	0.990	0.000	0.000
(0,0.925)	1.783	2.290	100.052	100.001	422.512	0.993	0.000	0.000
(0,0.900)	1.795	2.433	100.090	100.001	362.822	0.973	0.000	0.000
(0,0.875)	1.781	2.580	100.063	100.001	318.363	0.992	0.000	0.000
(0,0.850)	1.796	2.731	100.067	100.001	289.362	0.977	0.000	0.000
(0,0.825)	1.805	2.884	100.032	100.001	266.221	0.965	0.000	0.000
(0,0.800)	1.801	3.044	100.006	100.001	246.761	0.974	0.000	0.000
(0,0.775)	1.795	3.215	99.986	100.001	230.703	0.978	0.000	0.000
(0,0.750)	1.807	3.397	99.992	100.001	217.738	0.965	0.000	0.000
(0,0.725)	1.764	3.589	99.973	100.001	205.246	0.985	0.000	0.000
(0,0.700)	1.741	3.791	99.963	100.001	195.180	0.974	0.000	0.000
(0,0.675)	1.749	4.006	99.966	100.001	187.014	0.974	0.000	0.000
(0,0.650)	1.749	4.238	99.966	100.001	179.577	0.969	0.000	0.000
(0,0.625)	1.750	4.489	99.966	100.001	172.869	0.963	0.000	0.000
(0,0.600)	1.769	4.760	99.973	100.001	166.992	0.968	0.000	0.000
(0,0.575)	1.736	5.056	99.962	100.001	161.179	0.948	0.000	0.000
(0,0.550)	1.730	5.377	99.961	100.001	156.165	0.929	0.000	0.000
(0,0.525)	1.739	5.730	99.963	100.001	151.619	0.926	0.000	0.000

Note: Limits are applied by removing the top proportion from the simulated data (the parameters of which are shown in the first line in gray), retaining the data contained in the lower proportion of the distribution.

Table A B.2. Simulated Generalized Pareto Upper Limits

Estimated parameter values for simulated generalized Pareto data with upper truncation limit applied.

	Alpha		Sigma		Gamma			KS test p-value		Truncation
Limits	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Nu	Adjusted	Unadjusted	Test p-value
(0,1.000)	1.800		1.200		100.000		NA	NA		NA
(0,1.000)	1.798	1.793	1.219	1.221	100.000	100.000	3272234	0.995	0.994	0.154
(0,0.975)	1.818	1.459	1.217	1.303	100.000	100.000	573.156	0.996	0.047	0.000
(0,0.950)	1.796	1.240	1.219	1.340	100.000	100.000	248.330	0.993	0.003	0.000
(0,0.925)	1.813	1.061	1.217	1.361	100.000	100.000	170.394	0.995	0.001	0.000
(0,0.900)	1.779	0.903	1.218	1.373	100.000	100.000	142.716	0.988	0.000	0.000
(0,0.875)	1.805	0.760	1.221	1.378	100.000	100.000	127.737	0.986	0.000	0.000
(0,0.850)	1.760	0.632	1.216	1.373	100.000	100.000	120.177	0.984	0.000	0.000
(0,0.825)	1.702	0.510	1.210	1.366	100.000	100.000	115.244	0.981	0.000	0.000
(0,0.800)	1.675	0.398	1.206	1.353	100.000	100.000	111.775	0.981	0.000	0.000
(0,0.775)	1.665	0.293	1.205	1.334	100.000	100.000	109.338	0.974	0.000	0.000
(0,0.750)	1.582	0.192	1.193	1.314	100.000	100.000	107.628	0.965	0.000	0.000
(0,0.725)	1.685	0.100	1.212	1.286	100.000	100.000	106.183	0.960	0.000	0.000
(0,0.700)	1.757	0.100	1.227	1.159	100.000	100.000	105.153	0.947	0.000	0.000
(0,0.675)	1.698	0.100	1.212	1.048	100.000	100.000	104.401	0.953	0.001	0.000
(0,0.650)	1.658	0.100	1.203	0.950	100.000	100.000	103.776	0.945	0.001	0.000
(0,0.625)	1.585	0.100	1.187	0.863	100.000	100.000	103.261	0.943	0.000	0.000
(0,0.600)	1.433	0.100	1.146	0.785	100.000	100.000	102.844	0.956	0.000	0.000
(0,0.575)	1.445	0.100	1.152	0.715	100.000	100.000	102.464	0.948	0.000	0.000
(0,0.550)	1.378	0.100	1.132	0.652	100.000	100.000	102.159	0.946	0.000	0.000
(0,0.525)	1.199	0.100	1.080	0.594	100.000	100.000	101.901	0.965	0.000	0.000

Note: Limits are applied by removing the top proportion from the simulated data (the parameters of which are shown in the first line in gray), retaining the data contained in the lower proportion of the distribution.

Annex C. Netherlands (2013) Results

Table 10. Estimated Pareto distribution parameters, Netherlands 2013

	Distribution	α	γ	σ	ν	KS p-value	Truncation p-value
Top-Tail Threshold: 10%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	2.3054	9,667.62			0.0000	
	Generalized Pareto	***0.0833***	***9,667.62***	***32,490.71***		***0.3159***	
	Truncated Pareto Type 1	2.0756	30,569.68		333,495	0.0000	0.0099
	Truncated generalized Pareto	*0.1040*	*9,667.62*	*32,073.70*	*333,495*	*0.3741*	*0.3928*
D1R Compensation of employees							
	Pareto Type 1	***3.4363***	***83,463.00***			***0.6182***	
	Generalized Pareto	*0.2472*	*83,463.00*	*25,449.30*		*0.7387*	
	Truncated Pareto Type 1	*3.4324*	*83,533.13*		*1,163,956*	*0.5860*	*0.8337*
	Truncated generalized Pareto	*0.2487*	*83,463.00*	*25,425.51*	*1,163,956*	*0.7387*	*0.9241*
D4R Property income received							
	Pareto Type 1	1.3085	2,576.00			0.0000	
	Generalized Pareto	***0.6757***	***2,576.00***	***2,685.08***		***0.4660***	
	Truncated Pareto Type 1	1.2921	2,905.45		468,268.9	0.0010	0.2038
	Truncated generalized Pareto	*0.6935*	*2,576.00*	*2,668.43*	*468,268.9*	*0.4365*	*0.3370*
D4P Property income paid							
	Pareto Type 1	3.2718	3,003.39			0.0001	
	Generalized Pareto	***0.2369***	***3,003.39***	***1,091.29***		***0.4356***	
	Truncated Pareto Type 1	3.2320	3,089.30		26,944.4	0.0024	0.3088
	Truncated generalized Pareto	*0.2509*	*3,003.39*	*1,084.82*	*26,944.4*	*0.3810*	*0.5368*
Top-Tail Threshold: 5%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	2.3054	32,662.00			0.0000	
	Generalized Pareto	***0.0224***	***32,662.00***	***37,670.02***		***0.8558***	
	Truncated Pareto Type 1	2.0756	42,755.53		333,495	0.0008	0.0099
	Truncated generalized Pareto	*0.0379*	*32,662.00*	*37,336.47*	*333,495*	*0.8558*	*0.7424*
D1R Compensation of employees							
	Pareto Type 1	***3.6800***	***103,113.0***			***0.7939***	
	Generalized Pareto	*0.3000*	*103,113.0*	*27,761.85*		*0.7939*	
	Truncated Pareto Type 1	*3.6753*	*103,550.2*		*1,163,956*	*0.7454*	*0.8973*
	Truncated generalized Pareto	*0.3055*	*103,113.0*	*27,682.57*	*1,163,956*	*0.7939*	*0.8375*
D4R Property income received							
	Pareto Type 1	***1.3301***	***4,916.00***			***0.6908***	
	Generalized Pareto	*0.8132*	*4,916.00*	*3,576.76*		*0.5021*	
	Truncated Pareto Type 1	*1.3070*	*4,980.28*		*468,268.9*	*0.7298*	*0.2248*
	Truncated generalized Pareto	*0.8759*	*4,916.00*	*3,518.98*	*468,268.9*	*0.4582*	*0.1417*
D4P Property income paid							
	Pareto Type 1	***3.2718***	***3,846.58***			***0.8750***	

	Generalized Pareto	*0.3479*	*3,846.58*	*1,104.70*		*0.9732*	
	Truncated Pareto Type 1	*3.2320*	*3,827.92*		*26,944.4*	*0.9850*	*0.3088*
	Truncated generalized Pareto	*0.4049*	*3,846.58*	*1,074.908*	*26,944.4*	*0.9850*	*0.2268*
Top-Tail Threshold: 1%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	***3.2906***	***92,631.00***			***0.2973***	
	Generalized Pareto	*0.0190*	*92,631.00*	*40,812.95*		*0.9997*	
	Truncated Pareto Type 1	*2.8336*	*94,082.22*		*333,495*	*0.7052*	*0.1762*
	Truncated generalized Pareto	*0.0991*	*92,631.00*	*39,527.58*	*333,495*	*1.0000*	*0.7243*
D1R Compensation of employees							
	Pareto Type 1	***3.3793***	***160,069.0***			***0.2468***	
	Generalized Pareto	*0.5159*	*160,069.0*	*34,499.73*		*0.9860*	
	Truncated Pareto Type 1	*3.3460*	*156,894.3*		*1,163,956*	*0.8299*	*0.8348*
	Truncated generalized Pareto	*0.6176*	*160,069.0*	*33,272.53*	*1,163,956*	*0.9860*	*0.5110*
D4R Property income received							
	Pareto Type 1	***0.9087***	***15,388.56***			***0.1379***	
	Generalized Pareto	*1.1123*	*15,388.56*	*11,638.99*		*0.8527*	
	Truncated Pareto Type 1	0.6176	10,626.50		468,268.9	0.0006	0.0075
	Truncated generalized Pareto	*2.4423*	*15,388.56*	*10,731.36*	*468,268.9*	*0.9251*	*0.0155*
D4P Property income paid							
	Pareto Type 1	***3.1598***	***6,279.98***			***0.9539***	
	Generalized Pareto	*0.1838*	*6,279.98*	*2,236.26*		*0.6387*	
	Truncated Pareto Type 1	*2.9669*	*6,253.74*		*26,944.4*	*0.8994*	*0.2185*
	Truncated generalized Pareto	*0.3183*	*6,279.98*	*2,105.07*	*26,944.4*	*0.8240*	*0.5713*

Table 11. Estimated Pareto distribution parameters, Netherlands 2013

Item	Distribution	α	γ	σ	ν	**KS p-value**	**Truncation p-value**
B2G_B3GR Operating surplus and Mixed income	Generalized Pareto	0.0833	9,667.62	32,490.71		0.3159	
D1R Compensation of employees	Pareto Type 1	3.4363	83,463.00			0.6182	
D4R Property income received	Generalized Pareto	0.6757	2,576.00	2,685.08		0.4660	
D4P Property income paid	Generalized Pareto	0.2369	3,003.39	1,091.29		0.4356	

Table 12. Primary Income Component Inequality, Netherlands 2013

Country: Netherlands Year: 2013		Income Components				
Method	Share	B2G_B3GR Operating surplus and Mixed income	D1R Compensation of employees	D4R Property income received	D4P Property income paid	B5 Primary Income

Unadjusted	Bottom 50%	16.80	13.77	20.32	21.86	14.89
	Middle 40%	42.09	59.05	24.53	53.85	52.30
	Top 10%	41.11	27.18	55.15	24.29	32.81
	Coverage Ratio	69.00%	77.58%	21.60%	100%	68.95%
Pareto (10%)	Bottom 50%	16.77	13.67	19.42	21.84	14.70
	Middle 40%	41.95	59.04	22.71	53.75	52.05
	Top 10%	41.28	27.30	57.88	24.41	33.25
	Coverage Ratio	69.33%	77.62%	23.22%	100.16%	69.23%
Pareto (5%)	Bottom 50%	16.74	13.63	18.78	21.69	14.59
	Middle 40%	42.07	59.04	21.21	53.66	51.89
	Top 10%	41.19	27.33	60.01	24.65	33.52
	Coverage Ratio	69.14%	77.60%	24.66%	100.31%	69.36%
Pareto (1%)	Bottom 50%	17.02	12.73	9.52	21.34	12.83
	Middle 40%	41.88	29.24	7.27	53.85	50.29
	Top 10%	41.10	28.03	83.21	24.811	36.88
	Coverage Ratio	69.83%	77.62%	66.52%	100.38%	74.63%

Table 13. Inequality statistics, Netherlands 2013

Netherlands 2013	OECD[6]			Assumed Top-Tail Shares[1]			Other Literature				
Shares	Top 10%	Top 5%	Top 1%	50% gap[5]	75% gap[5]	100% gap[5]	Proportional Allocation[5]	Proportional Allocation[6]	WID[4]	PovcalNet[3]	Statistics Netherlands[7]
Bottom 50%	14.69	14.56	12.81	12.60	11.05	9.51	13.79	14.89	18.00	31.86	6.70
Middle 40%	52.05	51.91	50.29	50.60	44.38	38.17	55.35	52.30	50.70	47.03	58.50
Top 10%	33.26	33.52	36.90	36.80	44.56	52.32	30.86	32.82	31.34	21.11	34.80
Top 1%	*9.32*	*9.70*	*14.31*	*7.29*	*8.83*	*10.37*	*6.35*	*8.73*	*6.52*	*4.50*	*8.70*
Gini	0.5761	0.5786	0.6106				0.5758	0.5722		0.2810	

[1] Törmähleto (2017[38]) and Lakner and Milanovic (2015[27]) calculate top-tail adjustments assuming a Pareto Type 1 distribution, utilising $N_{0.1} = NE[X|X_0] = N\frac{\alpha}{\alpha-1}X_0$ to calculate the sum value of the tail and the proportions within the tail. This is therefore calculated assuming that the total share of the top 10% is calculated as the sum of survey data and the proportion of the gap, with proportional allocation applied to the rest of the distribution.

[2] Top-tail values are calculated from the moments of the distribution, using identities from the Pareto Type 1 Distribution, and combined with survey data for the rest of the distribution. To ensure that the shares sum to 100%, the Top 10% share is calculated as the remainder after the Bottom 90% is removed. This number may differ from the theoretical share given using the formula above, which is used to calculate the Top 1% share.

[3] Results from http://iresearch.worldbank.org/PovcalNet/home.aspx, using Post-Tax Disposable Income for closest survey year.

[4] Results from https://wid.world/, using Fiscal Income.

[5] Results shown for Pre-Tax equivalised primary income (B5).

[6] Results shown for Pre-Tax equivalised primary income from sum of linked components.

[7] Primary income (B5) shares, with micro-macro gap unresolved calculated by Statistics Netherlands.

Annex D. Japan (2013) Results

Table 14. Estimated Pareto distribution parameters, Japan 2013

	Distribution	α	γ	σ	ν	KS p-value	Truncation p-value
Top-Tail Threshold: 10%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	1.9023	1,653,381			0.0000	
	Generalized Pareto	***0.2010***	***1,653,381***	***2,023,105***		***0.3574***	
	Truncated Pareto Type 1	*1.7991*	*2,111,433*		*29,783,395*	*0.3124*	*0.1418*
	Truncated generalized Pareto	*0.2241*	*1,653,381*	*2,004,204*	*29,783,395*	*0.3574*	*0.6992*
D1R Compensation of employees							
	Pareto Type 1	***3.2362***	***8,256,344***			***0.1457***	
	Generalized Pareto	*0.0545*	*8,256,344*	*3,272,929*		*0.0256*	
	Truncated Pareto Type 1	*2.9374*	*8,260,332*		*30,377,306*	*0.0996*	*0.0041*
	Truncated generalized Pareto	*0.1234*	*8,256,344*	*3,178,390*	*30,377,306*	*0.0317*	*0.3661*
D4R Property income received							
	Pareto Type 1	0.9079	205,618.2			0.0000	
	Generalized Pareto	***0.6835***	***205,618.2***	***596,452.7***		***0.8640***	
	Truncated Pareto Type 1	0.8648	308,918.6		110,901,600	0.0015	0.2307
	Truncated generalized Pareto	*0.7069*	*205,618.2*	*589,524.3*	*110,901,600*	*0.9152*	*0.8014*
D4P Property income paid							
	Pareto Type 1	1.3622	162,022.80			0.0000	
	Generalized Pareto	22.6113	162,022.80	0.0001		0.0000	
	Truncated Pareto Type 1	1.1887	67,226.16		5,670,798	0.0000	0.3155
	Truncated generalized Pareto	24,201,520	162,022.80	0.0001	5,670,798	0.0000	0.0000
Top-Tail Threshold: 5%							
B2G_B3GR Operating surplus and Mixed income							
	Pareto Type 1	*1.8497*	*3,100,000*			*0.7908*	
	Generalized Pareto	*0.4336*	*3,100,000*	*1,732,989*		*0.7908*	
	Truncated Pareto Type 1	***1.6860***	***3,009,276***		***29,783,395***	***0.9303***	***0.0777***
	Truncated generalized Pareto	*0.7086*	*3,100,000*	*1,581,790*	*29,783,395*	*0.8688*	*0.2765*
D1R Compensation of employees							
	Pareto Type 1	4.5853	10,300,293			0.0000	
	Generalized Pareto	***0.0001***	***10,300,293***	***3,845,983***		***0.4291***	
	Truncated Pareto Type 1	4.1259	11,193,523		30,377,306	0.0030	0.1398
	Truncated generalized Pareto	*0.0799*	*10,300,293*	*3,780,769*	*30,377,306*	*0.4291*	*0.3926*
D4R							

Property income received	***Pareto Type 1***	***1.0180***	***759,480.3***			***0.5350***	
	Generalized Pareto	*0.5186*	*759,480.3*	*1130593*		*0.7269*	
	Truncated Pareto Type 1	*0.9786*	*751,232.5*		*110,901,600*	*0.6304*	*0.3930*
	Truncated generalized Pareto	*0.5337*	*759,480.3*	*1121028*	*110,901,600*	*0.7269*	*0.9284*
D4P Property income paid							
	Pareto Type 1	1.3622	210,629.6			0.0000	
	Generalized Pareto	0.6923	210,629.6	88077.57		0.0014	
	Truncated Pareto Type 1	1.1887	210,629.6		5,670,798	0.0000	0.3155
	Truncated generalized Pareto	649,732	210,629.6	0.0001	5,670,798	0.0000	1.0000
Top-Tail Threshold: 1%							
Item: B2G_B3GR	*Not enough data points*						
Item: D1R							
Item: D4R							
Item: D4P							

Table 15. Selected Pareto distribution parameters, Japan 2013

	Distribution	α	γ	σ	ν	KS p-value	Truncation p-value
B2G_B3GR Operating surplus and Mixed income	Generalized Pareto	0.2010	1,653,381	2,023,105		0.3574	
D1R Compensation of employees	Pareto Type 1	3.2362	8,256,344			0.1457	
D4R Property income received	Generalized Pareto	0.6835	205,618.2	596,452.7		0.8640	
D4P Property income paid	NA	NA	NA	NA	NA		

Table 16. Primary Income Component Inequality, Japan 2013

Country: Japan Year: 2013		Income Components				
Method	Share	B2G_B3GR Operating surplus and Mixed income	D1R Compensation of employees	D4R Property income received	D4P Property income paid	B5 Primary Income
Unadjusted	Bottom 50%	33.33	11.83	21.04	19.45	16.19
	Middle 40%	43.49	56.25	31.71	57.86	52.22
	Top 10%	23.18	31.92	47.25	22.69	31.59
	Coverage Ratio	79.09%	70.38%	49.19%	100%	70.14%
Pareto (10%)	Bottom 50%	32.27	11.57	19.17	19.87	15.67

	Middle 40%	44.19	54.71	28.35	57.55	50.93
	Top 10%	23.54	33.72	52.48	22.58	33.41
	Coverage Ratio	79.58%	72.60%	55.36%	100.00%	72.37%
Pareto (5%)	Bottom 50%	31.53	10.97	5.08	19.75	14.04
	Middle 40%	40.06	56.33	4.93	57.32	49.72
	Top 10%	28.41	32.70	90.00	22.93	36.24
	Coverage Ratio	86.57%	70.58%	318.49%	100.00%	91.44%
Pareto (1%)	Bottom 50%	*Not enough data points*				
	Middle 40%					
	Top 10%					
	Coverage Ratio					

Table 17. Inequality statistics, Japan 2013

Japan 2013	OECD[6]			Assumed Top-Tail Shares[1]			Other Literature					
Shares	Top 10%	Top 5%	Top 1%	50% gap[5]	75% gap[5]	100% gap[5]	Proportional Allocation[5]	Proportional Allocation[6]	WID[4]	PovcalNet[3]	ESRI[7]	Taxable Income[8]
Bottom 50%	15.64	14.03		14.98	13.22	11.45	16.32	16.12	19.52	28.34	26.44	14.06
Middle 40%	50.92	49.68		48.22	42.54	36.85	52.54	52.26	38.13	45.25	48.54	35.79
Top 10%	33.43	36.29		36.70	44.16	51.63	31.14	31.62	43.35	26.41	25.01	50.15
Top 1%	*8.57*	*12.50*		7.68	9.24	10.80	6.40	6.79	12.46			*20.35*
Gini	0.5427	0.5725					0.5254	0.5293	0.5100			

[1] Törmähleto (2017[38]) and Lakner and Milanovic (2015[27]) calculate top-tail adjustments assuming a Pareto Type 1 distribution, utilising $N_{0.1} = NE[X|X_0] = N\frac{\alpha}{\alpha-1}X_0$ to calculate the sum value of the tail and the proportions within the tail. This is therefore calculated assuming that the total share of the top 10% is calculated as the sum of survey data and the proportion of the gap, with proportional allocation applied to the rest of the distribution.

[2] Top-tail values are calculated from the moments of the distribution, using identities from the Pareto Type 1 Distribution, and combined with survey data for the rest of the distribution. To ensure that the shares sum to 100%, the Top 10% share is calculated as the remainder after the Bottom 90% is removed. This number may differ from the theoretical share given using the formula above, which is used to calculate the Top 1% share.

[3] Results from http://iresearch.worldbank.org/PovcalNet/home.aspx, using Post-Tax Disposable Income for closest survey year.

[4] Results from https://wid.world/, using Pre-Tax National Income.

[5] Results shown for Pre-Tax equivalised primary income (B5).

[6]Results shown for Pre-Tax equivalised primary income from sum of linked components.

[7]Retults for 2014 Primary Income (B5) shares reported in (Yamazaki and Sakamaki, 2018[39])

[8]Returns for 2013 Taxable Income (43) reported by National Tax Agency, table 2-4 Self-Assessment Income Tax by Income Type https://www.nta.go.jp/publication/statistics/kokuzeicho/h25/h25.pdf