Tracing the invisible rich: A new approach to modelling Pareto tails in survey data

Rafael Wildauer[a]

[a]University of Greewnich
Department of Economics and International Business
30 Park Row, London SE10 9LS, United Kingdom
Corresponding author, r.wildauer@gre.ac.uk

Jakob Kapeller[b]

[b]University Duisburg-Essen
Institute for Socio-Economics
Lotharstr. 65, 47057 Duisburg, Germany
jakob.kapeller@uni-due.de

Abstract

This paper is concerned with the problem of modelling the tail of the wealth distribution with survey data when the data does not adequately cover the tail of the distribution. In order to deal with the problem post data collection, it is standard practice to either fit a Pareto tail to the data or to combine wealth survey data with observations from rich lists before fitting such a Pareto tail. This paper proposes a novel approach ('rank correction') to address such cases which does not require additional data-sources. Applying the rank correction approach to wealth survey data (HFCS, SCF, WAS) yields estimates of top wealth shares, which are closely in line with estimates from the World Inequality Database and therefore represent a significant improvement over the raw survey data. While the paper focuses on the distribution of wealth as a case in point the rank correction approach might generally prove useful in contexts, where the tail of a Pareto-distributed variable is not adequately covered by the available data.

JEL Classification: D31, C46, C81

Keywords: differential nonresponse, Pareto tail, post data collection, survey data, wealth distribution

Acknowledgements: The usual disclaimer applies. We thank the editor, an anonymous referee, Rob Calvert Jump, Ines Heck, Karsten Köhler and Stefan Steinerberger for useful comments and discussions. Previous versions of this research benefited from funding from the Chamber of Labour Vienna.

1 Introduction

Much applied work on economic inequality makes use of survey data. And indeed, survey data comes with a series of advantages for applied researchers: surveys are usually designed to ensure that the sample adequately represents the underlying population, they can be carried out repeatedly to trace changes over time and they typically provide a rich set of contextual variables (e.g. information on age, gender, education and the like). The possibility to include such contextual variables is a major advantage of survey data as compared to administrative data from tax returns, which typically only provide limited information about the individual or household. Additionally, tax data is not available for many countries, which makes survey data the prime data source for conducting large-scale international comparisons.[1]

However, survey data on inequality of wealth or income does come with one essential drawback: in many cases it does not adequately cover the tail of the distribution, that is, the richest households or those with the highest incomes. While it might seem negligible at first sight that some survey does not adequately cover the top 5% or the top 1% as this is only a small share of the population of interest, it should be emphasized that this lack of coverage represents a non-random measurement error, that can have a huge impact on final estimates. For instance, even underestimating the wealth held by the top 1% can strongly bias our estimate of total wealth or wealth inequality as the top 1% typically hold a substantial fraction of total wealth.

The reasons why surveys do not adequately cover the top tail of the income or wealth distribution can be grouped into three rough categories: first, there are administrative constraints, like confidentiality rules, which will often lead to the introduction of an arbitrary cut-off, above which no income or wealth is reported. For instance, the *Survey of Consumer Finances* (SCF) in the US deliberately excludes all people listed on the list of billionaires published by the *Forbes* magazine, while the German *Einkommens- und Verbrauchsstichprobe* does not collect incomes greater than 216,000 €/year. Technically, we can conceive of this administrative constraints as a form of *binary non-response*, where observations with specific properties will be automatically excluded from the survey.

A second problem of survey-based estimates for highly skewed characteristics such as income or wealth is that they suffer from *non-observation bias*. Non-observation bias comes in the form of a median-bias implying that a given survey will underestimate the true value of total wealth or income with a probability greater than 50% as the extreme values at the tail will be overlooked in the majority of draws. Hence, the median estimate of a series of repeated surveys will be downward biased.[2]

[1] While Piketty (2014) frames tax and survey data as rival approaches, which both come with their own limitations, Saez & Zucman (2016, p. 569) point out that survey and administrative data can be used as complements in order to derive a more detailed and fine-grained assessment of the distribution of wealth and income. Indeed the fruitful effort of constructing Distributional National Accounts (Piketty et al. 2016) relies on a combination of survey, tax and national accounts data to arrive at a description of the distribution of wealth that is as accurate as possible. In addition the *Survey of Consumer Finances* (SCF) in the US is a prime example of using administrative data to improve the accuracy of surveys.

[2] Encouragingly Eckerstorfer et al. (2016) show that this bias can be reduced by fitting a Pareto model to the data.

A third source of bias is that the probability of participating in such surveys is negatively correlated with the variable of interest itself, which leads to a *differential non-response bias*, because observations are missing not at random (Little & Rubin 2019). The evidence for differential non-response is compelling and can be illustrated for the case of wealth with reference to the SCF, where tax data on capital incomes are used to identify affluent households prior to data collection. While the response rate in the stratified random sample is about 70%, it sharply decreases for the so-called list sample of affluent households, which are ex ante identified based on tax records. Among these affluent households, even the poorest stratum has a response probability of only 50%, which further decreases to 12% for the stratum of the wealthiest households (Bricker et al. 2016, p. 282). Similarly, D'Alessio & Faiella (2002) report a response rate of 26% for the lowest wealth group which declines to 9% in the highest wealth group when in 1998 data from a commercial bank was used to identify affluent individuals in an oversampling effort for a wealth survey conducted by the Italian central bank. For the HFCS Osier (2016) emphasizes that non-response rates are not random and that additional data especially on income or wealth would be desirable to improve sample designs across countries.[3]

These problems became more pertaining in recent years as the last decade saw the publication of several novel data sources suitable for studying the distributions of wealth and income. Among these are the World Inequality Database (www.wid.world), the Household Finance and Consumption Survey (HFCS) carried out under the auspices of the ECB, the UK's Wealth and Asset Survey (WAS) as well as efforts to exploit insights from data leaks on offshore wealth holdings (Alstadsæter et al. 2019). For the United States the Survey of Consumer Finances (SCF) has been conducted regularly and consistently since 1989 and is considered as the most reliable source for assessing the distribution of private wealth. In Europe, three waves of the HFCS (2011, 2014, 2017) have been conducted and provide information on the distribution of wealth for up to 22 EU countries. For many of these countries this data source is a true novelty as reliable alternative data sources for assessing the distribution of private wealth have not been available before.[4]

In practical terms there are three established ways to engage with these problems: The first is the raw data approach. It implies to do nothing and to use the data as it is. This amounts to assuming that the efforts made by the administrators of the survey were sufficient to ensure adequate coverage of the tail of the income or wealth distribution. Most importantly this would require a strong over-sampling strategy where information, which is available already prior to data collection, is used to identify households with high wealth or income and include a disproportionate amount of them in the gross sample to ensure enough responses despite a lower response probability. However, this is rarely implemented with the necessary diligence: for instance, in the HFCS, only a handful of countries implement a convincing oversampling strategy based on high quality data such as income tax records. The second way forward is the Pareto

[3] By summarizing the inherent biases in survey data in this way, we actually abstract from the possibility of (differential) underreporting of incomes or assets as another possible source of measurement error.

[4] In addition, recent works by Piketty et al. (2016) and Saez & Zucman (2016) complement these efforts as they move towards producing data on wealth that are consistent with micro (Survey of Consumer Finances, SCF) as well as macro (Financial Accounts) sources.

correction approach. It involves fitting a Pareto distribution to the tail of the survey data and to use the estimated distribution to describe the tail instead of the tail observations (Jayadev 2008, Eckerstorfer et al. 2016). This can partly account for binary non-response and is suitable for dealing with non-observation bias, but does not compensate for differential non-response. As a consequence researchers developed the rich list correction approach (e.g. Vermeulen (2018)), which extends the second approach by adding journalists' rich lists like the *Forbes 400* for the US or the rich list provided by the *Manager Magazin* for Germany to the original survey data. A Pareto distribution is then fitted to the combined data-set and the fitted distribution is used to describe the tail of the wealth distribution (Advani et al. 2020, Bach et al. 2019, Vermeulen 2018). The success of the rich list correction approach fundamentally depends on the availability, size and quality of the rich lists used, which comes with some limitations. For example, the Forbes list of billionaires includes less than 10 entries for 18 of the 22 countries in the HFCS.

Our paper aims to supplement these three methods by a fourth one, the rank correction approach. The key advantage of this new tool is that it does not require additional external information while it significantly improves upon the raw data Pareto correction approaches. This means it can be used even when the rich list correction approach is not feasible, either due to poor quality (Capehart 2014, Kopczuk 2015) or complete lack of rich list data.

The core idea of the rank correction approach is to correct the ranks of the sample observations (i.e. the cumulative sum of the survey weights) in order to take into account that the most affluent households are much less likely to be included in the sample. This correction aims to preserve the linearity of the relationship between logarithms of household wealth and rank underlying the Pareto distribution, which is exploited when fitting the distribution to the data. We demonstrate that this simple adjustment is able to substantially reduce the bias from differential non-response when fitting a Pareto distribution to the tail of wealth survey data.

Applying the rank correction approach to wealth survey data shows that the average estimate of the Pareto tail index declines from 2.1 obtained from the Pareto correction approach to 1.8 after implementing the rank correction procedure. In comparison, using rich lists Vermeulen (2018) reduces the average tail index on a similar set of countries from 2.1 to 1.6. Using these Pareto tails to replace the tail from the survey indicates that the rank correction approach significantly improves upon the raw data approach. For example the top 1% wealth shares for the US, France and Germany based on the WID and on Schröder et al. (2020) are 37%, 23.4% and 35.3%, respectively. These compare very well with the results obtained from applying the rank correction approach (36.2%, 23.3% and 32.2%) and represent a clear improvement relative to the top wealth shares obtained from the raw survey data (35.4%, 18.7% and 23.6%).

The remainder of the paper is organised as follows. Section 2 introduces the rank correction approach. Section 3 analyses its performance by means of Monte Carlo simulations. In Section 4 contains an application to data from the HFCS, SCF and WAS. Section 5 contains a summary and concludes.

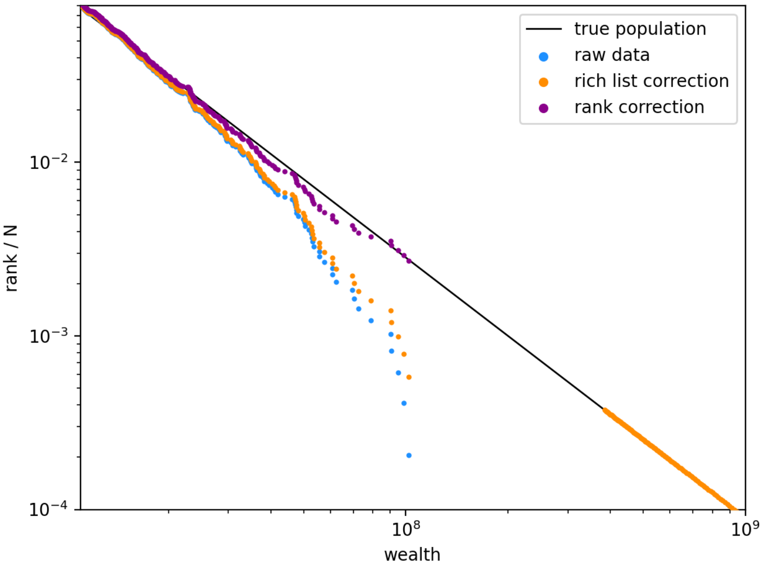
2 The Rank Correction Approach

In introducing the rank correction approach we focus mainly on those two sources of bias – binary non-response and differential non-response – that cannot be remedied by a standard Pareto correction and, hence, typically require rich lists to arrive at reliable estimates.[5] In doing so we first try to build some intuition for the underlying bias by means of a simplified example. We then revisit the standard approach of fitting a Pareto distribution to the tail of survey data. On this basis we explain the idea of the rank correction approach and show how to implement it by modifying the standard procedure.

2.1 A graphical motivation of the rank correction approach

In this example we focus on binary non-response, which is the simplest form of systematic bias in observing a given population. This setup yields an intuitive graphical explanation of why the naïve Pareto model fails under binary or differential non-response. We can model binary non-response by assigning a zero response probability to the richest households while assigning a nonzero probability to all other households. For our setup we assume a hypothetical tail population of 1 million households (N_T) which are described by a Pareto distribution with minimum level of wealth of € 1 million (x_m) and a shape parameter equal to 1.5 ($\alpha = 1.5$). For purposes of presentation we make use of a log-log plot in Figure 1, which shows the usual linear relationship between log(wealth) and log(rank) for the assumed population (black line). Thereby we normalize the ranks by either the total population (N_T) or the size of the sample (n) so that these re-scaled ranks represent the survival function, also known as complementary cumulative distribution function (CCDF).

Figure 1: Motivating the rank correction approach

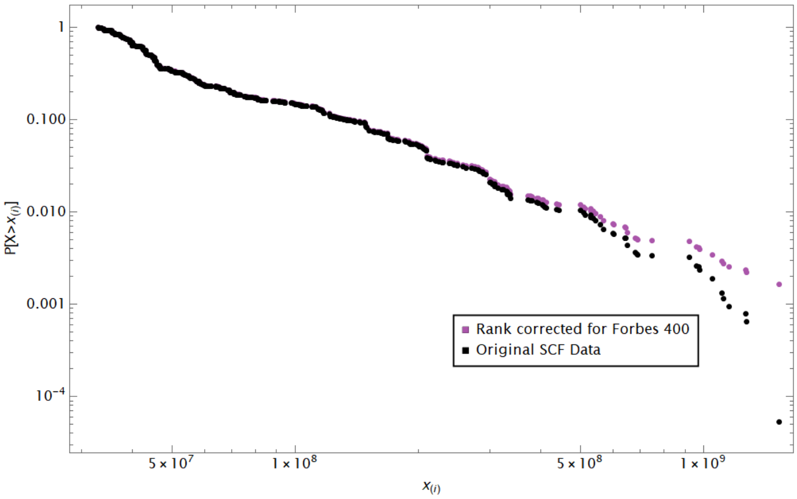


In Figure 1 the black line indicates the log linear population relationship between relative

[5] Vermeulen (2018, p. 377) shows that while the Pareto correction improves the estimate of the tail wealth, the average Pareto model still underestimates the actual tail wealth between 17% and 4%.

ranks and wealth (the slope of that line represents the shape parameter α). The blue dots represent a sample drawn from the population under the assumption that the most affluent 1000 households cannot be sampled due to privacy concerns. As a result the linear relationship between the log of the CCDF and the log of wealth, breaks down and thus regression based estimates of the slope parameter will be biased. Conceptually, this problem also persists in the case of a rich list correction, e.g. when adding the most affluent 100 observations to the random sample (which appear as an orange cluster towards wealth of 10^9 in Figure 1), if there remains a substantial gap between the maximum in the data and the minimum of the rich list as indicated by the orange dots in Figure 1[6]. In contrast, the purple dots take the fact that the most affluent 1000 observations are not observed into account by correcting the ranks accordingly and, hence, retain the log-linear relationship between wealth and ranks. Thereby the key strength of the rank correction approach (purple) is that it requires significantly less information about those households which are not observed compared to the rich list approach and thus is applicable in situations where rich lists are not available.

Figure 2: Rank correction applied to the SCF



The example in Figure 1 is a highly stylized illustration of why binary differential nonresponse leads to the breakdown of the log-linear relationship between household wealth and the empirical CCDF. In contrast, Figure 2 provides an illustration using data from the Survey of Consumer Finances (2016 wave). The SCF's design is tailored to protect the privacy of its participants and, hence, explicitly excludes individuals from the Forbes 400 List. Plotting the relationship between household wealth and household ranks for the observations representing the richest 250,000 US households based on the original SCF data reveals the breakdown of the log-linear relationship due to the exclusion of these richest 400 households. However, after adjusting the weights for the omission of these top 400 households, the purple dots again conform to a log-linear relationship. Therefore correcting the survey weights by taking the missing observations at the very top into account, will lead to improved estimates of the Pareto shape parameter.

This discussion focused on binary differential non-response. In practice, wealth survey data

[6] Of course if all households with a zero response probability were observed on a rich list, the problem would be solved. However, in practice this amounts to simply assuming that a perfect alternative data source exists.

will most likely suffer from more general forms of differential non-response, which, however, will create a bias with very similar effects to the one shown in Figure 1. Typically, it is observed that household responsiveness is a decreasing function of household wealth itself (Bricker et al. 2016, p. 282). While conceptually similar, addressing differential non-response bias in these cases is slightly more intricate, which is why we will rely on Monte Carlo simulations to study more elaborate forms of differential non-response in section (3). However, before doing so we will revisit the standard procedure for estimating the Pareto tail in data on household wealth or income to explain how to implement the rank correction approach in this procedure.

2.2 Fitting Pareto tails to wealth survey data: the standard approach

The standard approach of fitting a Pareto tail to wealth survey data is to fit the complementary cumulative distribution function (CCDF) of the Pareto distribution to the empirical CCDF derived from the available sample. This procedure effectively amounts to estimating the linear log-log relationship highlighted in Figure 1, where the slope of the estimated line corresponds to the shape parameter of the Pareto distribution. This correspondence becomes evident when keeping in mind that the CCDF is a natural way to express the ranks of the population as represented by the available data, especially in the context of survey weights dedicated to map insights from the data-set to the full population. However, to precisely compare the theoretical and the empirical CCDF we first need to introduce the respective definitions.

The theoretical $CCDF_T$ for a random variable X following a type I Pareto distribution above x_m is defined as[7]

$$CCDF_T(x_i) = Pr(X > x_i) = \left(\frac{x_m}{x_i}\right)^{\alpha}. \tag{1}$$

Technically, this definition asks for the probability to observe someone with wealth or income greater than x_i, which again indicates that the CCDF is just another way to express the underlying ranking of households or individuals. This intuition also guides the standard definition of the empirical $CCDF_E$: Let's assume a sample of households with net wealth $x = (x_1, \ldots, x_n)$ and corresponding survey weights $w = (w_1, \ldots, w_n)$, where the number of households represented by the available sample is defined as $N = \sum_{i=1}^{n} w_i$. Arranging the data in descending order (i.e., from the most to the least affluent observation) yields a data vector denoted as $x_d = (x_{(1)}, \ldots, x_{(n)})$ with the corresponding vector of weights $w_d = (w_{(1)}, \ldots, w_{(n)})$. Then the empirical $CCDF_E$ at some point x_i can be defined as the sum of weights assigned to households with $x \geq x_i$ divided by the full population, which gives the probability of observing someone at least as rich as x_i. Formally this can be stated as [8]

$$CCDF_E(x_{(i)}) = \frac{\sum_{j=1}^{i} w_{(j)}}{N}. \tag{2}$$

[7] Throughout the paper we refer to type I Pareto distributions when we talk about Pareto distributions.

[8] The precise reader will notice a slight inconsistency between the empirical $CCDF_E$ and the theoretical $CCDF_T$, which the standard approach silently accepts. $CCDF_E$ is defined as the probability to observe someone at least as rich as x_i when $CCDF_T$ is defined as the probability to observe someone richer than x_i. For a detailed discussion and workarounds see Wildauer & Kapeller (2019); also the derivation described in section 2.3 is suitable to avoid this inconsistency.

Setting the theoretical equal to the empirical CCDF leads to the following expression

$$\left(\sum_{j=1}^{i} w_{(j)}\right) = N \cdot \left(\frac{x_m}{x_i}\right)^{\alpha}. \tag{3}$$

Applying the logarithm to equation (3) and rewriting the resulting expression $\ln(N) + \alpha\ \ln(x_m)$ as c_1 naturally leads to the following regression equation, where the shape parameter of the Pareto distribution shows up as a regression slope parameter

$$\ln\left(\sum_{j=1}^{i} w_{(j)}\right) = c_1 - \alpha\ \ln(x_{(i)}) + \epsilon_i \tag{4}$$

where $c_1 = \ln(N) + \alpha\ \ln(x_m)$. Equation (4) is then estimated by OLS. The estimated Pareto tail index α is used to describe household wealth above x_m.

As has long been known in the more specialized literature (Aigner & Goldberger 1970), this standard approach to fitting Pareto distributions to tails of distributional data comes with a slight bias in the estimate for the shape parameter. Again this bias relates to how we define $CCDF_E$ (and, hence, our dependent variable) and it can be illustrated with a simple example: Assume a population of one-thousand people that is represented by two observations with equal weights of 500. In such a case our definition from equation (2) would assign a rank of 500 and 1000 to these observations, although both represent a certain range of households (one observation represents households with ranks 1 to 500, the other households with ranks 501 to 1000). Hence, taking the middle of these ranges – that is, 250 and 750 – can be considered as more apt to precisely describe the population under consideration relative to the naive sum of weights, and thus the endpoint of these ranges, used in equation (2).

Gabaix & Ibragimov (2011) have been the first to suggest a suitable method to correct for this bias, which, however, is formally rather intricate, which is why suggest to follow the next section when applying the standard approach. The formulations there are based on a supposedly more intuitive definition of the $CCDF_E$, that automatically applies the a bias-correction in the spirit of Gabaix & Ibragimov (2011) and, at the same time, avoids the complications mentioned in footnote 8 earlier (see Wildauer & Kapeller (2019) for a detailed discussion of these issues).

2.3 Deriving the rank correction estimator

Applying the rank correction estimator first requires a precise definition of the empirical CCDF, which we provide in Equation 5 below. As indicated before this definition generalizes the bias correction proposed by Gabaix & Ibragimov's (2011) to allow for an application to survey weights. The basic idea is closely related to our description of the intuition behind this bias and suggests to define the rank of an household not as the simple sum of weights, but as the sum of weights associated with more affluent households plus one half of its own weight. In other words, we take the median household represented by a certain observation as the anchor point for our regression instead of the household with lowest wealth represented by the same observation (which is what we arrive at when simply sum up weights). As this median household can be identified by taking the average between the cumulative weights of two consecutive data points, this argument corresponds to computing the empirical CCDF as

$$CCDF(x_{(i)})_{AV} = \frac{1}{N} \frac{\sum_{j=0}^{i-1} w'_{(j)} + \sum_{j=1}^{i} w'_{(j)}}{2} \tag{5}$$

where $w'_{(0)} = 0$. As Wildauer & Kapeller (2019) observe, this averaging procedure effectively amounts to defining the $CCDF(x_{(i)})_{AV}$ as the average between the empirical CCDF based on a data vector in descending order and the empirical CCDF based on a data vector in ascending order.

Based on this improved and more precise definition of the empirical CCDF, we can now apply the rank correction approach by increasing the most affluent household's sample weight by the rank correction factor (u). This means we shift all ranks up by u in order to account for super wealthy households which are excluded from the sample due to privacy concerns on the one hand (binary differential nonresponse), as well as for more general forms of differential nonresponse on the other. This results in the rank corrected empirical CCDF:

$$CCDF(x_{(i)})_{RC} = \frac{1}{N} \frac{\left[\sum_{j=0}^{i-1} w'_{(j)} + \sum_{j=1}^{i} w'_{(j)}\right] + 2u}{2} \tag{6}$$

where $w'_{(j)}$ are the elements from a vector of adjusted weights $w'_d = (w'_{(1)}, \ldots, w'_{(n)})$. The weight adjustment ensures that the number of households represented by the sample ($N = \sum_{j=1}^{n} w_{(j)}$) is unchanged after introducing the rank correction factor u. In order to achieve this the original weights are rescaled proportionally:

$$w'_{(i)} = w_{(i)} \left(1 - \frac{u}{N}\right) \tag{7}$$

Finally we can combine equation (6) with the theoretical CCDF (equation 1) and obtain the rank correction regression equation which can be estimated by OLS:

$$\ln\left(\left[\sum_{j=0}^{i-1} w'_{(j)} + \sum_{j=1}^{i} w'_{(j)}\right] + 2u\right) = c_2 - \alpha \ln(x_{(i)}) + \epsilon_i \tag{8}$$

where $c_2 = \alpha \ln(x_m) + \ln(2N)$ and $w'_{(0)} = 0$.

2.4 An algorithm for choosing u

In practice wealth survey data such as the SCF, HFCS or the WAS suffer from two forms of differential nonresponse. First, binary differential nonresponse in the form of the exclusion of the richest households from the sampling frame due to privacy concerns and second, general differential nonresponse in the form of response rates declining with increasing household wealth. Interpreting these two problems as deviations from linearity in the log rank, log wealth relationship, opens up a route to choosing an appropriate rank correction factor (u): Choose u such that the adjusted data is as close to the linear specification of equation (8) as possible. We use the root mean squared error (RMSE) of a regression based on equation (8) as a measure of deviation from linearity and choose u such that we minimize the RMSE

$$\min_u \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\ln\left(\left[\sum_{j=0}^{i-1} w'_{(j)} + \sum_{j=1}^{i} w'_{(j)}\right] + 2u\right) - \hat{c}_2 + \hat{\alpha} \ln(x_{(i)}) \right)} \tag{9}$$

where $\hat{c}_2$ and $\hat{\alpha}$ are the OLS estimates from equation (8). This one-dimensional minimization problem can be solved numerically.

In the next section we study the ability of the rank correction estimator to correct the bias when estimating type I Pareto tails under a restricted sample due to privacy concerns as well as under more general non-response problems. After that we apply it to EU, UK and US wealth survey data and assess to what extent the rank correction approach can improve the picture of the wealth distribution which we obtain from the data.

3 A Simulation Study

This section applies the rank correction approach in a Monte-Carlo setting. In doing so we analyze two main cases relevant for our argument: in the first scenario we apply the rank correction estimator to simulated data-sets where the richest members of the population were excluded from the sampling design due to privacy concerns as is the case in SCF data. In other words, the first scenario focuses on binary differential non-response. In the second scenario, which is presented in section 3.2, we introduce a more complex non-response mechanism inspired by non-response patterns observed in US data (Vermeulen 2018).

3.1 Addressing binary differential nonresponse

The simulation study is set up in the following way. We assume that the tail population (N) of interest consists of 1 million households and follows a Pareto distribution with scale parameter $x_{min} = 1,000,000$. This is roughly in line with French and German samples in the second wave of the HFCS, according to which there are 1.24 million millionaire households in Germany and 930,000 millionaire households in France[9]. Furthermore, our simulations are based on three different shape parameters $\alpha = (1.25,\ 1.5,\ 1.75)$ and four different net sample sizes $(s_1, ..., s_4) = (0.3‰,\ 0.8‰,\ 2‰,\ 6‰)$, which gives twelve different scenarios in total. Each scenario is analysed based on 1,000 draws from the population.

The exclusion of super rich households due to privacy concerns is modelled by setting the response probability of these households to zero, while assuming a response probability of 40% for the remaining population[10]. Thus we define the response mechanism as:

$$R_1(x_i) = \begin{cases} 0.4, & \text{for } x_{min} \leq x_i < x_{\text{NO}} \\ 0, & \text{for } x_{\text{NO}} \leq x_i \leq x_{max} \end{cases} \tag{10}$$

where $R_1(x_i)$ is the response probability of household i depending on its wealth (x_i). Here x_{NO} denotes the level of wealth of the poorest non-observed (NO) household that is excluded from the sampling frame due to privacy concerns and x_{max} is the richest of these excluded and non-observed households. For Germany and France the entries on national rich lists represent

[9] Vermeulen (2018) uses a similar setting and thus our simulations can be directly compared.

[10] This is roughly in line with available data. For example Bricker et al. (2016) report response rates for the richest strata in the SCF between 50% and 12%.

about 1,600 households and so we assume that the richest 1,600 households in the simulated tail population have a response probability of zero[11].

Table 1: Simulation results for response mechanism R_1

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
true	sample	data	PC	PC	RC	RC	RC	SRL	SRL	LRL	LRL
α	size (‰)	WD	$\hat{\alpha}$	WD	$\hat{\alpha}$	WD	u	$\hat{\alpha}$	WD	$\hat{\alpha}$	WD
1.5	0.03	-12.1	1.545	-4.4	1.473	4.3	2800	1.484	2.1	1.495	0.8
	0.08	-11.7	1.545	-5.4	1.483	3.0	2100	1.503	-0.5	1.498	0.3
	0.20	-11.4	1.547	-5.5	1.493	1.2	1800	1.523	-2.8	1.503	-0.3
	0.60	-11.5	1.550	-6.0	1.499	0.1	1600	1.539	-4.9	1.514	-1.8
1.25	0.03	-28.5	1.289	-9.6	1.228	8.9	2850	1.229	7.3	1.245	2.1
	0.08	-28.1	1.290	-10.8	1.240	3.3	1900	1.247	0.7	1.247	1.1
	0.20	-27.6	1.291	-11.2	1.246	2.1	1800	1.266	-4.8	1.251	-0.4
	0.60	-27.5	1.290	-11.1	1.248	0.8	1700	1.280	-8.6	1.260	-3.2
1.75	0.03	-6.8	1.805	-3.6	1.722	2.7	2600	1.739	0.8	1.746	0.5
	0.08	-6.2	1.806	-3.6	1.734	1.5	2000	1.759	-0.6	1.749	0.2
	0.20	-6.4	1.810	-4.2	1.748	0.3	1700	1.782	-2.4	1.755	-0.4
	0.60	-6.2	1.808	-4.1	1.748	0.1	1700	1.797	-3.3	1.767	-1.3

Simulation results based on different population alphas (column 1) and net sample sizes in ‰(column 2) for the Pareto correction (PC), rank correction (RC), a rich list correction with a short rich list (SRL) of 5 entries and a rich list correction approach with a long rich list (LRL) of 100 entries. For each approach the average estimated α is reported ($\hat{\alpha}$) as well as the median percent deviation from true tail wealth (WD). Results are based on 1,000 draws per sample size. For the RC approach the median of the chosen u's is reported in column 8. Tail population $N = 10^6$ and $x_{min} = 10^6$.

Table 1 presents the simulation results based on 1,000 draws for each sample size using the binary non-response mechanism $R_1(x_i)$ as specified above. Results are presented for different population shape parameters (true alphas, column 1) and net sample sizes (column 2). We compare the performance of the rank correction approach (columns 6 to 8) against the raw data approach (column 3), the Pareto correction approach (columns 4 and 5), and two different versions of a rich list correction approach (columns 9 to 12), which assume rich lists of different quality. The performance of the different strategies is evaluated with respect to two criteria: first, we compare the median deviation from true wealth in percent of the true population total (WD) occurring in the respective estimation for all available approaches. Second, we also present the shape parameter α of the estimated Pareto distribution, which is shown for all approaches except the raw data approach, where no such estimate is calculated.[12]

Since the numerical minimization problem of finding an optimal correction factor u for the rank correction approach is one dimensional we solve it computationally over the parameter

[11] Rich list entries often represent entire family clans and thus one entry often represents more than one household.

[12] Estimating a Pareto Distribution on the raw data, will lead to the estimates associated with the Pareto correction shown in columns 4 & 5.

range $u = [100; 10,000]$. The median of chosen u's across the 1000 draws is also reported in column (8). As indicated above, we analyzed two different scenarios to cover the rich list correction approach. One in which the researcher only has a very short rich list of the most affluent 5 observations (columns 9 and 10) and one in which the researcher observes the most affluent 100 missing households (columns 11 and 12). These two scenarios are motivated to span the range of situations researchers face in practice: On the one hand individual country rich lists can be rather exhaustive such as the *Manager Magazin*'s list for Germany. On the other end of the spectrum, however, are those cases where the analysis requires rich lists based on a consistent methodology across several countries. In that case the only option left is the *Forbes* global list. However, for many European countries the latter only includes 5 or less observations.[13]

For example in row 1 of Table 1 we see the results based on simulations with a shape parameter of $\alpha = 1.5$ and a sample size of 0.03‰ with the following results: Only using the raw sample data underestimates total wealth in the tail by 12.1%, using the PC approach yields an average alpha estimate of $\hat{\alpha} = 1.545$ and an underestimation of the tail of 4.4%, using the RC approach yields an average alpha estimate of $\hat{\alpha} = 1.473$ based on median correction factor $u = 2800$ and an overestimation of tail wealth of 4.3%, using the SRL approach yields an average estimate of $\hat{\alpha} = 1.484$ and an overestimation of tail wealth by 2.1% and using the LRL approach yields an average estimate of $\hat{\alpha} = 1.495$ and an overestimation of tail wealth by 0.8%. The following three rows show simulation results which are based on the same shape parameter ($\alpha = 1.5$) but increasing sample sizes, from 0.08‰to 0.6‰. The following rows provide simulation results for the same sample sizes but with different shape parameters $\alpha = 1.25$ and $\alpha = 1.75$.

The first main result from Table 1 is that the rank correction (RC) approach consistently outperforms the raw data approach. In column (3) we report the percent deviation of estimated aggregate wealth based on the raw sample from the true aggregate. The raw sample data underestimates the population wealth substantially whereas the RC approach (column 7) slightly overestimates aggregate wealth but yields a smaller absolute deviation from the true aggregate. This pattern holds for populations with different shape parameters and different sample sizes. Most importantly for large sample sizes the RC deviation converges towards zero while the raw data approach underestimates aggregate wealth even in the largest samples. The second main result from Table 1 is that the RC approach also consistently outperforms the Pareto correction (PC) approach across all shape parameters and population sizes. The absolute deviation from the true wealth aggregate for the PC approach, reported in column 4, is consistently larger than the RC deviation in column 7. Thirdly, the RC approach outperforms the rich list approach based on short rich lists (SRL), reported in column 10 for the two larger sample sizes of 0.2‰ and 0.6‰. Even when the researcher has access to a long rich list (LRL) of 100 observations (column 12), the rank correction approach still dominates for the largest sample size. The good performance of the long rich list approach is expected as it amounts to a situation where an alternative data source partially resolves the differential non-response problem at hand. However as the sample

[13] We also assume that rich lists are completely accurate and do not suffer from measurement error, which abstracts from the usual reliability problems that come with such lists.

size increases, the lack of information beyond the top 100 households becomes apparent and the RC approach starts to dominate.

The mean shape parameters based on which the wealth deviation calculations are based are reported in columns 4, 6, 9 and 11. Negative wealth deviations are due to an overestimated shape parameter and positive deviations correspond to an underestimated shape parameter. For the RC approach column 8 also reports the median value of the chosen correction factor u which approaches the true value of 1600 excluded households as the sample size increases.

Overall Table 1 shows that not taking the exclusion of super rich households due to privacy concerns into account induces a significant bias in the estimation of tail wealth even if no other differential non-response problem plagues the data. In this setting the rank correction approach unconditionally outperforms the raw data and the Pareto correction approach. Even for large sample sizes the RC approach performs similarly well or better than the rich list correction approach, however with the crucial advantage of not depending on additional exogenous information in the form of rich lists.

3.2 Addressing general forms of differential non-response

In practice wealth survey data is likely to suffer from more general forms of differential non-response. While the phenomenon is well documented for wealth surveys, only few countries implement convincing oversampling strategies in order to deal with it[14]. To illustrate how the rank correction approach can provide more reliable estimates of the top tail of the wealth distribution in the context of general forms of differential non-response, we conduct a Monte Carlo simulation which makes use of a response mechanism which expresses the probability of a household to respond (R) as a function of its wealth (x). Our starting point is the mechanism

$$R(x_i) = 0.903 - 0.036594 \ln(x_i) \tag{11}$$

, which is based on data from Kennickell & Woodburn (1997), who exploited the fact that the SCF uses high quality tax data to design the sample which allows for a comparison of ex ante information on wealth with ex post data on response behaviour. As such, this mechanism has a solid empirical basis[15]. We go one step further and combine this mechanism with the binary differential non-response mechanism $R_1(x_i)$ from the previous section. Thus for our Monte Carlo simulation we use the following non-response mechanisms:

$$R_2(x_i) = \begin{cases} 0.903 - 0.036594 \ln(x_i), & \text{for } x_{min} \leq x_i < x_{\text{NO}} \\ 0, & \text{for } x_{\text{NO}} \leq x_i \leq x_{max} \end{cases} \tag{12}$$

where R_2 is the response probability of household i and x_i is that household's net wealth. As before, x_{NO} denotes the level of wealth of the poorest non-observed (NO) household that it is excluded from the sampling frame due to privacy concerns and x_{max} is the richest household in

[14] Good practice examples include the US (SCF), France (HFCS), Spain (HFCS) and recently Germany (SOEP).

[15] See also (Vermeulen 2018). However, concerns regarding the extent to which this mechanism applies universally across countries and time remain. This is an important area for further research and depends crucially on access to tax data to allow for high quality oversampling strategies.

the population. This response mechanism $R_2(x_i)$ represents a situation where the sampling procedure suffers from differential non-response and the richest 1,600 households in the population are excluded due to privacy concerns as in the previous section. The main feature of response mechanism R_2 is that since the response probability is a logarithmic function of household net wealth, it falls off rather slowly. For example the maximum response probability of 40% (which corresponds to the poorest household in the population) is very close to the average of 37% across the entire population. However, for a very small number of households at the very top response rates are substantially lower. It is exactly this pattern, where response rates fall off very sharply in the tail that is also observed in practice.

Table (2) presents simulation results when using response mechanism R_2. We performed Monte Carlo simulations for different shape parameters (Pareto alphas, column 1) and net sample sizes (column 2) for the raw data approach (column 3), the Pareto correction approach (columns 4 and 5), the rank correction approach (columns 6 to 8) and the rich list correction approach (columns 9 to 12). Columns 4, 6, 9 and 12 contain the mean of the estimated shape parameter, calculated over 1,000 draws from the population. Columns 3, 5, 7, 10 and 12 contain the median percent deviation of estimated aggregate wealth from the true aggregate. Estimated aggregate wealth is either based on the raw data (column 3) or the estimated shape parameters. As in the previous section we analysed two rich list scenarios. The first is one in which the researcher observes a short rich list (SRL) containing only the most affluent 5 households (columns 9 and 10). For the second rich list correction approach scenario we assume the researcher observes a long rich list (LRL) with the most affluent 100 households (columns 11 and 12).

For example in row 1 of Table 2 we see the results based on simulations with a shape parameter of $\alpha = 1.5$ and a sample size of 0.03‰ with the following results: Only using the raw sample data underestimates total wealth in the tail by 18.8%, using the PC approach yields an average alpha estimate of $\hat{\alpha} = 1.642$ and an underestimation of the tail of 14.0%, using the RC approach yields an average alpha estimate of $\hat{\alpha} = 1.572$ based on median correction factor $u = 2200$ and an underestimation of tail wealth of 7.7%, using the SRL approach yields an average estimate of $\hat{\alpha} = 1.499$ and an overestimation of tail wealth by 0.1% and using the LRL approach yields an average estimate of $\hat{\alpha} = 1.489$ and an overestimation of tail wealth by 1.6%. The following three rows show simulation results which are based on the same shape parameter ($\alpha = 1.5$) but increasing sample sizes, from 0.08‰to 0.6‰. The following rows provide simulation results for the same sample sizes but with different shape parameters $\alpha = 1.25$ and $\alpha = 1.75$.

While Table (2) reveals that the rank correction approach is not able to fully correct the bias induced by the combined response mechanism R_2, it still consistently outperforms the raw data (column 3) and the Pareto correction (PC) approach (column 5). For example, for a population shape parameter of $\alpha = 1.25$ and a net sample size of 0.2‰, the raw data approach underestimates aggregate wealth by 36.4%, and the PC approach yields a mean alpha estimate of 1.392 and thus underestimates total wealth by 28.7%. In comparison the RC approach yields an average alpha estimate of 1.354 and underestimates the tail by 23.4%. This pattern holds across all sample sizes and shape parameters. When the researcher has access to short rich

Table 2: Simulation results for response mechanism R_2

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
true	sample	data	PC	PC	RC	RC	RC	SRL	SRL	LRL	LRL
alpha	size	WD	$\hat{\alpha}$	WD	$\hat{\alpha}$	WD	u	$\hat{\alpha}$	WD	$\hat{\alpha}$	WD
1.5	0.03	-18.8	1.642	-14.0	1.572	-7.7	2200	1.499	0.1	1.489	1.6
	0.08	-18.4	1.642	-14.5	1.587	-9.6	1500	1.538	-4.7	1.496	0.7
	0.20	-18.2	1.643	-14.6	1.597	-10.6	1300	1.58	-9.1	1.508	-1.1
	0.60	-18.2	1.648	-15.2	1.607	-11.6	1100	1.62	-12.9	1.539	-4.9
1.25	0.03	-37.4	1.393	-28.3	1.334	-19.5	2200	1.243	2.30	1.238	4.2
	0.08	-37.0	1.391	-28.5	1.346	-21.8	1400	1.281	-8.9	1.244	2.0
	0.20	-36.4	1.392	-28.7	1.354	-23.4	1200	1.325	-18.2	1.257	-2.1
	0.60	-36.4	1.392	-28.9	1.359	-24.2	1000	1.363	-24.7	1.286	-10.
1.75	0.03	-11.4	1.899	-9.0	1.820	-4.8	2100	1.754	-0.3	1.740	1.0
	0.08	-10.9	1.899	-9.1	1.835	-5.7	1500	1.794	-3.1	1.746	0.4
	0.20	-11.1	1.905	-9.7	1.851	-6.7	1300	1.840	-6.1	1.760	-0.8
	0.60	-11.0	1.903	-9.6	1.854	-6.9	1200	1.876	-8.2	1.792	-3.0

Simulation results based on different population alphas (column 1) and sample sizes (column 2) for the raw data, Pareto correction (PC), rank correction (RC) and rich list correction (SRL and LRL) approaches. For all approaches we report the median percent deviation from true tail wealth (WD); for the latter three approaches we also show the average estimate for α ($\hat{\alpha}$). Results are based on 1,000 draws per sample size. Sample sizes are net sample sizes. For the RC approach the median of the chosen u's is reported in column 8. Tail population $N = 10^6$ and $x_{min} = 10^6$.

lists (SRL) with 5 entries (columns 9 and 10), the rich list correction approach dominates the RC approach except for the largest sample size. This is due to the fact that as the sample size increases the rich list becomes less important relative to the number of observations in the sample and thus helps less to correct the differential non-response bias. Thus for large sample sizes, the RC approach produces aggregate wealth estimates which are closer or similarly close to the true aggregate compared to the SRL approach. When the researcher has access to a long rich list (LRL) of 100 observations, very precise estimates of the shape parameter (column 11) and the tail aggregate (column 12) are obtained, especially for the two smaller sample sizes. Similar to the SRL approach, the performance of the LRL approach declines with larger samples because the amount of additional exogenous information in the form of the rich list declines in relation to the number of observations in the sample.

Taken together the rank correction approach is not a *deus ex machina* which is able to resolve all forms of non-response problems. Nevertheless, it represents a robust improvement over the raw data and standard Pareto correction approach and is equivalent or even superior to the rich list approach in certain contexts (e.g. when richlists are short and/or erroneous as well as in the context of large samples). Crucially, this improvement comes at little additional cost, as the implementation of the rank correction approach is quite simple and does not rely on additional information such as journalists' rich lists. This lower information requirement makes the rank correction approach feasible in situations where the rich list approach would struggle for example due to limited rich list data availability or situations where rich list data is not available at all as is the case for the distribution of income.

4 Application to Wealth Survey Data

In this section the rank correction approach as laid out in sections 2.3 and 2.4 is applied to empirical survey data and compared to other approaches towards assessing the tail of the data. In terms of data, we focus on the second wave of the Household Finance and Consumption Survey (HFCS), the 2013 wave of the Survey of Consumer Finances (SCF) and the fourth wave (2012-2014) of the UK's Wealth and Asset Survey (WAS). We use the aggregate measures of net wealth from the HFCS (variable DN3001), the SCF (variable networth; SCF summary dataset) and the WAS[16].

4.1 Estimation results

To apply the different correction methods to survey data we first have to choose some scale parameter x_{min}, above which wealth is assumed to follow a Pareto distribution. For the majority of countries we choose x_{min} so that the most affluent 3% of all households in the samples are subjected to this assumption. Exceptions are given by the US, France and Spain, which all have extraordinarily high oversampling rates. In these cases we only analyze households within the richest percentile. After estimating the respective shape parameter α, the data in the chosen upper segment of the data greater than x_{min} can be replaced by an estimate derived from the fitted Pareto distribution.

Table 3 compares the estimates that emerge from different correction approaches: columns (1) to (3) report the estimated shape parameter $\hat{\alpha}$ for the three different approaches that rely on such an estimate. Similarly, columns (5) to (7) contain the ratio of total net wealth after correcting the data relative to the total emerging from the raw survey data, which is shown in absolute terms in column (4). The correction factors u obtained when applying the algorithm described in section 2.4 are reported in column (8), while column (9) reports the number of entries on the Forbes billionaire list, which is the rich list we used for estimating columns (3) and (7). Finally column (10) reports he number of observations with net wealth beyond 10 million in the sample. We report plots of the fitted Pareto tails in the Appendix.

By comparing columns (5)-(7) we can infer some structural features of the approaches under study: first, all correction methods suggest that estimated net wealth based on the corrected data is greater than or equal to the corresponding estimate from the raw survey data. Only a

[16] The WAS exhibits two important differences compared to the HFCS and the SCF: First, it does not provide information on the value of privately held businesses due to a high number of missing answers. This is a serious shortcoming which limits its ability to provide comprehensive information about the tail of the wealth distribution. Second, it also includes model based estimates of future pension claims. The amount of wealth the WAS adds in the form of claims on future pensions is substantial. For example the number of millionaires in wave 4 based on TotWlthW4 amounts to 2.75 million while there are 887,209 millionaires based on the net wealth variable which excludes pension wealth. To make our wealth measure comparable across surveys, we exclude pension wealth in the WAS and define netwealth as (TotWlthW4 − TOTPENw4_aggr).

Table 3: Pareto tails for SCF, WAS and HFCS survey data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Country	α PC	α RC	α RL	raw	PC/raw	RC/raw	RL/raw	U	RL	10^7
US_{2013}*	1.789	1.671	1.496	66.8	0.98	1.02	1.10	360	442	598
$\text{UK}_{2012\text{-}14}$	1.958	1.949	1.547	6.59	1.00	1.00	1.10	100	37	11
Austria	1.404	1.390	1.342	1.00	1.10	1.11	1.13	100	10	4
Belgium	2.348	1.403	1.687	1.58	0.99	1.22	1.08	11,780	3	0
Cyprus	1.644	1.427	1.211	0.12	1.09	1.58	1.61	322	4	3
Germany	1.597	1.397	1.340	8.50	1.00	1.13	1.19	10,760	85	10
Spain*	1.724	1.334	1.582	4.77	1.00	1.12	1.02	1,560	16	153
Finland	2.140	2.012	1.731	0.51	1.00	1.02	1.07	100	4	4
France*	1.525	1.423	1.351	7.03	1.03	1.06	1.10	1,200	43	90
Greece	3.382	2.146	1.401	0.44	0.98	1.05	1.27	10,300	3	0
Italy	2.417	2.288	1.372	5.59	1.00	1.01	1.26	1,300	35	2
Luxem.	1.578	1.306		0.16	1.04	1.22		100		12
Malta	1.291	1.168		0.06	1.32	1.52		64		2
Nether.	3.082	1.867	1.419	1.15	0.99	1.08	1.24	22,800	6	0
Portugal	2.259	2.183	1.669	0.63	1.02	1.03	1.13	100	3	4
Slovenia	1.267	1.126		0.11	1.41	1.71		300		2
Slovakia	1.959	1.836		0.12	1.03	1.05		100		0
Estonia	1.727	1.607		0.06	1.01	1.05		100		1
Hungary	1.719	1.552		0.21	1.02	1.07		900		0
Ireland	2.396	2.304	1.403	0.37	1.00	1.01	1.29	100	5	1
Latvia	1.806	1.106		0.03	0.98	2.38		1,604		0
Poland	2.193	2.027	1.463	1.31	1.00	1.02	1.16	1,920	5	0

Columns 1 to 3 report the estimated shape parameters based on the Pareto Correction (PC), Rank Correction (RC) and Rich List (RL) approach. Column 4 contains aggregate net wealth in trillion USD¡ GPB and EUR for the US, UK and EU countries. Column 8 contains the chosen correction factor u for the RC apparoch and column 9 contains the number of entries on the 2014 Forbes global list of billionaires used for the RL approach. Columns 5 to 7 contain the ratio of aggregate wealth after adding the estimated tail to aggregate wealth based on raw survey data. Column 10 reports he number of observations with net wealth beyond 10 million in the sample. *For the US, Spain and France x_{min} is defined as the 99th percentile due to their high oversampling rates. For the remaining countries x_{min} is defined as the 97th percentile.

few applications of the weakest correction – the simple Pareto correction (PC) – actually would suggest that true wealth is below the value implied by the raw survey data. In a similar vein, the simple PC correction typically suggests rather modest corrections of the underlying data. Second, the rich list approach can only be applied in cases, where a rich list exists, which is why columns (3) and (7) show missing values for those countries, which are not represented in the Forbes billionaires list. Third, the rank correction approach quite consistently suggests corrections that lie between the simple Pareto correction and the richlist approach. Hence, it occupies some middle ground in empirical terms that might lead to significantly higher estimates of total wealth, but typically not as high as a richlist approach towards correcting the data would suggest. Hence, these findings are qualitatively well in line with the results obtained in section 3 and confirm the impression that the rank correction approach might offer a significant improvement over both, an approach based on raw data only as well as a simple Pareto correction approach.

More generally we observe that the results in Table 3 show a large amount of heterogeneity. This reflects, among other things, significant differences in sample sizes, country wealth and the extent to which survey administrators were able to observe the right tail of the wealth distribution. The latter factor is especially pronounced in countries where $\alpha\hat{_{PC}}$ is greater than or close to 2, such as UK, Belgium, Germany, Finland, Italy, Greece, Netherlands and Poland. Closer inspection shows that in these countries the upper tail is not well represented by the data as evidenced by, e.g., the small number of observations with net wealth in excess of 10 million (column 10). Given that the standard Pareto correction approach (column 5) often leads to no correction at all (ratio of 1), or even implies a thinner tail than the raw data (ratios of less than 1) in those cases, we interpret this as an indication that the Pareto correction approach requires a minimum amount of observations in the tail to work well. Put differently, if the differential non-response problem is too extreme, not enough observations are left to trace out the tail of the Pareto distribution. Third, in line with the previous point, the rank correction approach has a tendency to correct the survey data especially in those countries where the PC approach yielded an upwards correction already. Examples include Austria, France, Luxembourg, Slovakia or Hungary. Since the RC approach does not add additional data to the sample it faces the same shortcomings as the PC approach in cases where very few tail observations are available.

The US result is particularly noteworthy with respect to the rank correction approach, because we know that the SCF data excludes the richest 400 Americans. In line with this, the RC approach selects a correction factor of $u = 360$ which is very close to the expected result as we would not expect substantially higher correction factors for the US sample due to its high quality oversampling.

Table 4 reports top wealth shares based on the raw survey data and after the tail had been replaced by an estimated Pareto distribution using all three correction approaches. Corrected top wealth shares increase substantially for countries like Belgium, Austria and Germany, which are those countries for which in Table 3 the aggregate estimates were corrected substantially. These results are a reminder of how sensitive distributional measures such as top wealth shares are to differential non-response problems. The rank correction approach represents a simple but still highly useful tool to (partially) correct it post data collection. The fact that some countries exhibit unrealistically low top wealth shares based on raw survey data and rank corrected data such as Greece or the Netherlands, indicates that the practice of fitting Pareto tails to survey data is still fundamentally dependent on the quality of the underlying data. The more of the tail is already missing, the more difficult it becomes to fit a meaningful Pareto tail to the data. The rank correction approach is not able to resolve this fundamental problem. It is in these situations of very limited tail data where the rich list approach is most useful as it aims to directly compensate for missing data at the top of the distribution.

Table 4: Household top 1% wealth shares

Country	raw	PC	RC	RL
US_{2013}	35.4%	34.1%	36.2%	40.6%
$\text{UK}_{2012\text{-}2014}$	15.1%	14.8%	14.9%	21.6%
Austria	25.4%	29.9%	30.6%	31.8%
Belgium	12.0%	11.6%	26.0%	17.9%
Cyprus	20.3%	25.5%	36.5%	43.7%
Germany	23.6%	25.3%	32.6%	35.6%
Spain	16.3%	16.1%	24.2%	17.9%
Finland	13.3%	13.4%	14.5%	17.9%
France	18.7%	20.7%	23.1%	25.4%
Greece	9.2%	9.1%	14.0%	26.9%
Italy	11.7%	11.6%	12.3%	27.8%
Luxembourg	18.8%	20.7%	30.1%	
Malta	19.9%	30.9%	36.8%	
Netherlands	9.8%	9.6%	16.1%	25.8%
Portugal	14.4%	15.4%	16.0%	22.8%
Slovenia	22.9%	33.3%	42.5%	
Slovakia	9.3%	11.2%	12.3%	
Estonia	21.2%	22.3%	24.9%	
Hungary	17.2%	18.6%	22.2%	
Ireland	14.2%	14.9%	15.5%	31.7%
Latvia	21.4%	23.3%	56.5%	
Poland	11.7%	12.1%	13.3%	22.7%

Household net wealth shares based on raw survey data and Pareto correction methods, expressed in % of total aggregate household wealth.

4.2 Reconciliation with other data sources

While the results in the previous subsection suggest that the rank correction approach delivers empirically plausible results that represent an improvement over raw survey data as well as a simple Pareto correction, a firmer conclusion about the quality of the rank correction approach in particular can be drawn when comparing our results with other exogenous data sources. The two crucial sources of exogenous information against which we compare our results are first, the World Inequality Database (WID) as well as a new dataset for Germany (Schröder et al. 2020) and second, journalists' rich lists for individual countries.

The methods used to construct the WID series for the US are discussed in Piketty et al. (2016) and the accompanying data appendix (Tables II-E1 to E13 contain the wealth share estimates). The country specific details for applying this methodology to France are discussed in Garbinti et al. (2016) and in the accompanying appendices. The methods used for the UK series are discussed in Alvaredo et al. (2018), its working paper version and the online appendix.

The most important difference between the WID concentration measures and the survey based concentration measures is that the former are based on net personal wealth, which means that the unit of analysis is the individual instead of the household. One of the first steps of the WID methodology is to split married couples in survey or tax data into two observations with equal net wealth shares. This means some differences in the results stem from these methodological differences.

Against this background, Table 5 compares wealth concentration ratios from the WID (rows 1, 4 and 7) with the results from the rank correction approach (rows 2, 5 and 8) and raw survey based measures (rows 3, 6 and 9). The rank correction results for France, the US and Germany clearly represent an improvement over the raw survey data and are closer to the WID measures than the raw counterparts. In the case of France the RC measures are very well in line with WID values except for the top 0.1% share. For the US case, the rank correction based top shares are also very close to WID results except for the top 0.1%. For the UK, the WID does not provide an entry for the top 0.1% share in 2012. The comparison of the UK data to the WID emphasizes the underestimation of the tail of the wealth distribution in the raw survey. The shortcoming of the WAS to adequately capture privately held business wealth seems to be a fundamental problem which RC cannot do anything about. For Germany the estimated top 1% share is firmly in line with the measure provided by Schröder et al. (2020). The advantage of the latter datasource is that it employs a highly promising oversampling strategy based on share holdings. Overall the RC based measures help to close the gap between the WID and the raw survey measures which we interpret as support for the rank correction approach.

Table 5: Top wealth shares: WID vs rank correction

			(1)	(2)	(3)
	country	data and method	top 0.1%	top 1%	top 10%
(1)	France	World Inequality Database	8.2	23.4	55.3
(2)	France	rank correction estimator	11.6	23.1	53.4
(3)	France	uncorrected survey data (HFCS)	7.3	18.7	50.8
(4)	USA	World Inequality Database	20.3	37.0	73.2
(5)	USA	rank correction estimator	14.4	36.2	75.3
(6)	USA	uncorrected survey data (SCF)	13.1	35.4	75.0
(7)	UK	World Inequality Database		19.9	51.9
(8)	UK	rank correction estimator	4.8	14.9	45.7
(9)	UK	uncorrected survey data (WAS)	5.6	15.1	45.7
(10)	DE	Schröder et al. (2020)		35.3	67.3
(11)	DE	rank correction estimator	16.8	32.6	64.0
(12)	DE	uncorrected survey data (HFCS)	6.3	23.6	59.8

Source: Authors' computations based on data from the Household Finance and Consumption Survey (HFCS), Survey of Consumer Finances (SCF), Wealth and Asset Survey (WAS) and the World Inequality Database (WID). Comparison of French, US and UK wealth shares for the years 2014 and 2013 and 2012-2014 respectively.

As has been emphasized, another alternative source of information about the top tail of the wealth distribution are journalists' rich lists. Table 6 column (1) lists the number of billionaires in the population according to the raw survey data (i.e. billionaire observations times their weight) which indicates that no country except the US has an oversampling strategy in place which is suitable to capture billionaires[17]. Column (5) reports the number of billionaire entries on Forbes global rich list from 2014. Columns (2(to (4) report the number of billionaires according to the estimated Pareto tail using the three correction approaches under investigation. The Pareto correction approach in column (2) almost always yields a number of billionaires which is far below the number of entries on the Forbes rich list. This is in line with the results from the Monte Carlo simulations, that the PC approach underestimates the tail of the distribution in a situation of differential non-response. For most countries the rank correction approach implies fewer billionaires than are reported on rich lists (exceptions are Austria, Belgium and Spain as well as a few very small countries). We interpret this result as general support for the claim that the rank correction approach is a rather conservative tool for addressing the non-response bias in survey data as the results it provides probably still underestimate the actual degree of concentration. The probable underestimation of the RC approach becomes apparent for countries like Italy, the Netherlands and the UK, all three of which do not have an oversampling strategy in place. The RL approach improves upon some of these results with two caveats. First, the RL approach might overestimate total wealth in cases, where a high-quality oversampling strategy is already in place. The US is an example for such a case: here the RL approach implies almost twice as many billionaires as on the Forbes list. Second, the RL approach is not feasible for many countries for which no rich list data is available.

5 Summary and Conclusion

For many countries, household surveys are the only available source of data on the distribution of wealth. Against this background, this paper presents a new approach, which we labelled the rank correction approach, for tackling the problem that surveys tend to underestimate household wealth due to differential non-response. The key advantage of the rank correction approach over similar existing methods such as the use of rich lists, is that it requires no additional information beyond the survey data. This is especially important since rich lists are not necessarily accurate (Capehart 2014, Kopczuk 2015) and often not available. Hence, the rank correction approach serves not only as a substitute for the richlist approach, but also as a complement in the form of a robustness check applicable in cases where rich list data is available.

Applying the rank correction approach to data from the SCF and the HFCS results in significant corrections of top wealth shares. Our results are more closely aligned with other existing top wealth share estimates than the raw survey estimates. For example the WID which relies on tax data and is thus less prone to differential non-response, provides top 1% wealth shares for the US and France of 37% and 23.4%, respectively. The rank correction approach by comparison yields 36.2% and 23.1%, representing a clear improvement over raw survey estimates

[17] Another factor is the larger size of the US.

Table 6: Number of billionaire households

Country	(1) raw	(2) PC	(3) RC	(4) RL	(5) Forbes
US_{2013}	43	220	391	904	492
$\text{UK}_{2012\text{-}2014}$	0	1	1	19	47
Austria	0	11	12	12	10
Belgium	0	0	16	2	3
Cyprus	0	0	4	5	4
Germany	0	20	80	117	85
Spain	0	4	44	8	26
Finland	0	0	0	0	4
France	0	19	36	56	43
Greece	0	0	0	3	3
Italy	0	0	0	55	35
Luxembourg	0	1	3	NA	0
Malta	0	1	2	NA	0
Netherlands	0	0	0	7	7
Portugal	0	0	0	1	3
Slovenia	0	3	4	NA	0
Slovakia	0	0	0	NA	0
Estonia	0	0	0	NA	0
Hungary	0	0	0	NA	0
Ireland	0	0	0	4	5
Latvia	0	0	2	NA	0
Poland	0	0	0	4	5

Forbes refers to the Forbes list of billionaires from 2014.

of 35.4% and 18.7%. The rank correction approach also allows us to reproduce recent estimates of the top 1% wealth share for Germany by Schröder et al. (2020) who implement a highly promising oversampling strategy based on share holdings. The latter estimate a top 1% share of 35.3% compared to our 32.6% based on the rank correction approach.

Another important question is for what kind of applications the rank correction approach might be used other than wealth surveys. In this context we have noted that differential non-response also occurs in the context of surveys aimed at measuring the distribution of income or wages. However, income-based equivalents to rich lists are hard to come by, which means the rank correction approach can be a valuable tool for analyzing tails in income survey data. Obvious examples for the potential merit of such an approach include empirical tax simulation models such as EUROMOD [18] or TAXBEN (Giles & McCrae 1995) for which a fitted Pareto tail can improve the tail coverage of the wage, income or wealth data used. Another potential application is the growing literature on distributional national accounts (DINA) which aims

[18] See https://euromod-web.jrc.ec.europa.eu/.

to produce aggregate macro time series which are consistent with distributional micro data (Piketty et al. 2016, Saez & Zucman 2016). Moreover, the example of the US – in which the rank correction approach closely estimates the number of missing households excluded from the survey – shows that the rank correction approach might also be useful in cases, where data on income or wages is capped at a certain level. The latter is often the case for administrative data (e.g. social security records) or survey data. Moreover, the rank correction approach could be of help to infer information about the tail in cases where income is collected by assigning households to specific income brackets.

While we think this shows that the rank correction approach is an important improvement over simply ignoring differential non-response problems, it is of the utmost importance to address the root cause of the problem as part of the data collection by improving oversampling strategies in existing wealth surveys. From a European perspective, the introduction of the HFCS (2011) and WAS (2006) were massive steps forward, but political reluctance to provide the national central banks with tax information to implement effective oversampling strategies unnecessarily undermines data quality[19]. While granting access to tax information is a sensitive issue, the success of the SCF demonstrates that it is feasible. Until these improvements are made, the rank correction approach provides a simple way to address the persisting problem of under-representation of highly affluent households in survey data.

References

Advani, A., Bangham, G. & Leslie, J. (2020), 'The uk's wealth distribution and characteristics of high-wealth households', *Wealth Tax Commission Evidence Paper no. 1* .

Aigner, D. J. & Goldberger, A. S. (1970), 'Estimation of pareto's law from grouped observations', *Journal of the American Statistical Association* **65**(330), 712–723.

Alstadsæter, A., Johannesen, N. & Zucman, G. (2019), 'Tax evasion and inequality', *American Economic Review* **109**(6), 2073–2103.

Alvaredo, F., Atkinson, A. B. & Morelli, S. (2018), 'Top wealth shares in the uk over more than a century', *Journal of Public Economics* **162**, 26–47.

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

Bricker, J., Henriques, A., Krimmel, J. & Sabelhaus, J. (2016), 'Measuring income and wealth at the top using administrative and survey data', *Brookings Papers on Economic Activity* **2016**(1), 261–331.

Capehart, K. W. (2014), 'Is the wealth of the world's billionaires not paretian?', *Physica A: Statistical Mechanics and its Applications* **395**, 255–260.

[19] In addition, the approach of Schröder et al. (2020) might offer another promising approach to oversampling that could be applied to a greater number of countries.

D'Alessio, G. & Faiella, I. (2002), 'Non-response behaviour in the bank of italy's survey of household income and wealth', *Temi di discussione (Bank of Italy Economic working papers)* **2002**(462).

Eckerstorfer, P., Halak, J., Kapeller, J., Schütz, B., Springholz, F. & Wildauer, R. (2016), 'Correcting for the missing rich: An application to wealth survey data', *Review of Income and Wealth* **62**(4), 605–627.

Gabaix, X. & Ibragimov, R. (2011), 'Rank - 1/2: A simple way to improve the ols estimation of tail exponents', *Journal of Business & Economic Statistics* **29**(1), 24–39.

Garbinti, B., Goupille-Lebret, J. & Piketty, T. (2016), 'Accounting for wealth inequality dynamics: Methods, estimates and simulations for france (1800-2014)', *WID Working Paper Series* **2016/5**.

Giles, C. & McCrae, J. (1995), 'Taxben: the ifs microsimulation tax and benefit model', *IFS Working Paper* (19).

Jayadev, A. (2008), 'A power law tail in india's wealth distribution: Evidence from survey data', *Physica A: Statistical Mechanics and its Applications* **387**(1), 270–276.

Kennickell, A. B. & Woodburn, R. L. (1997), CONSISTENT WEIGHT DESIGN FOR THE 1989, 1992 AND 1995 SCFs, AND THE DISTRIBUTION OF WEALTH, Technical report, Federal Reserve Board Survey of Consumer Finances Working Papers.

Kopczuk, W. (2015), 'What do we know about the evolution of top wealth shares in the united states?', *Journal of Economic Perspectives* **29**(1), 47–66.

Little, R. J. A. & Rubin, D. B. (2019), *Statistical Analysis with Missing Data*, John Wiley and Sons Ltd.

Osier, G. (2016), 'Unit non-response in household wealth surveys: Experience from the eurosystem's household finance and consumption survey', *European Central Bank Statistics Paper Series* **2016**(15).

Piketty, T. (2014), *Capital in the Twenty-First Century*, Harvard University Press.

Piketty, T., Saez, E. & Zucman, G. (2016), 'Distributional national accounts: Methods and estimates for the united states', *NBER Working paper* **22945**.

Saez, E. & Zucman, G. (2016), 'Wealth inequality in the united states since 1913: Evidence from capitalized income tax data', *The Quarterly Journal of Economics* **131**(2), 519–578.

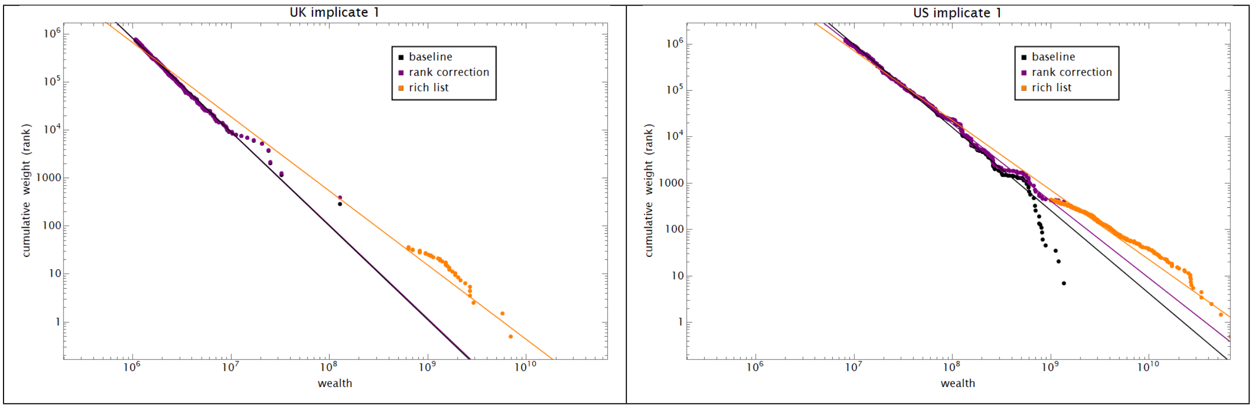
Schröder, C., Bartels, C., Göbler, K., Grabka, M. M. & König, J. (2020), 'Millionaires under the microscope: Data gap on top wealth holders closed; wealth concentration higher than presumed', *DIW Weekly Report* .

Vermeulen, P. (2018), 'How fat is the top tail of the wealth distribution?', *Review of Income and Wealth* **64**(2), 357–387.

Wildauer, R. & Kapeller, J. (2019), ‘A comment on fitting pareto tails to complex survey data’.

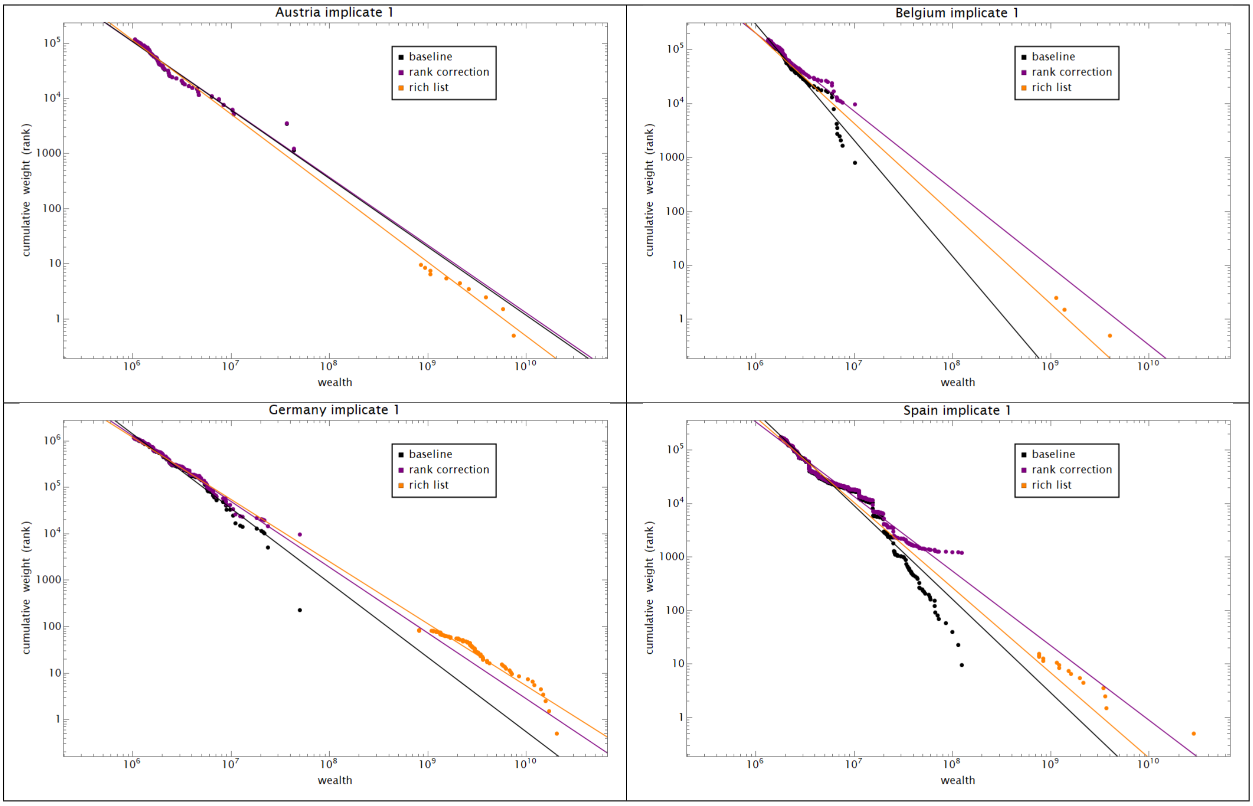
Appendix

Table A1: Regression results UK and US



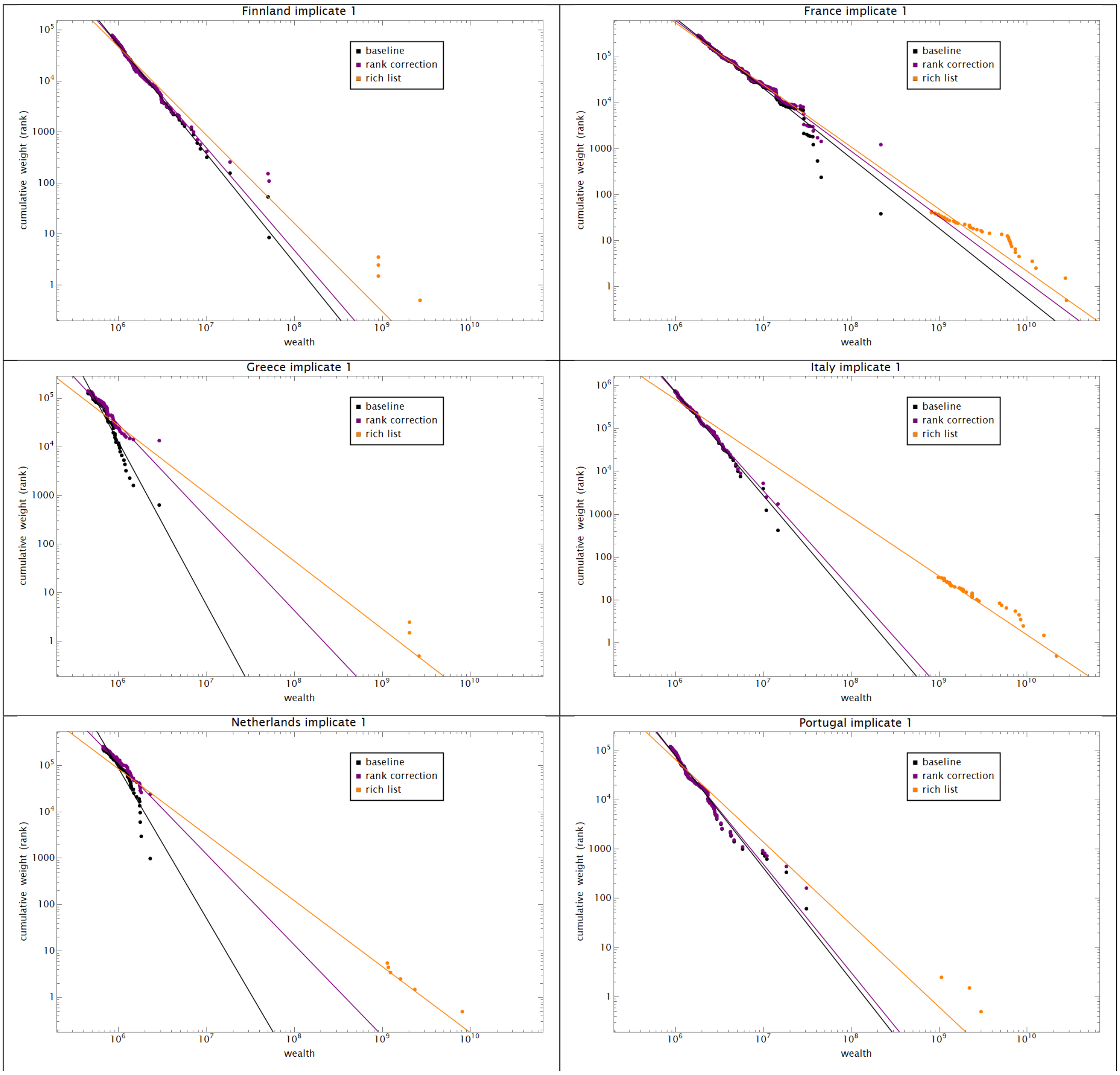
Results for the UK in the left panel and for the US in the right panel. Results from fitting pareto tails to the top 3% of households (US top 1%) using the RC, RL and baseline approaches. US results are based on the first implicate. Results in the paper are always based on all implicates.

Table A2: Regression results HFCS part I



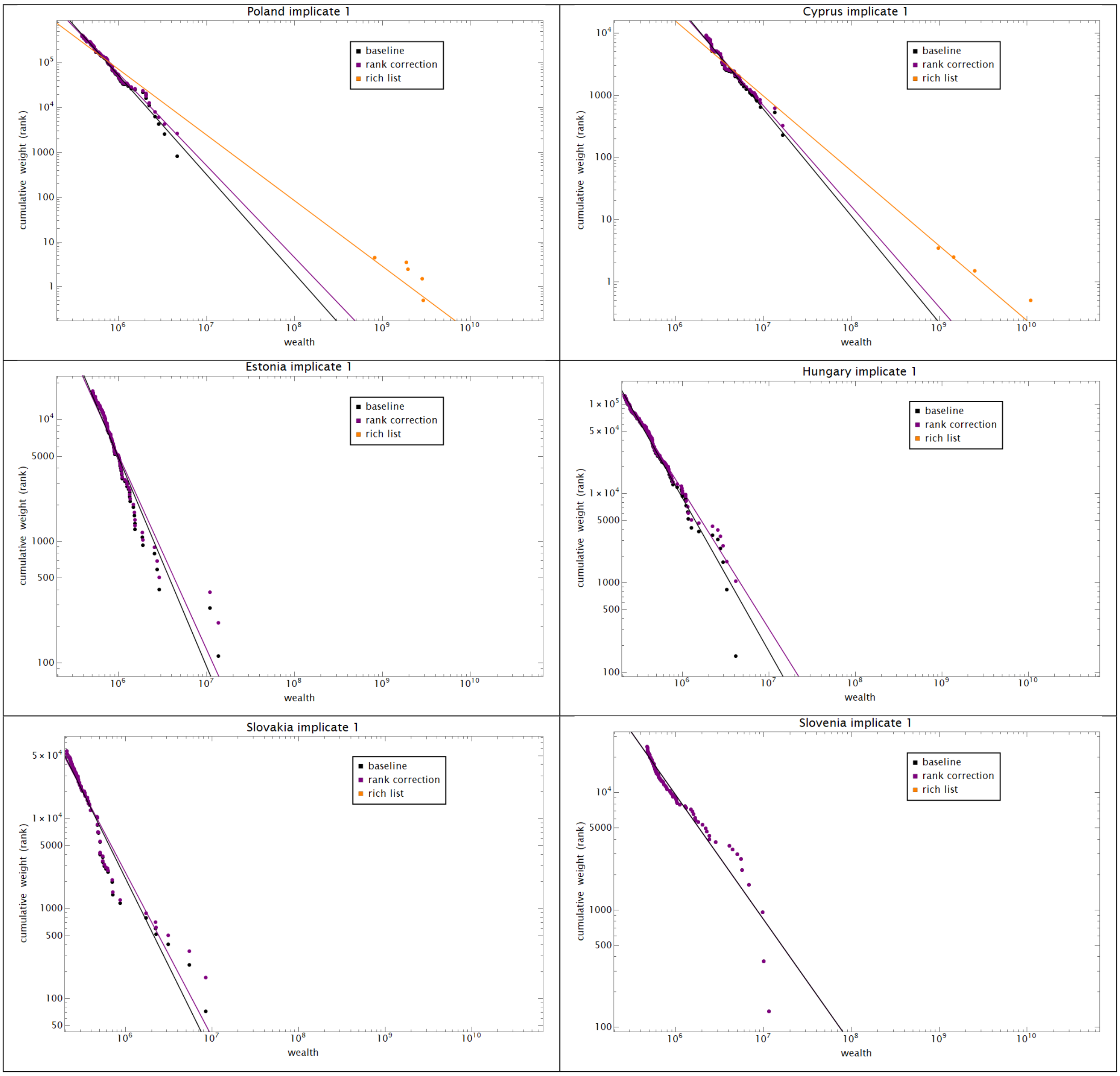
Results from fitting pareto tails to the top 3% of households (Spain top 1%) using the RC, RL and baseline approaches based on the first implicate. Results in the paper are always based on all implicates.

Table A3: Regression results HFCS part II



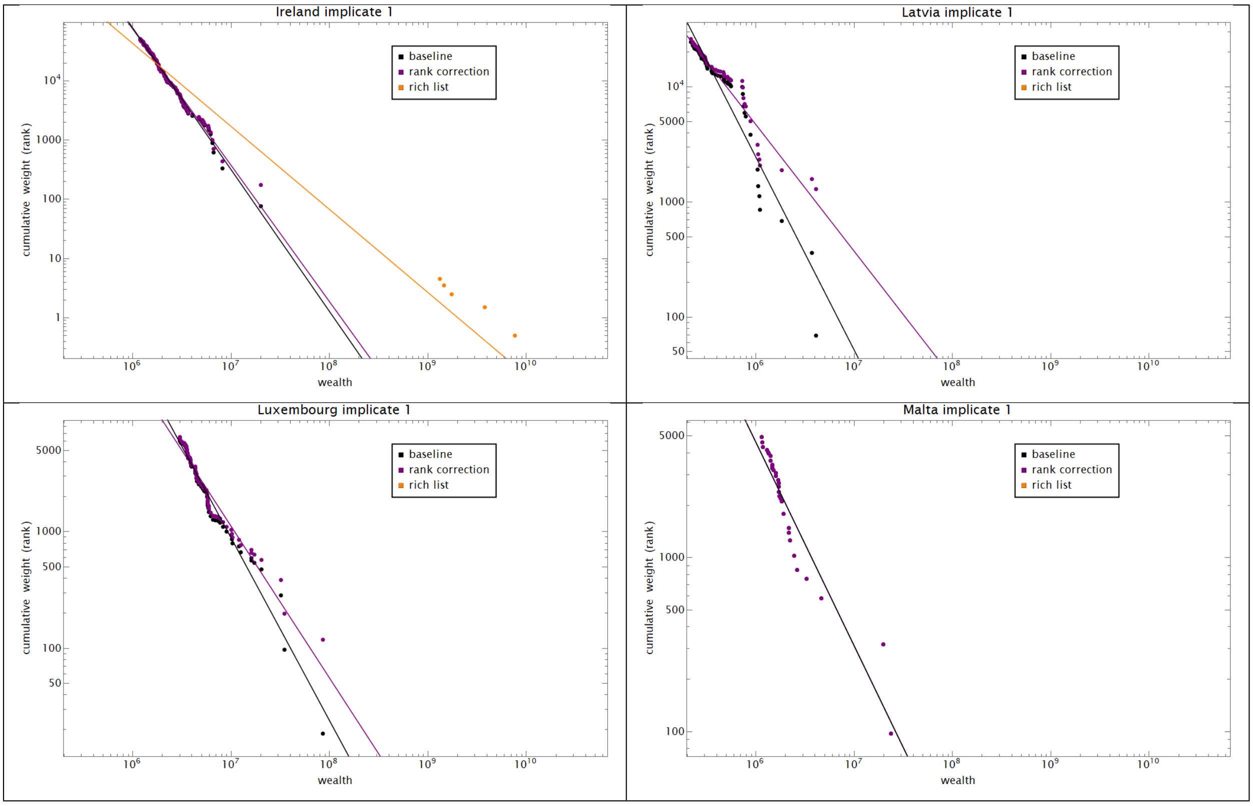
Results from fitting pareto tails to the top 3% of households (France top 1%) using the RC, RL and baseline approaches based on the first implicate. Results in the paper are always based on all implicates.

Table A4: Regression results HFCS part III



Results from fitting pareto tails to the top 3% of households using the RC, RL and baseline approaches based on the first implicate. Results in the paper are always based on all implicates.

Table A4: Regression results HFCS part IV



Results from fitting pareto tails to the top 3% of households using the RC, RL and baseline approaches based on the first implicate. Richlist data only available for Ireland. Results in the paper are always based on all implicates.