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

On Top of the Top. Adjusting Wealth Distributions Using National Rich Lists

Franziska Disslbacher

(Austrian Federal Chamber of Labor)

Michael Ertl

(Austrian Federal Chamber of Labor)

Emanuel List

(Vienna University of Economics and Business and Macroeconomic Policy Institute)

Patrick Mokre

(Austrian Federal Chamber of Labor and The New School)

Matthias Schnetzer

(Austrian Federal Chamber of Labor)

Paper prepared for the 36th IARIW Virtual General Conference
August 23-27, 2021
Session 2: The Potential and Challenges of Big Data and other Alternative Data in the Production of Prices, National Accounts, and Measures of Economic Well-Being
Time: Tuesday, August 24, 2021 [14:00-16:00 CEST]

Working Paper Series

#20

Franziska DISSLBACHER
Michael ERTL
Emanuel LIST
Patrick MOKRE
Matthias SCHNETZER

On Top of the Top - Adjusting wealth distributions using national rich lists

Original Citation:
Disslbacher, F., Ertl, M., List, E., Mokre, P., Schnetzer, M. (2020) On Top of the Top. Adjusting wealth distributions using national rich lists. INEQ Working Paper Series, 20. WU Vienna University of Economics and Business, Vienna.

ECONOMICS OF INEQUALITY
Research Institute
WU Vienna

Welthandelsplatz 1
1020 Wien
www.ineq.at

On Top of the Top. Adjusting wealth distributions using national rich lists

Franziska Disslbacher[a] · Michael Ertl[a] · Emanuel List[b,c] · Patrick Mokre[a,d*] · Matthias Schnetzer[a]

[a]Department of Economics, Austrian Federal Chamber of Labor
[b]Research Institute Economics of Inequality, WU Vienna
[c]Macroeconomic Policy Institute (IMK)
[d]The New School

December 23, 2020

Abstract

Poor coverage of the top in wealth surveys conceals the extent of wealth inequality. The literature mitigates this shortcoming by enriching survey data with rich lists and estimating the top tail with a Pareto distribution. However, recent studies rely on ad-hoc assumptions for some of the required parameters. We suggest a unified regression approach to estimate all parameters of a Pareto distribution jointly and extend our analysis with a more flexible three-parameter Generalized Pareto estimation. We introduce a new database of national rich lists (ERLDB) as an alternative to commonly used global rich lists to combine with survey data from the Household Finance and Consumption Survey (HFCS 2017). Our findings for 14 European countries show that wealth is more concentrated than surveys suggest, with almost doubling Top 1% shares in the most extreme cases. In contrast, countries with successful oversampling strategies tend to experience only minor changes in inequality metrics.

Keywords: Generalized Pareto estimation, national rich lists, missing rich, wealth shares, oversampling, HFCS
JEL Classification: C46, D31

1 Introduction

The distribution of wealth is infamously top-heavy, thus the richest parts of the population are crucial for a comprehensive picture of economic inequality. A profound understanding of the top tail is also important for economic policy design with regard to wealth taxation which usually affects richest households most. Survey data is the most important source of information in these matters but comes with flaws concealing the extent of inequality.

In this paper, we adopt Vilfredo Pareto's (1965 [1896]) intuition that inequality within segments increases with wealth and estimate the wealth of the richest percentiles in 14 European countries. We present a new unified approach to estimating all aspects of the standard two-parameter Pareto distribution as well as the three-parameter Generalized Pareto distribution without the need for arbitrary choices. To add very rich households to our analysis with Household Finance and Consumption Survey (HFCS) data, we introduce the most comprehensive data collection of journalistic rich lists so far, the European Rich List Database (ERLDB). Interpolating the top tail, we find that the wealth share of the Top 1% substantially increases in all, and almost doubles in the most extreme cases.

The literature on wealth inequality recognizes a striking pattern at the top percentiles, which closely resemble a Pareto distribution (Davies and Shorrocks, 2000). This distribution is characterized by the fact that inequality between observations increases with their relative wealth, such that wealth in the 99th percentile is more unequally distributed than in the 98th, which in turn is more unequal than in the 97th. As the availability and quality of administrative tax data deteriorated in past decades (Krenek and Schratzenstaller, 2018), the investigation of wealth distribution relies on surveys such as the Survey of Consumer Finances in the United States or the HFCS for Eurozone countries. However, the richest households are less likely to be captured correctly than their lower percentile counterparts due to (1) a higher likelihood to refuse participation (Kennickell and Woodburn, 1997) and (2) more complex financial portfolios favoring underreporting (Vermeulen, 2016b). This leads to a gap between aggregate survey wealth and assets recorded in National Accounts (Chakraborty and Waltl, 2018), as well as between the highest survey observations and journalistic evidence, such as the Forbes list of billionaires. The literature dubbed this the problem of "the missing rich".

The recent empirical literature approximates the tail of the distribution by using survey data as the lower and rich list observations as the upper bound to interpolate a Pareto distribution (Vermeulen, 2014). Vermeulen (2016a) finds that in eight European countries, the wealth share of the Top 1% is underestimated by between one (lower bound in Spain) and 11 percentage points (upper bound in Austria). More recent studies provide detailed results for a set of European countries based on the same methodological fundament (Eckerstorfer et al., 2016; Chakraborty and Waltl, 2018; Bach et al., 2019; Brzezinski et al., 2020). The estimation of a Pareto distribution hinges on two decisive parameters (location parameter w_{min} and shape parameter α), and a threshold w_0 that determines the wealth above which

observations become unreliable and are replaced by simulated new entries (Dalitz, 2016). The literature so far relied on best guesses or visual inspection for some of these parameters, which makes their methods hard to replicate with respect to time and location, and subject to methodological scrutiny.

In this paper, we combine techniques to estimate the parameters of Pareto distributions in linear regressions and for the first time provide a unified regression approach to estimate all aspects of the distribution jointly. We present results for the most comprehensive collection of countries so far by introducing a new European Rich List Database (ERLDB). Since we abstain from visual inspection or arbitrary parameter choices, our method is easily scaled to a large sample and seamlessly adapt to different survey designs or wealth regimes between countries. We furthermore extend the Pareto approach by estimating a three-parameter Generalized Pareto distribution, which is more flexible to varying inequality with increasing percentiles due to its additional scale parameter. Our unified approach closes the gap between survey and ERLDB data, and shows how severely surveys underestimate the wealth of the super-rich. The robustness of our approach is emphasized by the fact that inequality metrics hardly change in those countries that use administrative tax data or apply extensive oversampling in their surveys.

The remainder of the paper is organized as follows. Section 2 describes the HFCS survey data and introduces the ERLDB. Section 3 compares and combines estimation techniques to Pareto distributions in the literature and extends the methodology by using a Generalized Pareto distribution. Section 4 gives the results from both estimation procedures before section 5 concludes.

2 Data

Surveys on assets and liabilities like the Household Finance and Consumption Survey (HFCS) are valuable sources for the distributional analysis of household wealth. Despite enormous compiling efforts, the data is biased for two main reasons. First, wealth distributions are heavily skewed and a large fraction of total wealth is concentrated at the top. A small random sample thus may not adequately represent the full distribution of wealth when very rich households are not drawn into the sample. Second, the participation rate of households in wealth surveys decreases with wealth resulting in differential nonresponse (Davies and Shorrocks, 2000). Both lead to an under-representation of rich households in survey data, and thus an underestimation of total wealth and distributional indicators that focus on the top. We address these shortcomings of wealth surveys in section 2.2 and introduce the European Rich List Database (ERLDB) as a complementary data source. It covers information on the top of the wealth distribution for 20 European countries based on national rich lists. The combination of HFCS and ERLDB constitute the basis of our estimation strategy, which aims for a more comprehensive picture of aggregate wealth and wealth inequality. In total, we are able to include 14 countries for which both HFCS and ERLDB data are available.

2.1 Household Finance and Consumption Survey (HFCS)

The Household Finance and Consumption Survey (HFCS) is a harmonized survey of household finances coordinated by the European Central Bank (ECB). Its third wave was mainly conducted in 2017 by national central banks and covers all 19 Eurozone countries as well as Croatia, Hungary, and Poland. The survey sample of nearly 28,000 observations represents 167 million households and provides detailed information on real assets, financial assets, as well as liabilities (ECB, 2020). For many countries in Europe, the HFCS is the first and only, for others the most comprehensive micro dataset on wealth that enables distributional analysis.

The preferred survey mode of the HFCS is based on computer assisted personal interviews (CAPI) supported by consistency checks and automatic data storage. Only Poland, Finland, and the Netherlands decided to conduct other non-face-to-face survey modes. Participating households may refuse to answer difficult or sensitive questions which leads to item nonresponse. Mis-information about the financial situation or the wish to conceal, on the other hand, might lead to factually wrong answers, i.e. under- or overreporting. Both item nonresponse and misreporting are problematic if they are not uniformly distributed as this results in systematic biases in statistical analysis. The HFCS surveyors counter these problems with a multiple imputation strategy where missing and implausible values are estimated five times. The differences between the five sets of observations account for the underlying level of imputation uncertainty (ECB, 2020).

Despite various efforts to comprehensively cover the whole population, Davies and Shorrocks (2000) point out that wealth surveys struggle to accurately describe the top of the distribution. Many reasons for nonresponse apply to all sampled households (survey nonresponse) and can partially be addressed by adjusting survey weights *ex post*. However, there is evidence that unit nonresponse is correlated with household wealth and thus results in differential unit nonresponse (Kennickell and Woodburn, 1997; D'Alessio and Faiella, 2002). The reasons why it is more difficult to contact wealthy respondents are manifold (ECB, 2020): they are more likely to be absent for longer time periods, live in several residences, and are more able to protect their privacy. Furthermore, perceived and actual time restrictions of wealthy respondents and the reluctance to disclose information on the financial situation increase their refusal rate. Against this backdrop, it is necessary to oversample affluent households in wealth surveys to adequately capture the whole distribution (Kennickell, 2008; Bricker et al., 2016; Pfeffer et al., 2016).

An efficient survey design includes a disproportionally high number of wealthy households *ex ante* in the sampling frame. Most of the countries participating in HFCS implemented such oversampling strategies to correct for differential unit nonresponse. The success heavily depends on the ability to identify and reach wealthy households in order to actually interview them. Oversampling strategies fall back on individual or group-wise information correlated with wealth and influence the sample design as well as post-survey adjustments. While

some countries, for example France, use individual information on personal wealth from register data, others rely on proxies like income (Finland), electricity consumption (Cyprus), or dwelling size (Portugal). Countries without access to individual data consider regional information like property prices to oversample wealthy street sections, like Germany and Slovakia (ECB, 2020). However, oversampling of households with estimated wealth above a certain threshold does not ensure accurate coverage of the top if it is outweighed by unit nonresponse.

The extent of oversampling is measured in the effective oversampling rate which captures the number of unweighted households with wealth above a certain percentile that has been derived from the weighted data. When the net sample includes a relatively large number of affluent households with small average weights, this indicates high effective oversampling (ECB, 2020). Table A.1 presents the oversampling strategies and the effective Top 5% oversampling rates for the countries in our study. Values range from −15% in Austria with no oversampling to +278% in France where oversampling is based on administrative wealth tax data.

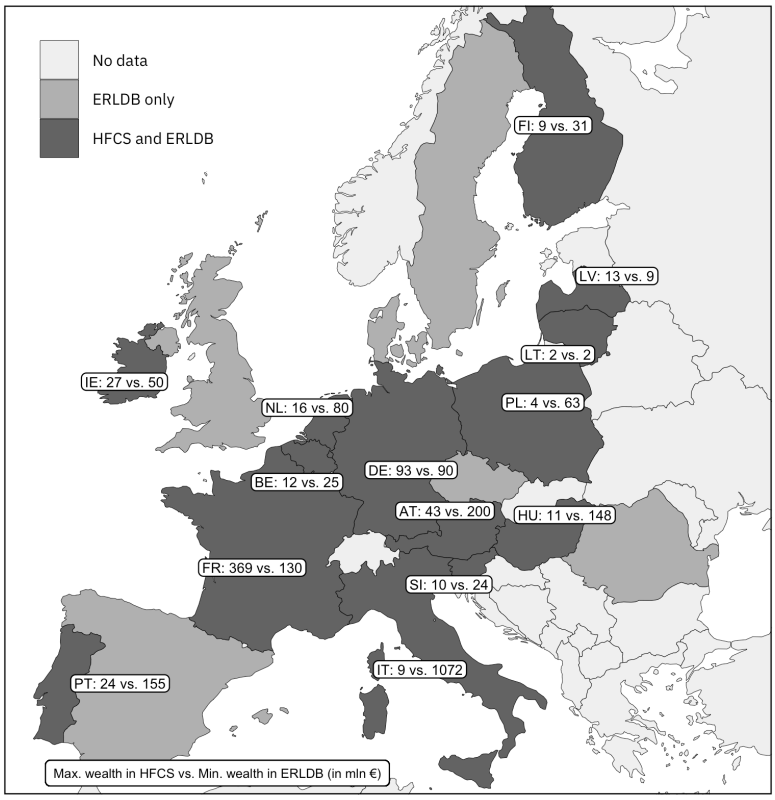
2.2 European Rich List Database (ERLDB)

Despite the various oversampling attempts in the HFCS, survey aggregates on household wealth usually are considerably lower than macroeconomic aggregates from National Accounts (Vermeulen, 2016a; Chakraborty and Waltl, 2018). This corresponds to the observation that the highest fortunes in HFCS data are substantially smaller than evidence from journalistic rich lists. Thus, studies focusing on the top of the wealth distribution often rely on such lists to enrich survey data with information on very affluent households. For this reason, we have collected lists from 20 countries with roughly $9,000$ observations and make them publicly available for research as European Rich List Database. Figure 1 shows the geographical coverage of the ERLDB and compares the maximum values in the HFCS with minimum values in the rich lists.

While it is convenient for cross-country studies to use the international billionaires list by US-magazine Forbes, national rich lists feature some important advantages. First, the Forbes list only contains US-Dollar billionaires, whereas rich lists compiled by national magazines or newspapers mostly comprise observations with much less wealth. Second and related, national rich lists provide significantly more entries and list up to $1,000$ observations for a single country while the Forbes list totals roughly $2,100$ observations worldwide. National rich lists might thus improve wealth estimates particularly in countries with only a few entries in the Forbes list (Bach et al., 2019). Third, local journalists might have better insights and intuition for the wealth portfolios of the rich in a country than an international team of reporters. Although editors and journalists do their best to accurately cover the super-rich, these national lists also have weaknesses.

First, it is questionable whether rich lists are exhaustive. Individuals can opt out for

Figure 1: Wealth in Surveys and Rich Lists



privacy issues or simply do not appear in the list although they would qualify (Kennickell, 2003). Second, investigators rely on public information which may be flawed particularly with regard to private assets and liabilities. Moreover, some assets are difficult to assess, for instance art collections and business wealth in form of non-traded corporate shares. Additionally, debts are less visible than assets, and potentially cause net worth to be overstated (Kopczuk, 2015; Atkinson, 2008; Davies and Shorrocks, 2000). Third, rich lists often mix individuals, families, and family clans. Accordingly, Bach et al. (2019) show for Germany's Manager Magazin list that some individual observations actually consist of several households.

Despite these methodological deficiencies, rich lists provide important information on individuals and families that are not captured in wealth surveys. We thus enrich survey data from HFCS with ERLDB data for 14 countries. In some cases, the interview period of the HFCS and the reference period of the national rich list do not match and we have to choose the closest year. When the interview period in HFCS was not restricted to a calendar year, we select the year in which the most interviews were conducted. Table A.1 presents detailed information on the number of observations and reference years of the ERLDB.

3 Method

3.1 Pareto Distribution

Spanning over times and places, the distribution of income and wealth in capitalist societies takes a remarkably similar form. It was early Italian economist Vilfredo Pareto (1965 [1896]) who recognized that 80 % of Italian land was owned by the richest 20 % of the population. He quickly extended this observation for different forms of property to brand *Pareto's Law*. The 80 - 20 ratio can be generalized to a power law distribution

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

with wealth w and w_{min} as the lower bound of observations closely following a Pareto's Law. The distribution has a linear relationship between the logarithm of the complementary cumulative distribution $log(1 - F(w))$ and the logarithm of the wealth variable $log(x)$ as in

$$1 - F(w \mid w_{min}, \alpha) = \left(\frac{w_{min}}{w}\right)^{\alpha} \tag{2}$$

A corresponding log-log plot reveals the characteristic pattern at a glance, adding to the theory's popularity. More recently, the inequality literature has found that the Pareto distribution is a good approximation for the tail of the wealth distribution, i.e. the inequality among rich households (see e.g. Davies and Shorrocks, 2000 or Gabaix, 2016).

In this paper, we combine various elements from the literature estimating the Pareto tail. We exploit the linear relationship of the logarithms and apply linear regression, but include Gabaix and Ibragimov (2011)'s rank correction of the left-hand side as well as Chakraborty and Waltl (2018)'s insight that median quantile regression is more robust to outliers (Koenker and Bassett, 1978) to retrieve point estimates for shape parameter α in

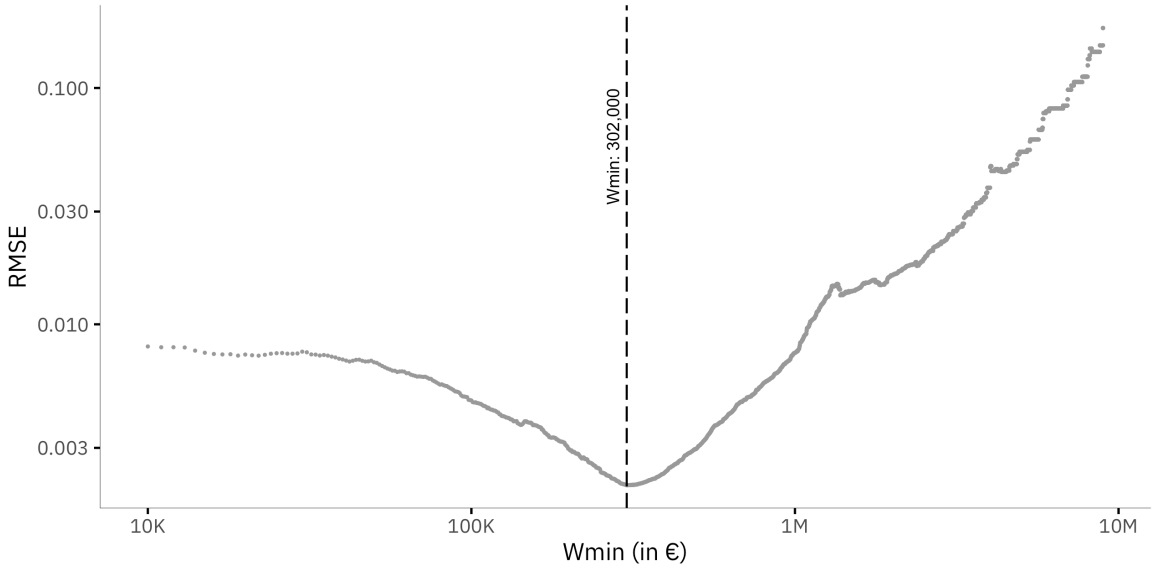
$$\log((i - 0.5)\frac{\bar{N_{fi}}}{\bar{N}}) = \underbrace{\log(\frac{\bar{N_i}}{\bar{N}}) + \alpha \log(w_{min})}_{\text{constant}} - \alpha \log(w_i) \tag{3}$$

where i is a decreasing ranking with $i = 1$ indicating the richest household, N is the sum of total weights, $\bar{N}$ is the average weight of an observation, and $\bar{N_{fi}}$ denotes the average weight of the first highest i observations. α gives the slope of the log-linearized plot, it is the inequality parameter of the distribution. A smaller α corresponds to higher inequality within the tail.

We use both HFCS survey and ERLDB data for the regression. We decide on the location parameter w_{min} by comparing the root mean squared errors (RMSE) for a wide

range of potential values. This makes use of the interpretation of the RMSE as a measure of linearity and is based on Langousis et al. (2016)'s estimation procedure for Generalized Pareto distributions. Figure 2 illustrates the procedure.

Figure 2: Determination of w_{min}



Note: This figure is based on 1st implicate of HFCS 2017 data for Germany.

There are other approaches of estimating w_{min}, Vermeulen (2016b) does his calculations for a set of potential location parameters, while Clauset et al. (2009)'s method of calculating a Kolmogorov-Smirnov metric of distance between the empirical and theoretical distribution for a number of w_{min} candidates applies a similar logic as RMSE minimization.

The motivation of our analysis is the gap between survey and anecdotal evidence on wealth, which we rationalize by suspecting differential underreporting and nonresponse bias following Vermeulen (2016b). A third parameter w_0 indicates a threshold in the tail above which we do not trust the survey data to be complete. We make use of Eckerstorfer et al. (2016)'s argument that this point should coincide with the transition from continuous to discrete survey observations. We adapt Dalitz (2016)'s intuition that this transition can be found where the empirical density function of the data falls below the theoretical probability density function. Equations 4 and 5 define the equality condition for $\hat{w}_0$, which we determine numerically:

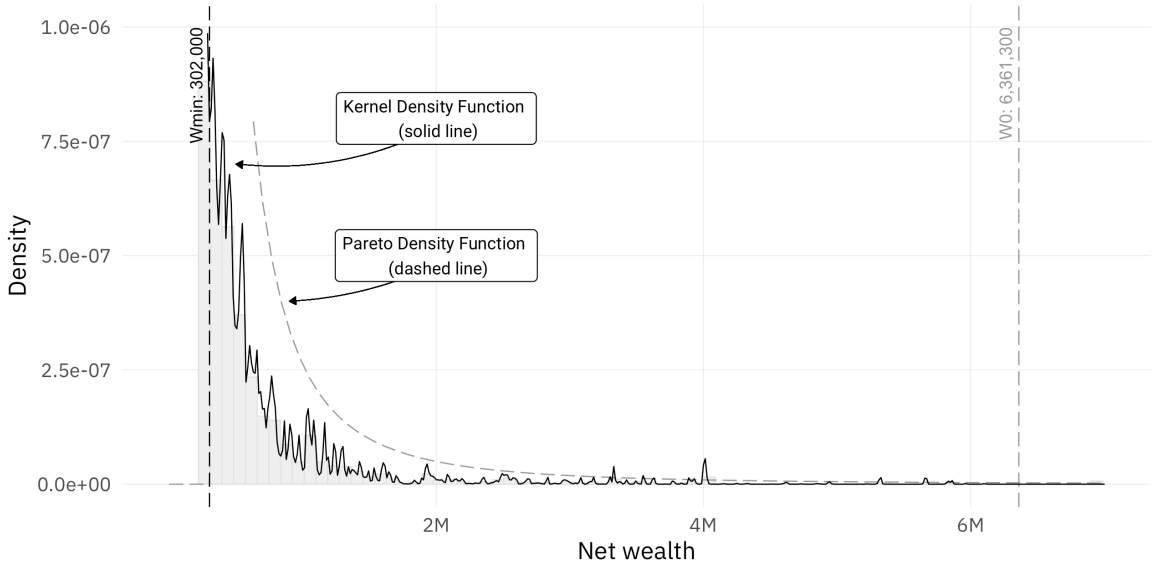
$$\hat{w}_0 = w_0 : \hat{f}_{kern}(w_0) = \frac{1}{Nh}\sum_i n(w_i)K(\frac{w_0 - w_i}{h}), \tag{4}$$

$$\hat{f}_{kern}(w_0) - \underbrace{\alpha w_{min}^{\alpha}\frac{1}{N}\sum_{w_i > w_{min}} n(w_i)}_{\text{normalizing constant C}} \times w_0^{-(\alpha+1)} = 0, \tag{5}$$

Let $n(w_i)$ be the weight of some household i, and h be the bandwidth for the kernel estimation, which we choose using Sheather and Jones (1991)'s procedure.

Note that the procedure includes a normalizing constant C, which adjusts the number of tail observations such that the sum of weights in the population before and after re-estimation remains the same (Eckerstorfer et al., 2016). C shifts the theoretical PDF up or down, and is crucial for finding the intersection of theoretical and empirical densities. Figure 3 illustrates the numerically derived result for one HFCS implicate in Germany.

Figure 3: Tail histogram with w_{min} and w_0



Note: This figure is based on 1st implicate of HFCS 2017 data for Germany.

We prefer this algorithmic approach to visual inspection, because the latter is somewhat arbitrary but also impractical for analysis over multiple countries and implicates. Furthermore, Dalitz (2016) points out that inequality metrics of data with re-estimated Pareto tails vary substantially for different values of w_0. The distance between $\hat{w}_{min}$ and $\hat{w}_0$ is an indicator of how well surveyors were able to battle differential biases among the richest households; as different participating central banks of the HFCS apply different oversampling strategies we do expect some variation here. This too emphasizes the need for a flexible and unambiguous procedure.

Finally, we simulate a new tail above w_0. We calculate the number of households with wealth above w_0 according to a $Pareto(\hat{\alpha}, \hat{w}_{min})$ distribution by extrapolating the number of households between $\hat{w}_{min}$ and $\hat{w}_0$ with the cumulative density function above $\hat{w}_0$ ($1-F(\hat{w}_{min})$ gives the theoretical share of tail observations above $\hat{w}_0$). The "tail length" is defined in

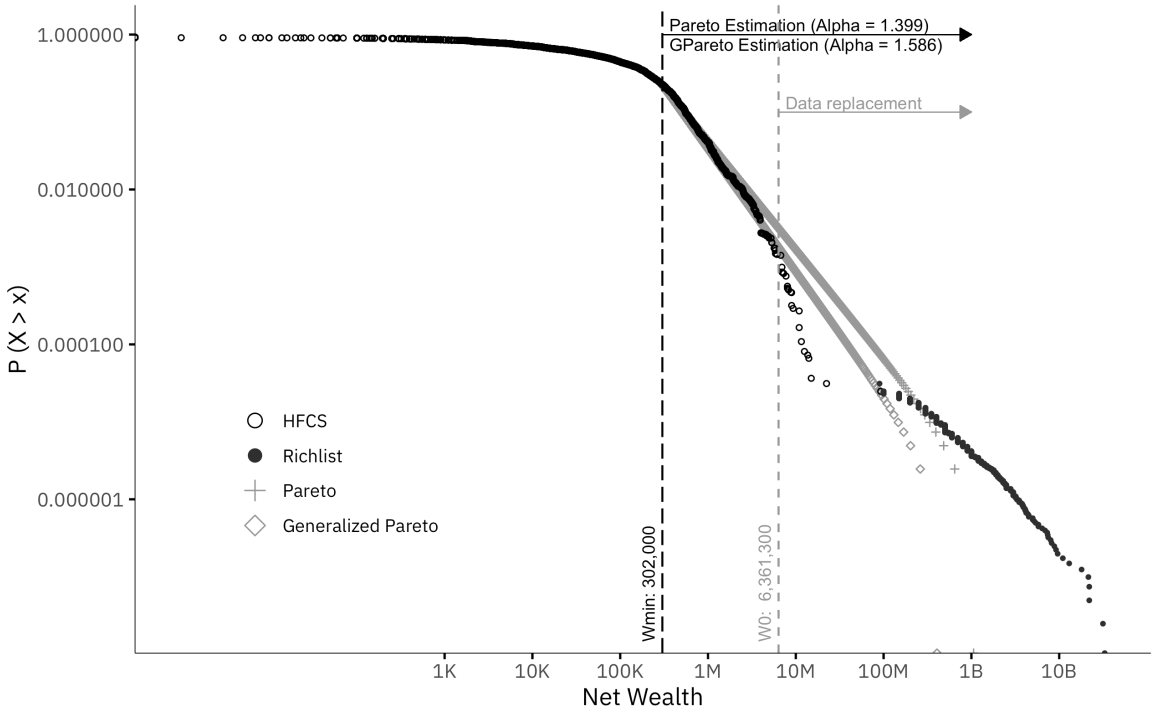
$$\sum_{w_i > w_0} n(w_i) = [\sum_{w_i \in (w_{min}, w_0)} n(w_i)] * \frac{1 - F(w_0)}{F(w_0)} \tag{6}$$

We rank the new, theoretical, observations and respectively assign wealth values according to

$$w_i = w_{min} \left(\frac{\sum_{w_i > w_{min}} n(w_i)}{\sum_{w_j > w_i} n(w_j)} \right)^{1/\alpha} . \tag{7}$$

and a uniform household weight of 1. The combination of simulated observations and all survey entries below $\hat{w}_0$ gives our re-estimated population. We calculate inequality metrics, such as the share of wealth held by the Top 1%, Top 5%, and Top 10%, P99/P50 quantile ratio, or the Gini coefficient. Figure 4 illustrates the individual steps as well as the main parameters.

Figure 4: Cumulative Density Function of HFCS, Rich List, and Pareto Simulation



Note: This figure is based on the 1st HFCS implicate and ERLDB data from 2017 for Germany.

3.2 Generalized Pareto Approach

Pareto's law approximates the tail of observable phenomena surprisingly well, but the simplicity of the two-parameter Pareto distribution implies a certain rigidity. Atkinson (2017) points out that Vilfredo Pareto envisioned an upper tail distribution that requires a rejection of a constant shape parameter α and and calls for a richer functional form to approximate economic phenomena at the top.

A recent approach by Blanchet et al. (2017) uses a non-parametric definition of power laws and implements Generalized Pareto Curves with varying α values along the top tail to interpolate tabulations of exhaustive tax data with significantly higher precision compared to non-exhaustive survey data.

We rely on survey and ERLDB data and improve the functional form of the standard Pareto distribution by introducing the Generalized Pareto distribution. It is more flexible, as it is defined by a three-parameter complementary cumulative density function (CCDF) as in

$$\left(1 + \xi \frac{w - \mu}{\sigma}\right)^{\frac{-1}{\xi}} \tag{8}$$

with a location parameter μ, a shape parameter ξ, and a scale parameter σ. Its shape parameter ξ relates to Pareto's α such that $\xi = \frac{1}{\alpha}$ (Jenkins, 2017). The location parameter μ has the same interpretation as w_{min} and indicates the threshold above households' wealth approximately follows a Generalized Pareto (GP) distribution. We adopt the standard Pareto notation and use α_{GP} and w_{min} instead of ξ and μ because the two parameters share a similar interpretation. The scale parameter σ determines the drift towards the end of the tail and defines a higher or lower wealth concentration compared to the two-parameter Pareto distribution, which is a nested special case of the Generalized Pareto distribution with $w_{min} = \frac{\sigma}{\xi}$ and therefore no drift by definition.

Our Generalized Pareto approach is an extension of our efforts outlined in section 3.1 to approximate the top tail of the wealth distribution. We build on our already detected threshold w_{min} from the standard Pareto and estimate a Generalized Pareto distribution that may fit the tail better. For a given w_{min} we estimate the scale and shape parameter.

Langousis et al. (2016) show that if the scaled excesses of a random variable over some location parameter w_{min} follow a Generalized Pareto distribution, the scaled excesses for any threshold $u \geq w_{min}$ are also Generalized Pareto distributed with the same shape parameter $\frac{1}{\alpha_{GP}}$. Furthermore, the scale parameters σ_u depends linearly on the scale parameter of the threshold w_{min}, the shape parameter, and the excess over u. The scaled excess of a random variable over any threshold u is defined as $e(u) = E[W - u \mid W > u]$. Equation 9 gives the linear relationship for σ_u, equation 10 the expected value of the excess over u.

$$\sigma_u = \sigma_\mu + \frac{1}{\alpha_{GP}}(u - w_{min}) \tag{9}$$

$$e(u) = E[W - u \mid W > u] = \frac{\sigma_u}{1 - \frac{1}{\alpha_{GP}}} = \frac{\sigma_\mu + \frac{1}{\alpha_{GP}}(u - w_{min})}{1 - \frac{1}{\alpha_{GP}}} = \beta_0 + \beta_1 u \tag{10}$$

The linear relationship in equation 10 allows for a linear regression based estimation

of both the scale and shape parameters, since $\beta_1 = \frac{1}{\alpha_{GP}}/(1 - \frac{1}{\alpha_{GP}})$ and $\beta_0 = (\sigma_u - \frac{1}{\alpha_{GP}} w_{min})/(1 - \frac{1}{\alpha_{GP}})$. Then, $\frac{1}{\alpha_{GP}} = \beta_1/(1 + \beta_1)$ and $\sigma_{w_{min}} = \beta_0(1 - \frac{1}{\alpha_{GP}}) + \frac{1}{\alpha_{GP}} w_{min}$.

We deploy the empirical strategy of Langousis et al. (2016) to our weighted survey data. We estimate the weighted mean excesses $e(w) = E[W - u \mid W > u]$ above different thresholds $u_i = W_{i,n}$ with $i = 1, 2, ..., n - 20$. Omitting the last (i.e. largest) 20 observations ensures that mean excesses are calculated based on at least 20 observations. This effectively pairs every observation w_i with a mean excess value $e(w_i) = E[W - w_i \mid W > w_i]$. For each observation w_i, $i = 1, 2, ..., n - 20$, we calculate the conditional weighted excess variance $Var[W - w_i \mid W > w_i]$ to account for the increasing estimation variance of $e(w_i)$ in w_i. The weights are calculated as $v_i = (N - i)/(Var[W - w_i \mid W > w_i]$. Finally, we perform a weighted least squares estimation corresponding to equation 10 with weights v_i.

From the literature on Pareto α estimations, we know that quantile median regressions of linear relationships derived from a distribution's density function is often more robust to outliers than ordinary least squares regression. Therefore we perform weighted least square and quantile regressions and compare the results.

Our survey-data-driven approach indicates the transition threshold w_0 where the empirical density suggests that we should not trust the survey data above w_0 to fully capture the tail. Therefore we utilize the w_0 estimates from the two-parameter Pareto approach and obtain the length of the Generalized Pareto tail in the same way as in the previous section.

As a last step we simulate the tail above w_0 according to our estimates and assign wealth values to our new theoretical observations as in $GPareto\,(\hat{\alpha}_{GP}, \hat{\sigma}, \hat{w}_{min})$

$$w_i = w_{min} + \alpha_{GP}\sigma \left[\left(\frac{\sum_{w_i > w_{min}} n(w_i)}{\sum_{w_j > w_i} n(w_j)} \right)^{-1/\alpha_{GP}} - 1 \right]. \tag{11}$$

This allows us to obtain the same standard metrics of inequality as in the standard Pareto approach.

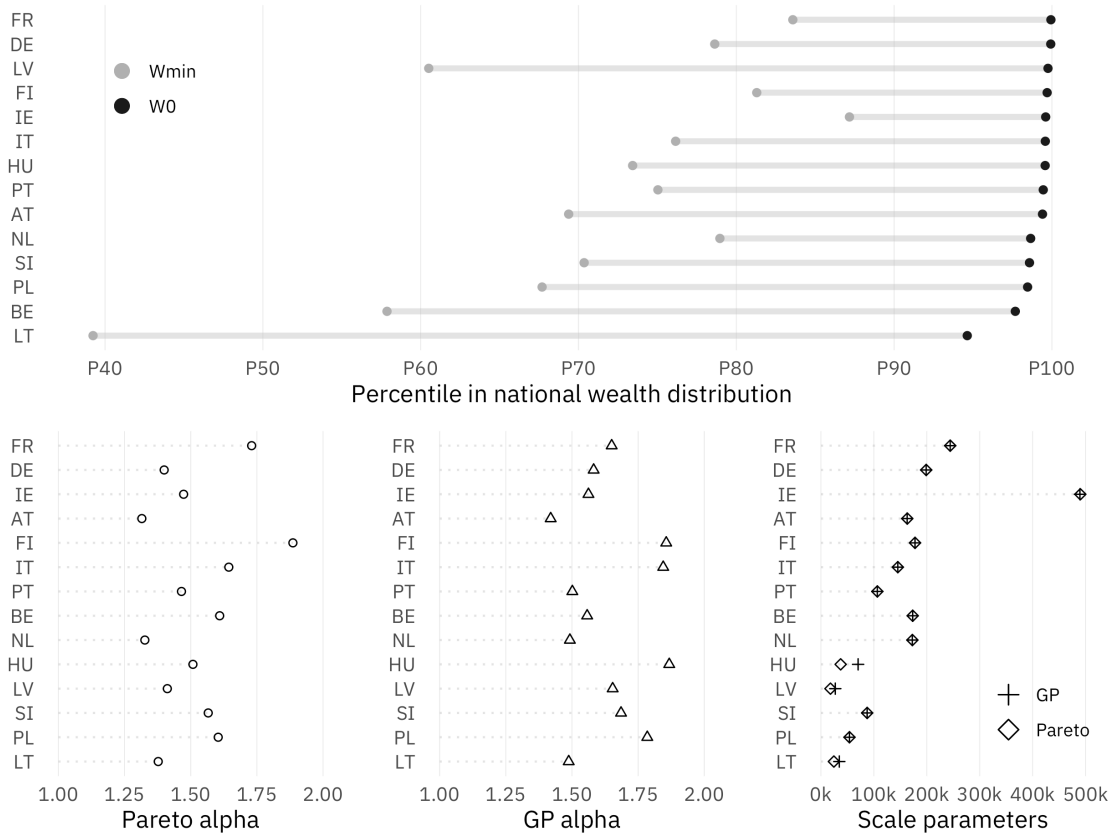
4 Results

We combine HFCS and ERLDB data to tackle underreporting and differential nonresponse that is disproportionally prevalent among the super-rich and to gain better insight in the wealth concentrated at the top. We apply a novel unified regression approach to the Pareto distribution and incorporate findings of the recent literature on linearized parameter estimation. While the Pareto approach allows us to close the gap between survey and rich list observations, we extend the framework and estimate a three-parameter Generalized Pareto distribution which is able to capture a "drift" deviation from the linear relationship between the logarithms of the complementary cumulative distribution function and wealth levels.

The Generalized Pareto approach comes with a trade-off since it is more flexible and more robust where differential underreporting is particularly prevalent but also more complex and arduous to estimate.

With regard to the Pareto distribution, we estimate a location parameter w_{min}, that marks the threshold where the data starts following a Pareto distribution, and a shape parameter α that defines the degree of inequality in the tail. We estimate the Pareto parameters sequentially. First, we apply the median regression approach by Chakraborty and Waltl (2018) to determine point estimates of α for a sequence of w_{min}s. Then, we minimize the regressions' root mean squared error $RMSE(w, \alpha \mid w_{min})$ and extract the corresponding parameter values. Figure 5 illustrates average parameter estimates between implicates, while we report the full list in Appendix B.

Figure 5: Parameters of estimation strategies



Note: This figure is based on all five implicates of HFCS 2017 data.

Our estimates for the location parameter w_{min} vary considerably across countries. For Lithuania, the starting point of the Pareto tail is as low as the 39th percentile (€36,400) of the national net wealth distribution, while the threshold is at the 87th percentile (€765,600) in Ireland. This wide range of location parameters indicates a considerable variety of wealth accumulation regimes in Europe and mirrors different oversampling strategies. The variety of

best-fit location parameters also underline the advantage of a unified and rule-based approach over arbitrary choices of w_{min}, especially when dealing with a multi-country panel.

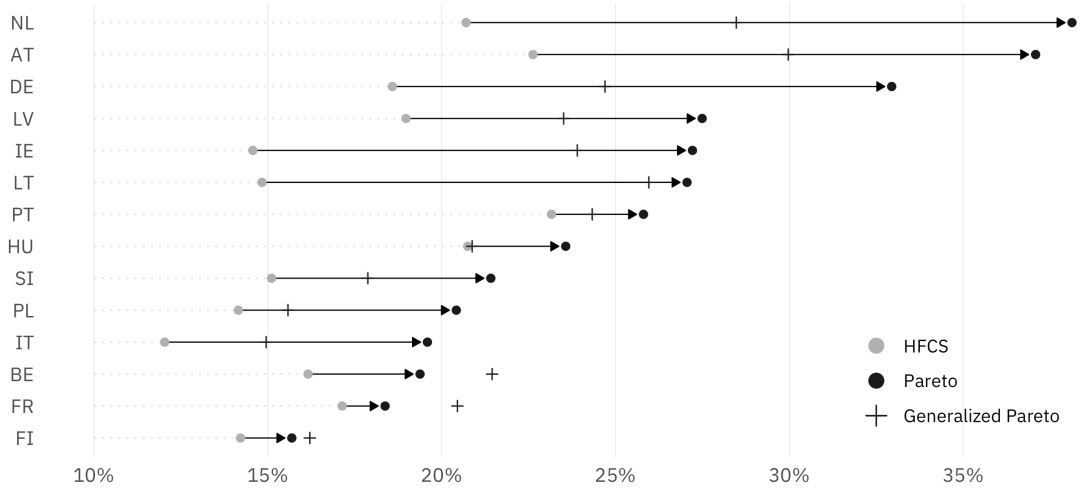
The Pareto α in the lower left-hand panel of figure 5 determines the heaviness of the tail. A smaller α implies higher inequality within the tail and, for a given location parameter, also higher inequality across the whole population. Our estimates for the shape parameter range from 1.32 in Austria to 1.89 in Finland. The results are consistent with the assertion in Gabaix (2016) that parameter values around 1.5 are the norm for wealth distribution. Our estimations are also in the range of values presented in the literature (Ferschli et al., 2017; Vermeulen, 2018; Brzezinski et al., 2020), although our sample is based on the latest HFCS wave, contains more countries, and applies a different estimation strategy than any of these papers.

The parameters of the Pareto distribution are based on a combination of HFCS and ERLDB data. In a next step, we determine a transition value w_0 above which we discard the data and simulate observations from the estimated Pareto distribution. The position of the transition value mirrors the success of oversampling strategies to include very rich observations and tackle differential nonresponse. The better the survey data is able to cover the top, the higher is the threshold for data replacement. We find considerable correlation between our estimated w_0 and the effective HFCS oversampling rates in figure A.2, indicating that successful oversampling significantly reduces the need to simulate wealth observations in the top tail.

For the Generalized Pareto distribution, we build on the location parameter from the Pareto estimation as both distributions share the same interpretation of w_{min}, that is the threshold where the data starts to follow a (Generalized) Pareto distribution. The same is true for the threshold w_0 beyond which we believe differential nonresponse and under-reporting render the survey data problematic. The additional scale parameter σ increases the flexibility and determines the drift in the tail. When $\sigma = w_{min}/\alpha_{GP}$, the Generalized Pareto equals a Pareto distribution. For a given α_{GP}, a scale parameter $\sigma > w_{min}/\alpha_{GP}$ implies that the heaviness of the tail increases towards the top and results in a higher degree of inequality. The shape and scale parameters for the Generalized Pareto distribution are depicted in the lower right-hand panels of figure 5. For most countries, the scale parameter is very close to the Pareto equivalent with no drift, except for Hungary, Latvia, and Lithuania. In these countries, the heaviness decreases slightly towards the top of the tail in the Generalized Pareto framework.

For each country and survey implicate, we simulate the tail above threshold w_0 for both the Pareto and the Generalized Pareto distribution by first calculating the number of observations above w_0 using the cumulative density function, and then assigning the appropriate theoretical quantile to each observation. We combine the simulated tail with survey observations and derive inequality metrics and top wealth shares for Pareto and Generalized Pareto. Figure 6 shows wealth shares for the Top 1%, whereas table 1 provides an overview of several inequality metrics for the raw and augmented survey data respectively.

Figure 6: Net Wealth Share of Top 1%



Note: This figure is based on all five implicates of HFCS 2017 data.

It is noteworthy that countries with the highest oversampling rates, such as Finland, France, Hungary, and Portugal experience the smallest changes in inequality measures. These countries either rely on wealth tax data or available dwelling information for survey oversampling, or obtain information on specific assets directly from administrative registers. In this regard, Germany is an exception as top shares increase substantially with the Pareto estimation even though the oversampling rate in HFCS is among the highest. This is however not surprising, as we can see in figure B.5 that the regional oversampling does not close the gap between the HFCS and the rich list observations, compared to France in figure B.7, where the oversampling is similarly high, but based on register wealth data. Additionally, we observe very large changes in countries like Ireland, Netherlands, and Lithuania. Here, the Top 1% shares almost double while Bottom 50% shares substantially decrease. Top 5% and 10% shares resemble the patterns of Top 1%, albeit to a lesser extent.

Surprisingly, the more flexible Generalized Pareto simulation leads to smaller increases in top shares. On average, they increase half as much compared to standard Pareto estimates with two notable exceptions. First, France, Finland, and Belgium show even higher top shares than in the Pareto approach. Second, Generalized Pareto estimates for Hungary do not deviate from HFCS top shares indicating that there is no value added with the Generalized Pareto simulation in this case. In a cross-country perspective, Generalized Pareto top shares are closer together than Pareto top shares, which confirms our intuition that the greater flexibility leads to higher robustness. However, it seems as if the strength of the Generalized Pareto is to intervene when the survey deviates from the Pareto distribution, which comes with a disadvantage at closing the linear gap.

The adjustment of the top tail of wealth distribution clearly has considerable effects

on wealth aggregates, which can be seen in figure A.1. Particularly in Austria and in the Netherlands, the Pareto estimation increases total wealth by more than 30 and 40 per cent respectively. Countries with high oversampling rates and little need for data correction at the top show only small changes in total wealth. While estimates for the Generalized Pareto approach are mostly below the Pareto figures, they show a similar pattern between countries.

In sum, our estimates underscore that oversampling affluent households in survey data makes a substantial difference and contributes to higher accuracy of survey-based wealth estimates. But not only the oversampling rate, also the quality of the data basis for the oversampling crucially determines this accuracy. Our non-discretionary algorithmic approach proofs to be suited to correct for differences in the methodological differences in the surveys. For countries where wealth-correlated information is limited, *ex post* adjustments by means of national rich lists significantly increase aggregate wealth, top shares, and alternative measures of inequality - regardless of the specific estimation method. As survey data underestimates inequality at the top, such efforts provide direly needed insights into wealth inequality for evidence based policies.

Table 1: Inequality Metrics For Pareto and Generalized Pareto Estimations

	AT	BE	DE	FI	FR	HU	IE	IT	LT	LV	NL	PL	PT	SI
Gini coefficient														
HFCS	73.0	63.2	73.9	66.2	67.4	65.0	67.0	60.6	58.9	67.9	78.2	56.7	67.9	59.4
Pareto	78.8	64.0	78.7	66.8	67.9	66.2	72.6	64.6	62.8	71.6	83.6	61.4	69.1	63.1
GPareto	75.9	65.2	75.9	67.1	68.7	64.9	71.1	62.1	63.6	69.8	80.4	57.8	68.4	60.6
Share Top 1%														
HFCS	22.6	16.2	18.6	14.2	17.1	20.7	14.6	12.0	14.8	19.0	20.7	14.2	23.2	15.1
Pareto	37.1	19.4	32.9	15.7	18.4	23.6	27.2	19.6	27.1	27.5	38.1	20.4	25.8	21.4
GPareto	30.0	21.5	24.7	16.2	20.5	20.9	23.9	15.0	26.0	23.5	28.5	15.6	24.3	17.9
Share Top 5%														
HFCS	43.1	35.0	40.8	32.9	35.5	39.4	35.5	30.0	36.0	38.7	42.0	29.6	41.6	32.2
Pareto	55.4	36.3	51.8	34.1	36.5	41.5	46.3	37.0	43.2	45.6	57.0	36.9	43.8	38.2
GPareto	49.2	38.5	45.4	34.6	38.2	39.3	43.5	32.5	43.7	42.2	48.2	31.2	42.5	34.2
Share Top 10%														
HFCS	56.4	47.2	55.4	46.8	49.2	51.4	50.0	43.4	47.9	52.1	56.6	41.3	53.9	44.0
Pareto	66.2	48.2	63.7	47.7	50.0	53.0	58.7	49.2	52.8	57.7	68.2	47.8	55.6	49.2
GPareto	61.3	50.1	58.8	48.2	51.3	51.3	56.4	45.4	53.7	55.2	61.4	42.8	54.5	45.8
Share Bottom 50%														
HFCS	3.6	9.2	2.7	6.1	5.8	9.8	7.0	9.9	13.7	7.1	0.5	13.1	8.1	12.0
Pareto	2.8	9.0	2.2	5.9	5.7	9.4	5.8	8.9	12.6	6.3	0.4	11.6	7.8	10.9
GPareto	3.2	8.7	2.5	5.9	5.6	9.8	6.1	9.6	12.2	6.7	0.5	12.7	8.0	11.6
Ratio P99/P50														
HFCS	25.6	14.6	35.3	14.6	15.0	16.9	16.0	12.1	20.9	18.4	27.5	10.7	17.0	11.9
Pareto	35.9	13.1	39.1	14.7	15.2	16.5	21.1	14.4	15.8	21.3	37.6	13.9	17.6	13.9
GPareto	29.4	14.1	36.1	14.9	15.2	16.3	19.3	12.7	17.5	20.5	29.1	11.2	17.0	11.9

Note: This table is based on all five implicates of HFCS 2017 data.

5 Conclusion

While the top of wealth distribution is particularly important for understanding economic inequality, household surveys tend to cover the top percentiles insufficiently. We adjust Eurozone wealth data from the HFCS for differential nonresponse and underreporting at the top by adding observations from journalistic rich lists and suggesting a new approach to Pareto estimation.

This approach substantially changes the sensitive top share metrics. We find that wealth inequality is vastly underestimated by raw survey data. In the extreme cases of the Netherlands and Austria, the adaptation almost doubles the Top 1% shares to 38% and 37% respectively. At the same time, inequality metrics change much less in countries like France and Finland, where administrative information complements survey data. Generally, there is a significant relationship between oversampling of rich households and the precision of inequality metrics. This also implies a severe under-estimation of aggregate wealth in survey data, ranging from only 2% in France, Finland or Belgium up to 25% in the Netherlands, and Austria.

As observations of very rich individuals are essential for the quality of the tail estimation, we introduce the most comprehensive compilation of journalistic rich lists to date as European Rich List Database (ERLDB). This way, we are able to include several countries for the first time in Pareto estimations based on national rich lists. We combine ERLDB and HFCS data and present a unified algorithmic approach to parameterize the Pareto tail of the distribution. Linearization of the cumulative density function allows for the intuitive but robust median regression approach as our workhorse estimation technique, with the location parameter, survey weight correction and thresholds for simulation being derived only from the stochastic definition of the distribution or regression results. As we do not have to rely on graphical inspection or discretionary decisions, our method is easily scaleable to a large dataset and deals well with heterogeneity between countries. Furthermore, we can seamlessly extend the two-parameter Pareto approach to its three-parameter generalization, which allows for the distribution to drift towards decreasing or increasing inequality in the "tail of the tail". Our results suggest that the more flexible three-parameter estimation shows a better fit in some countries but does not add value in other countries.

The paper highlights the potential of using journalistic evidence and meticulous survey designs to improve our understanding of wealth inequality. We combine multiple data sources with a unified and robust estimation technique, which allows us to make use of all available information. At the same time, closing the gap at the top shows that inequality in many European countries is much higher than previously understood. This has important implications for policy design with respect to wealth distribution such as in fiscal planning, where affluent households might play a particularly important role.

References

Atkinson, A. B. (2008). Concentration Among The Rich. *Personal wealth from a global perspective*. Ed. by J. B. Davies. UNU-WIDER studies in development economics. Oxford: Oxford University Press, 64–89.

— (2017). Pareto and the Upper Tail of the Income Distribution in the UK: 1799 to the Present. *Economica* 84 (334), 129–156.

Bach, S., A. Thiemann, and A. Zucco (2019). Looking for the Missing Rich: Tracing the Top Tail of the Wealth Distribution. *International Tax and Public Finance* 26 (6), 1234–1258.

Blanchet, T., J. Fournier, and T. Piketty (2017). *Generalized Pareto Curves : Theory and Applications*. Working Papers 201703. World Inequality Lab.

Bricker, J., A. Henriques, J. Krimmel, and J. Sabelhaus (2016). Measuring Income and Wealth at the Top Using Administrative and Survey Data. *Brookings Papers on Economic Activity* 47 (1), 261–331.

Brzezinski, M., K. Sałach, and M. Wroński (2020). Wealth inequality in Central and Eastern Europe: Evidence from household survey and rich lists' data combined. *Economics of Transition and Institutional Change* 28 (4), 637–660.

Chakraborty, R. and S. R. Waltl (2018). *Missing the wealthy in the HFCS: micro problems with macro implications*. Working Paper Series 2163. European Central Bank.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-Law Distributions in Empirical Data. *SIAM Review* 51 (4), 661–703.

D'Alessio, G. and I. Faiella (Dec. 2002). *Non-response behaviour in the Bank of Italy's Survey of Household Income and Wealth*. Working Paper Series 462. Banca D'Italia.

Dalitz, C. (2016). *Estimating Wealth Distribution: Top Tail and Inequality*. Tech. rep. 2016-01. Hochschule Niederrhein, Fachbereich Elektrotechnik & Informatik.

Davies, J. B. and A. E. Shorrocks (2000). The Distribution of Wealth. *Handbook of Income Distribution Volume 1*. Ed. by A. B. Atkinson and F. Bourguignon. North Holland, 605–675.

Eckerstorfer, P., J. Halak, J. Kapeller, B. Schütz, F. Springholz, and R. Wildauer (2016). Correcting for the Missing Rich: An Application to Wealth Survey Data. *Review of Income and Wealth* 62 (4), 605–627.

European Central Bank (2020). The Eurosystem Household Finance and Consumption Survey: Methodological Report for the 2017 Wave. *ECB Statistics Paper Series* 35.

Ferschli, B., J. Kapeller, B. Schütz, and R. Wildauer (2017). *Bestände und Konzentration privater Vermögen in Österreich*. Working Paper Series 167. Arbeiterkammer Wien.

Gabaix, X. (2016). Power Laws in Economics: An Introduction. *Journal of Economic Perspectives* 30 (1), 185–206.

Gabaix, X. and R. Ibragimov (2011). Rank-1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business Economics and Statistics* 29 (1), 24–39.

Jenkins, S. P. (2017). Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. *Economica* 84 (334), 261–289.

Kennickell, A. B. (2003). A Rolling Tide: Changes in the Distribution of Wealth in the U.S., 1989-2001. *The Levy Economics Institute Working Paper* 393.

— (2008). The role of over-sampling of the wealthy in the survey of consumer finances. *The IFC's contribution to the 56th ISI Session, Lisbon, August 2007*. Ed. by B. f. I. Settlements. Vol. 28. Bank for International Settlements, 403–408.

Kennickell, A. B. and R. L. Woodburn (1997). Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth, Appendix A: Weighting Adjustments for 1989, 1992, and 1995. *Survey of Consumer Finances Working Paper Series*, 37–63.

Koenker, R. and G. Bassett (1978). Regression Quantiles. *Econometrica* 46 (1), 33–50.

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.

Krenek, A. and M. Schratzenstaller (2018). *A European Net Wealth Tax*. WIFO Working Papers 561. WIFO.

Langousis, A., A. Mamalakis, M. Puliga, and R. Deidda (2016). Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resources Research* 52 (4), 2659–2681.

Pareto, V. (1965 [1896]). La Courbe de la Repartition de la Richesse. *Écrits sur la courbe de la répartition de la richesse*. Ed. by G. Busino. Vol. 3. Librairie Droz, 1–15.

Pfeffer, F. T., R. F. Schoeni, A. Kennickell, and P. Andreski (2016). Measuring wealth and wealth inequality: Comparing two U.S. surveys. *Journal of Economic and Social Measurement* 41 (2), 103–120.

Sheather, S. J. and C. M. Jones (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 53 (3), 683–690.

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

— (2016a). *Estimating the Top Tail of the Wealth Distribution*. Working Paper Series 1907. European Central Bank.

— (2016b). Estimating the Top Tail of the Wealth Distribution. *American Economic Review* 106 (5), 646–650.

— (2018). How Fat Is the Top Tail of the Wealth Distribution? *Review of Income and Wealth* 64 (2), 357–387.

A Appendix

Figure A.1: Aggregate Wealth

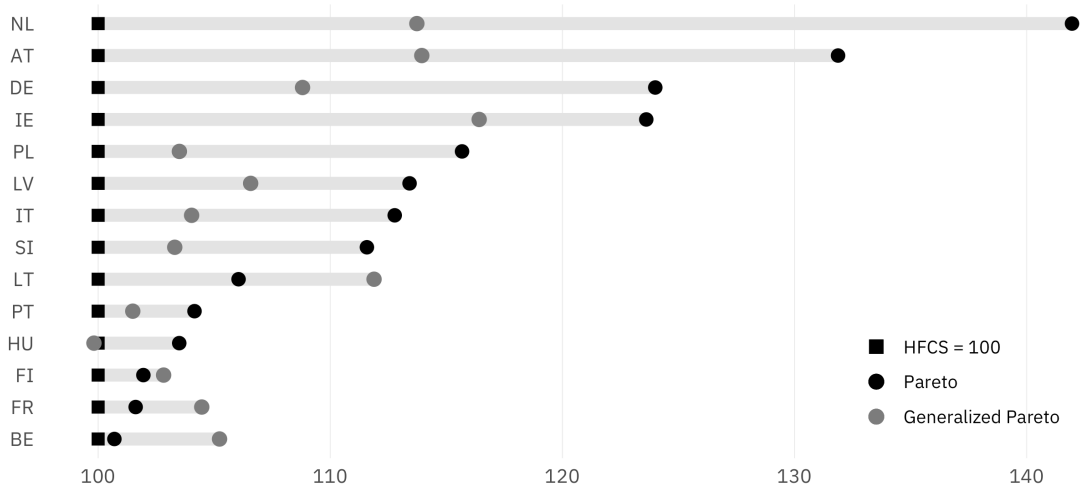


Figure A.2: Correlation of w_0 and Survey Oversampling Rate

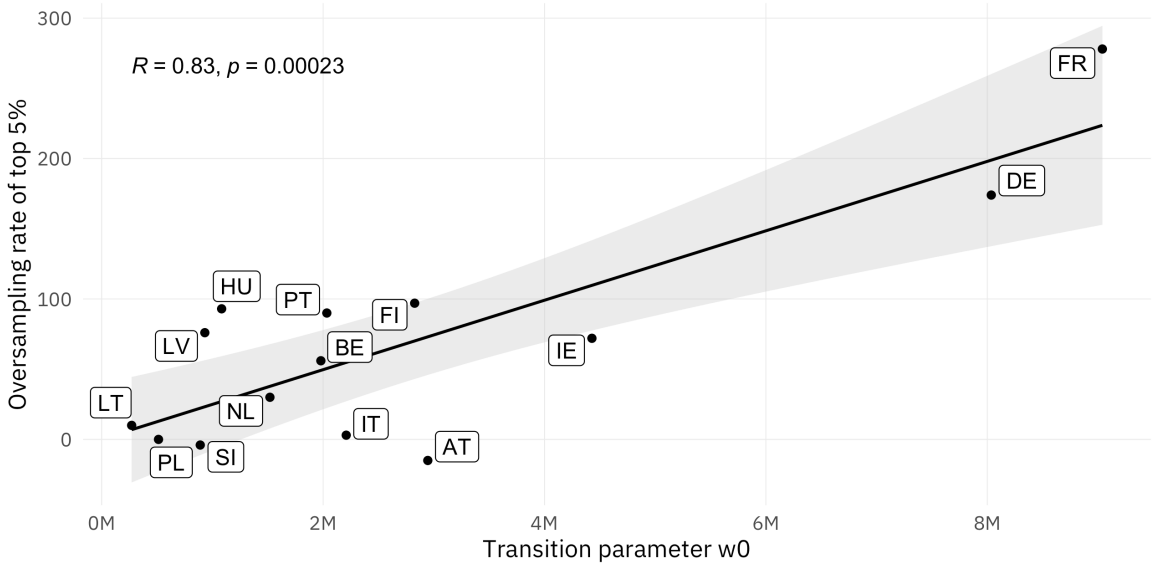


Table A.1: Summary table of HFCS 2017 and ERLDB

	HFCS					ERLDB				
			Max.	Oversampling				Wealth (mln €)		
Country	**Interview Period**	**Year**	**(mln €)**	**Type**	**Top 5%**	**Year**	**Obs.**	**Min.**	**Max.**	**Source**
Austria	11/2016 - 07/2017	2017	24	no	-15 %	2017	100	200	35,400	Trend
Belgium	01/2017 - 09/2017	2017	10	[regional] units with higher number of households and bigger dispersion of income	56 %	2018	600	25	17,295	De Rijkste Belgen
Finland	*01/2017 - 06/2017	2016	9	[income] register data	97 %	2016	50	31	1,490	Arvopaperi
France	09/2017 - 01/2018	2017	369	[wealth] register data	278 %	2017	500	130	46,900	Challenges
Germany	03/2017 - 10/2017	2017	93	[regional] wealthy street sections in cities, municipalities with a high share of taxpayers with a certain income	174 %	2017	1001	90	33,000	Manager Magazin
Hungary	10/2017 - 12/2017	2017	11	[dwellings]	93 %	2019	25	148	1,107	Napi
Ireland	04/2018 - 01/2019	2018	27	[regional] areas that scored highly on a wealth index based on home-ownership rates and local property tax bands	72 %	2018	232	50	15,600	Sunday Independent
Italy	*01/2017 - 09/2017	2016	9	no	3 %	2019	35	1,072	20,018	Forbes Italia
Latvia	09/2017 - 11/2017	2017	13	[income] register data	76 %	2017	80	9	172	Dienas Bizness
Lithuania	*12/2017 - 05/2018	2016	3	[wealth] real assets from register data	10 %	2019	500	2.1	1,400	Alfa
Netherlands	05/2017 - 07/2017	2017	16	no	30 %	2018	550	80	12,800	Quote
Poland	09/2016 - 11/2016	2016	4	[income, property] property size and register data on income	0 %	2016	100	63.4	3,641	wprost
Portugal	05/2017 - 09/2017	2017		[dwellings] size of dwelling	90 %	2018	39	155	4,502	Forbes
Slovenia	04/2017 - 10/2017	2017	10	no	-4 %	2018	100	24.2	689	Finance Manager

Source: ECB, 2020

*) Selected countries evaluated assets and liabilities not at the time of interview but on 31.12.2016

B CCDF and Pareto parameters

Figure B.3: CCDF and estimated parameters: AT

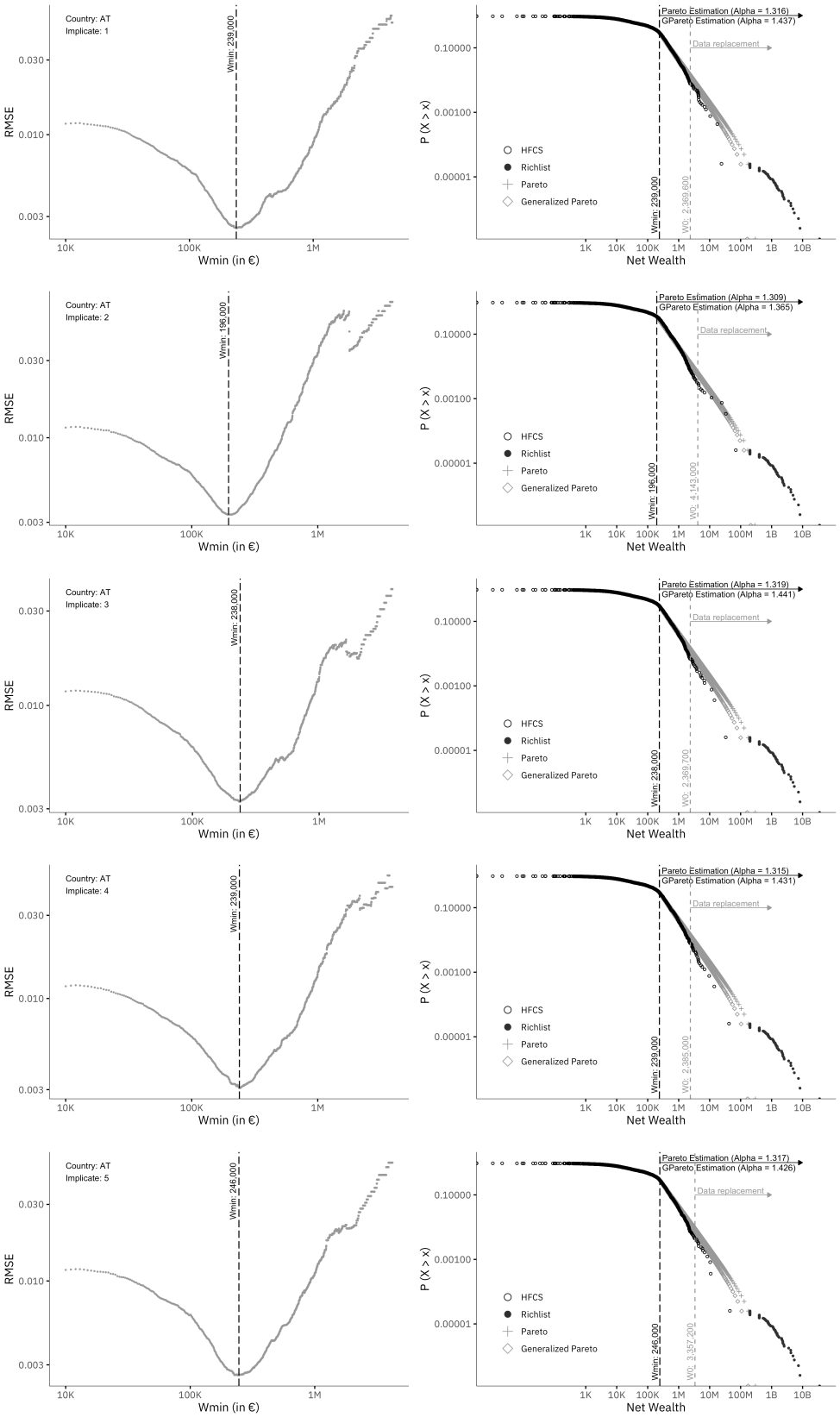


Figure B.4: CCDF and estimated parameters: BE

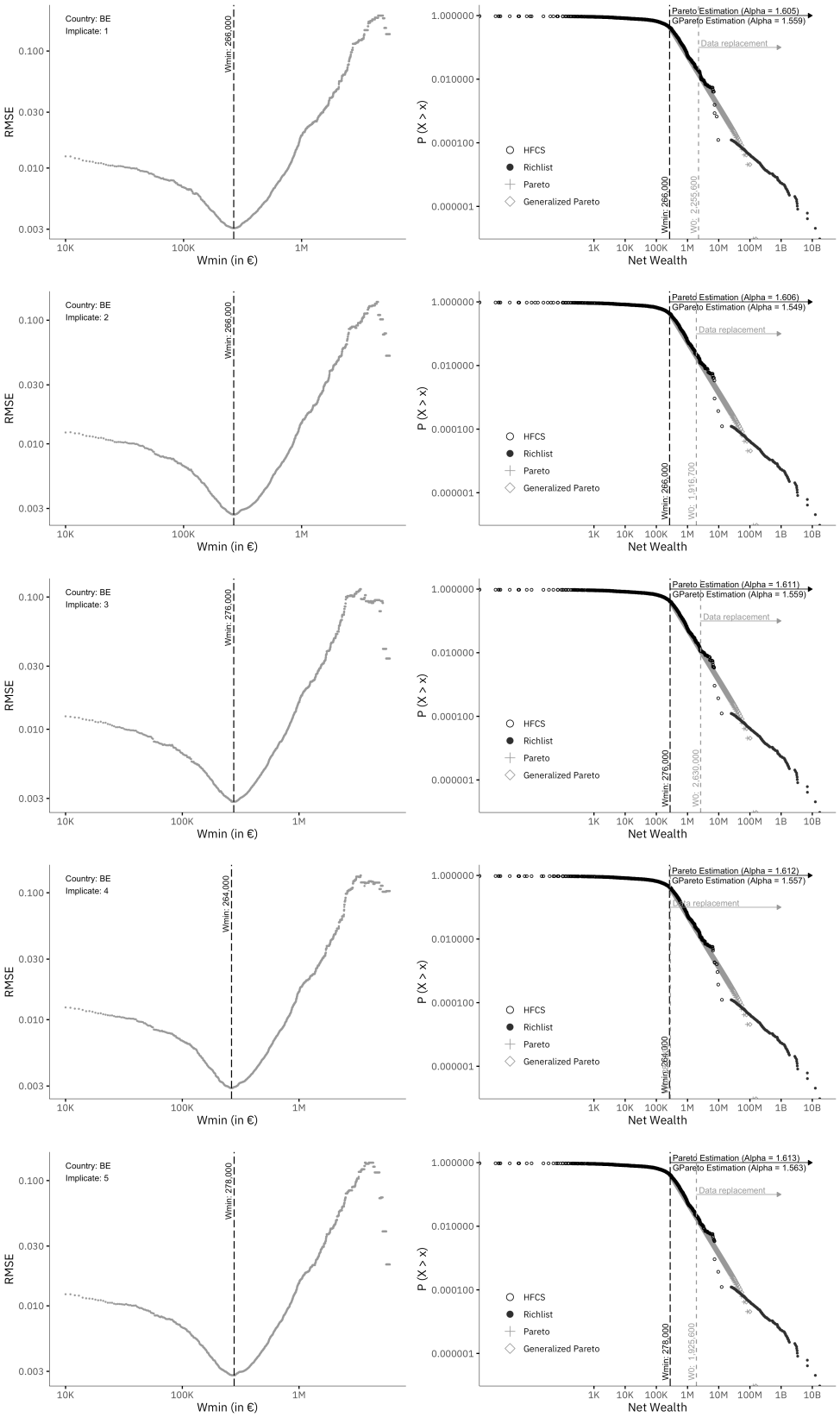


Figure B.5: CCDF and estimated parameters: DE

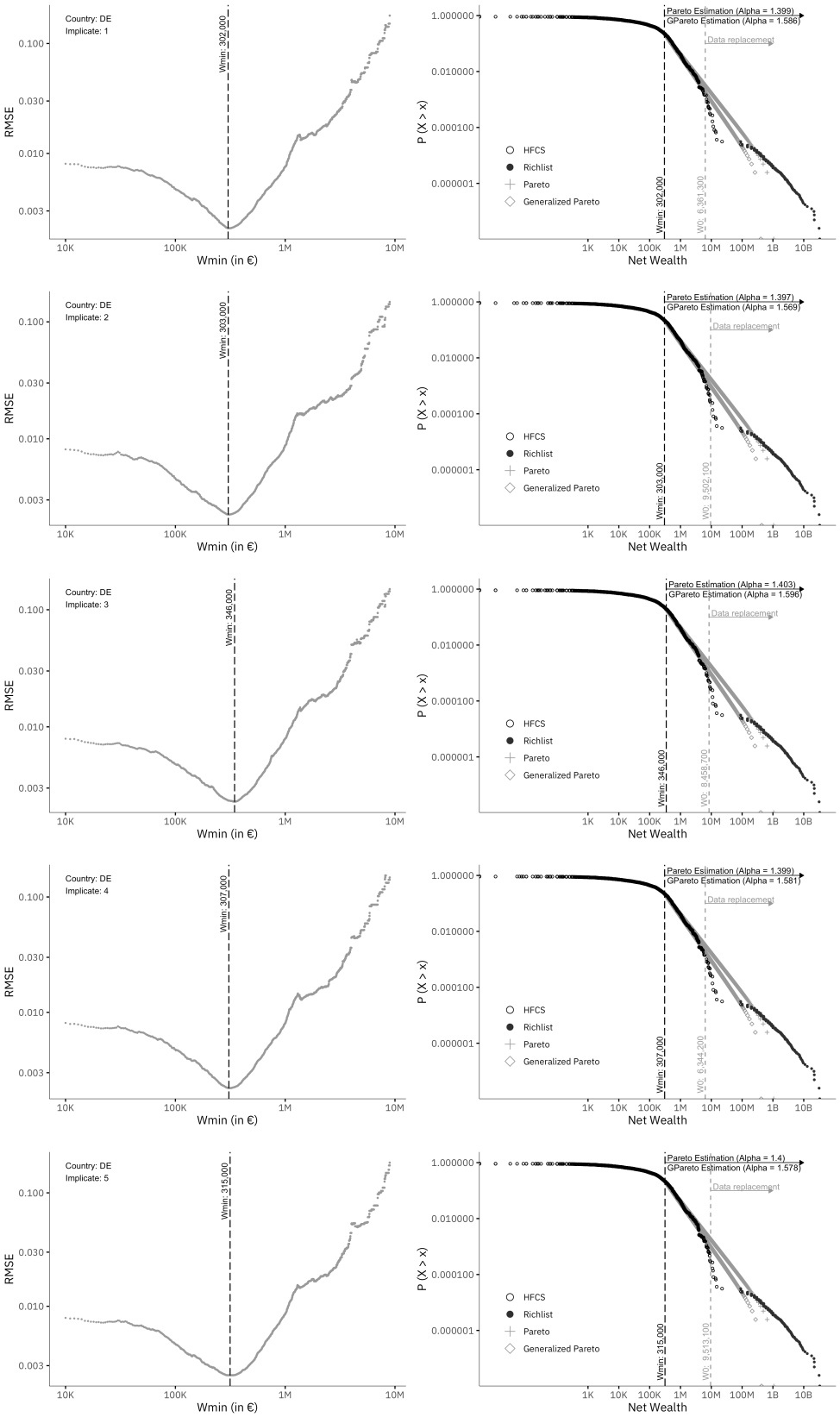


Figure B.6: CCDF and estimated parameters: FI

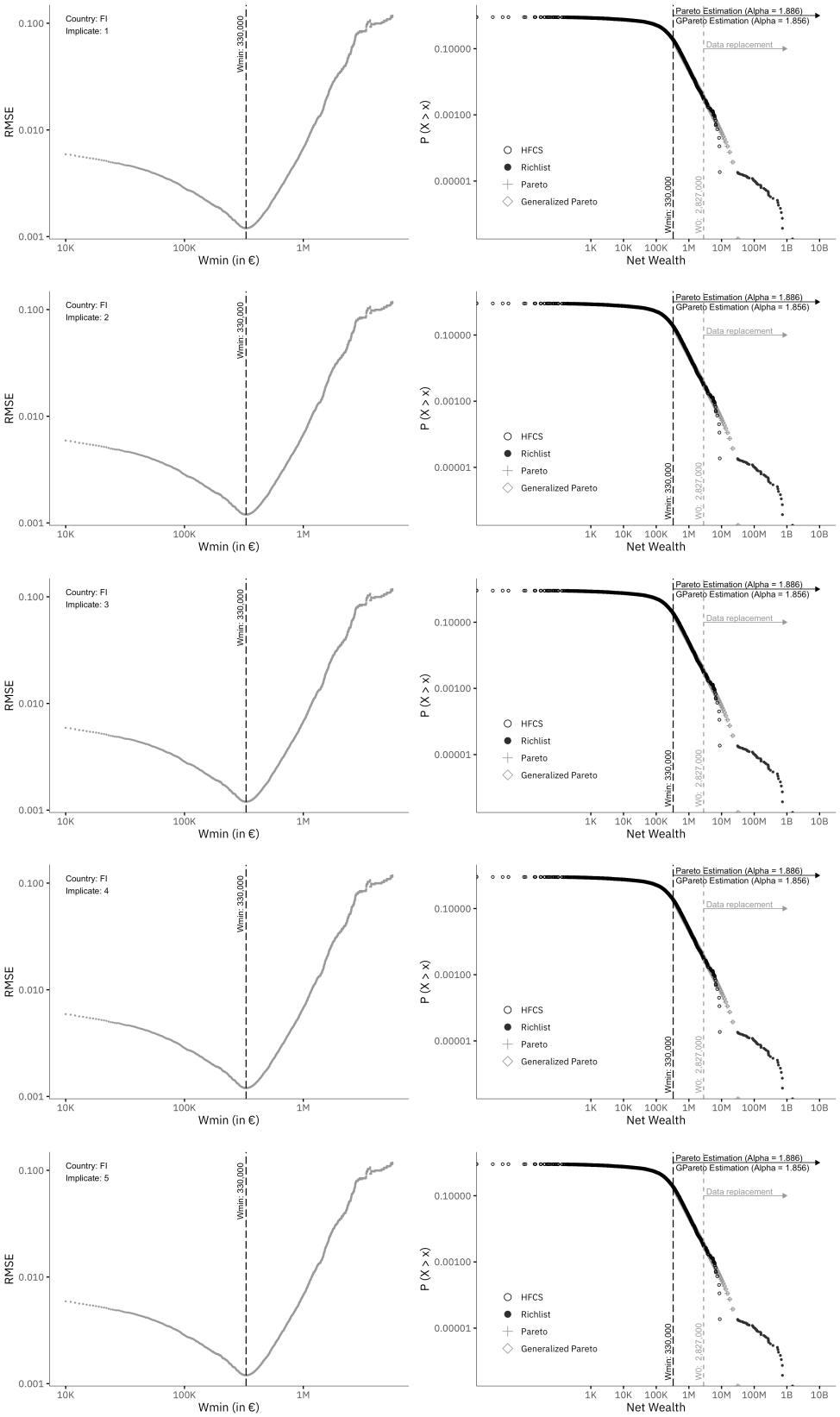


Figure B.7: CCDF and estimated parameters: FR

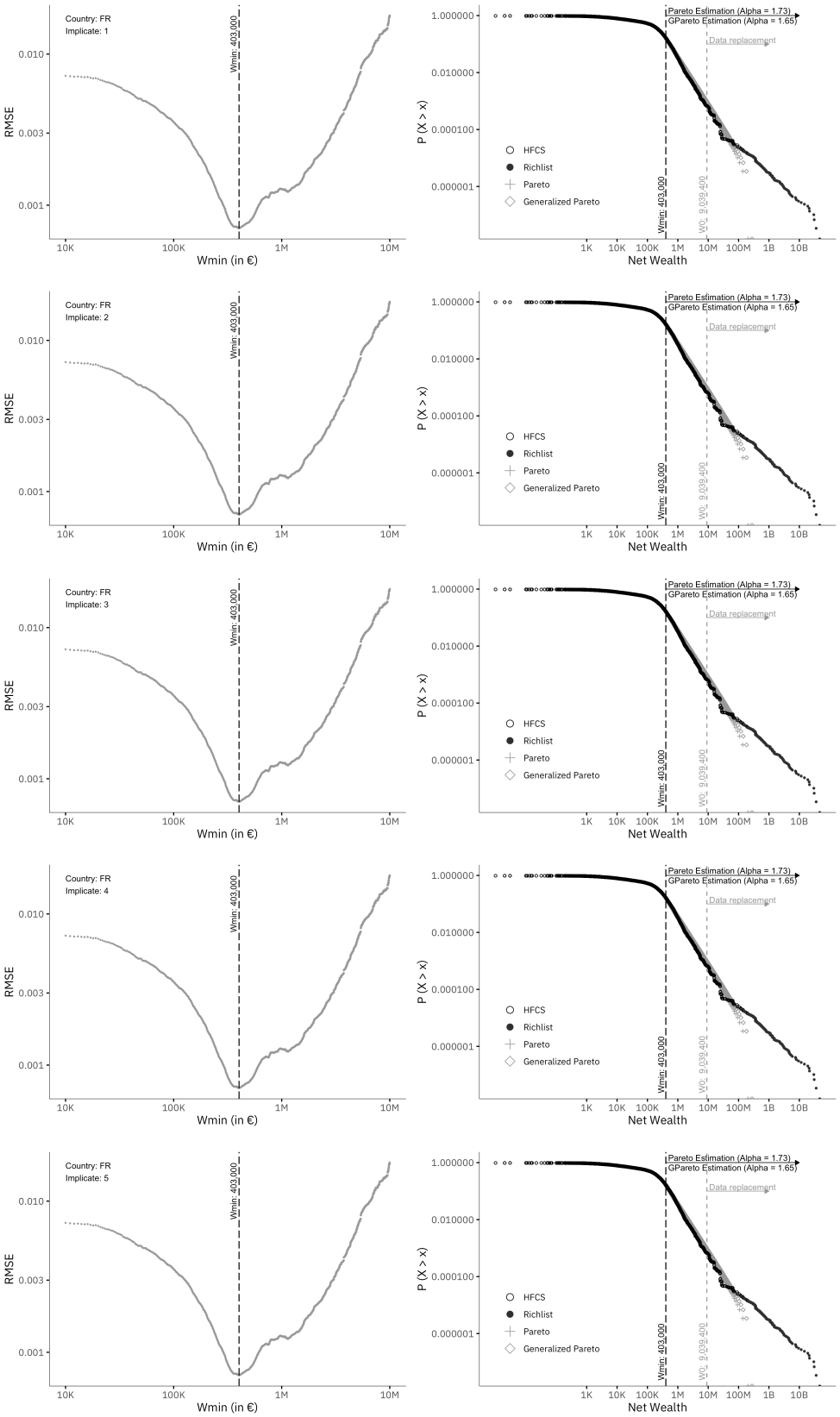


Figure B.8: CCDF and estimated parameters: HU

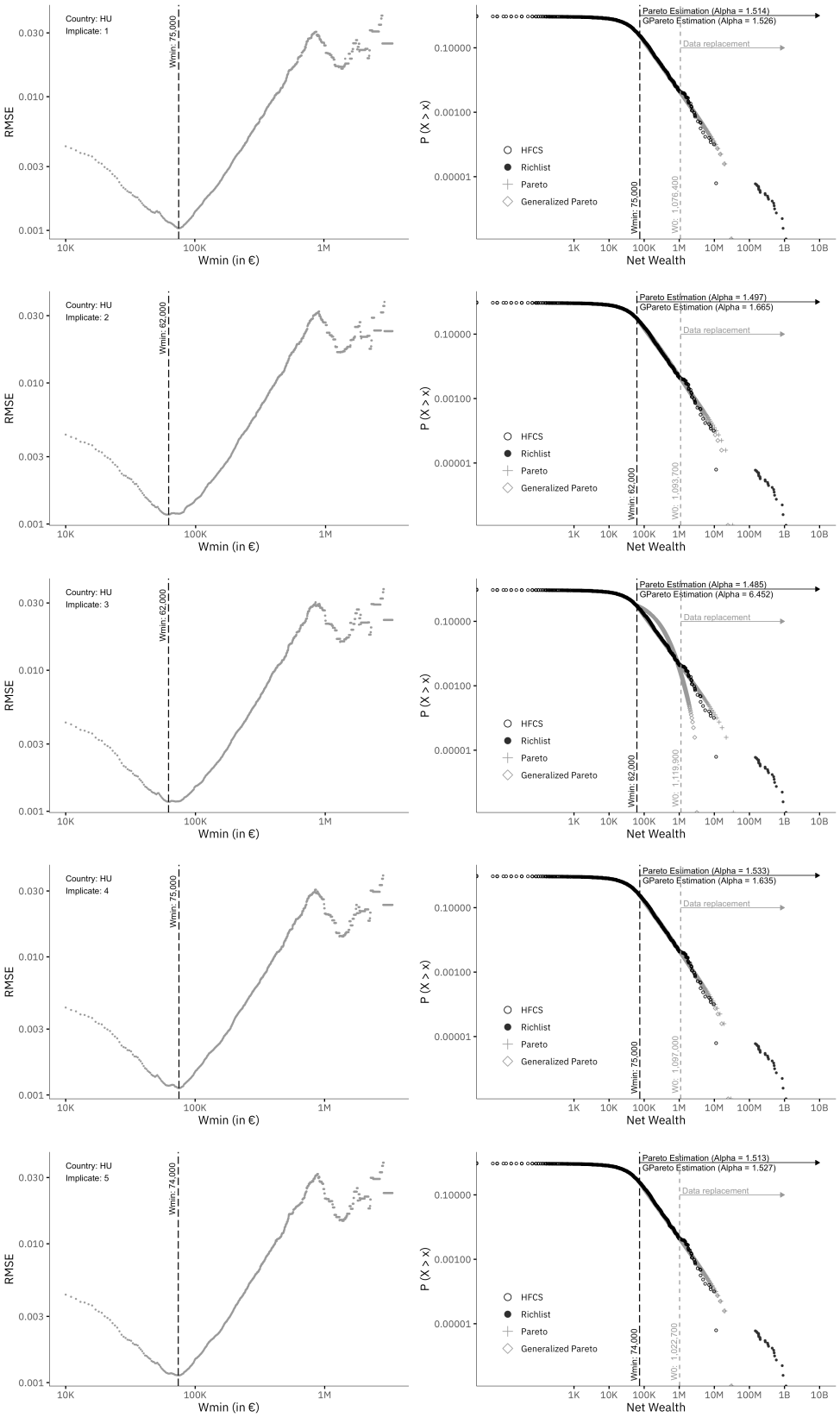


Figure B.9: CCDF and estimated parameters: IE

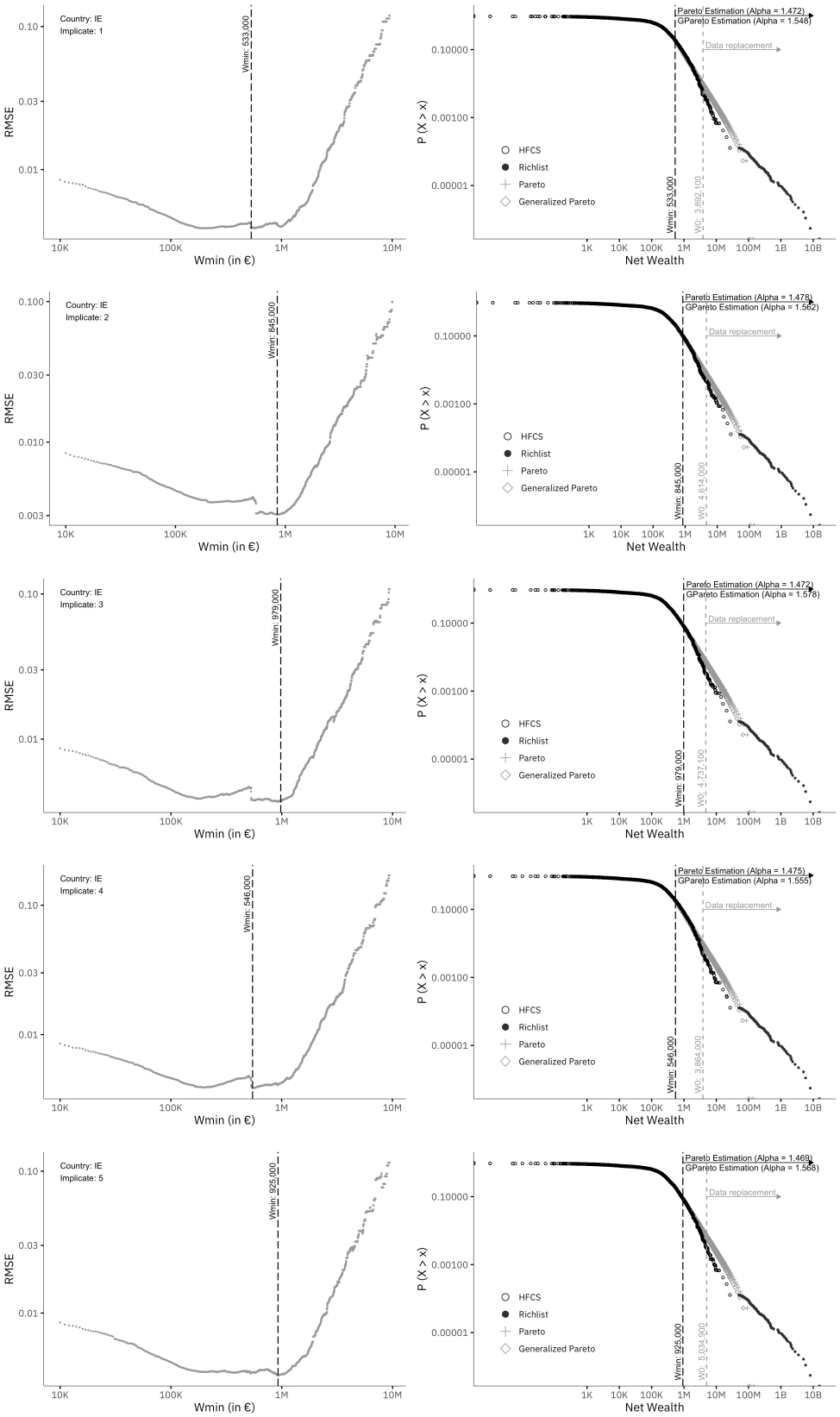


Figure B.10: CCDF and estimated parameters: IT

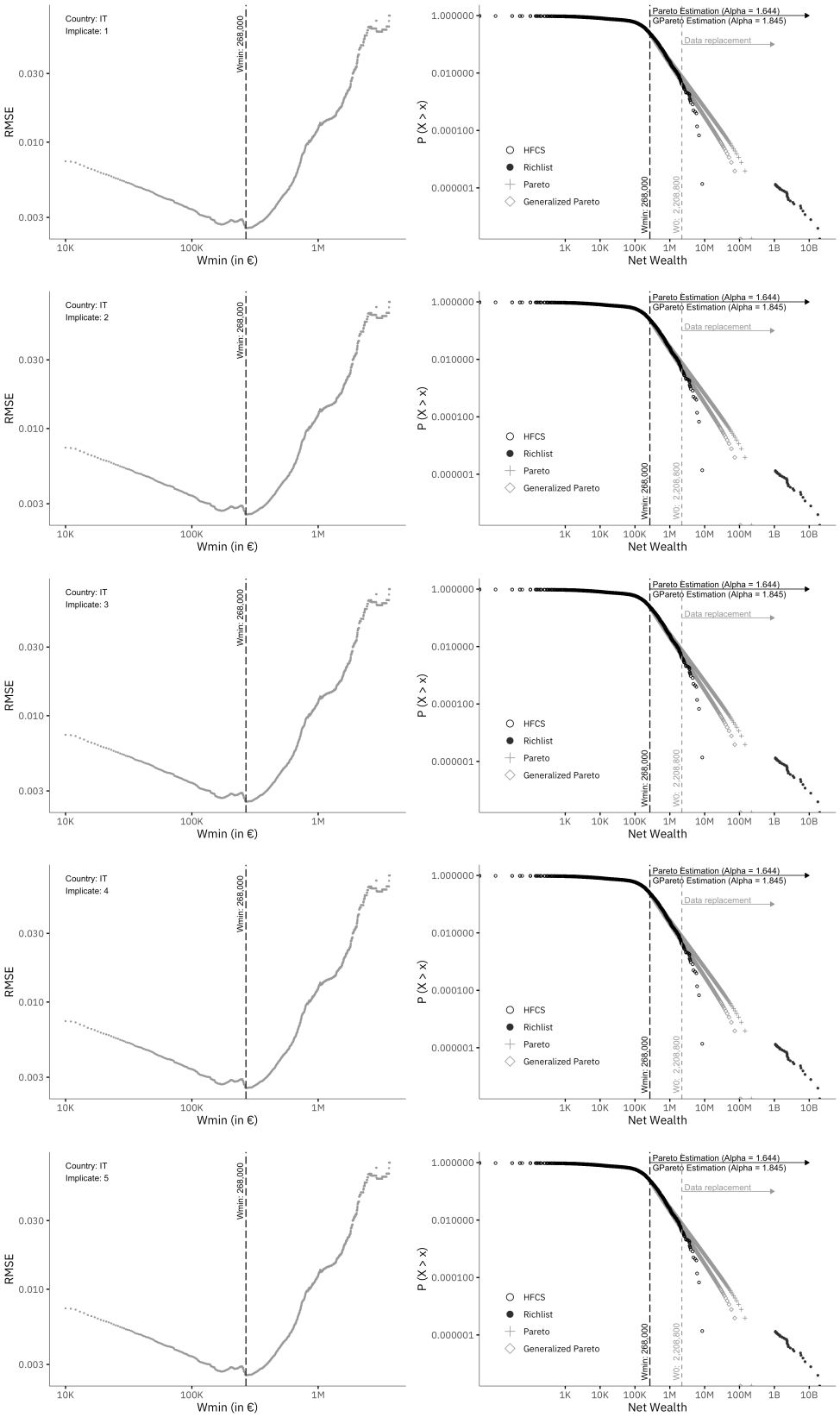


Figure B.11: CCDF and estimated parameters: LT

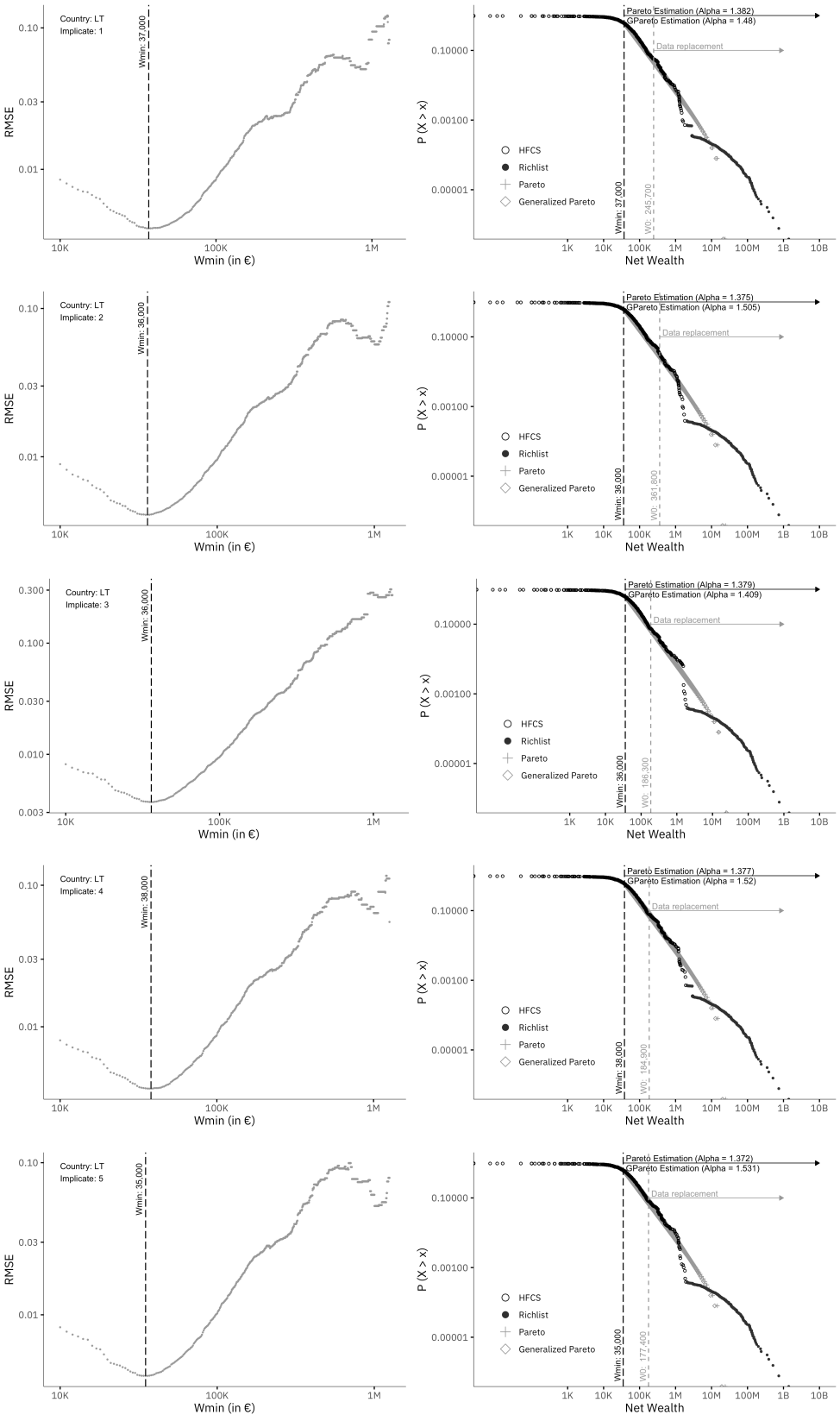


Figure B.12: CCDF and estimated parameters: LV

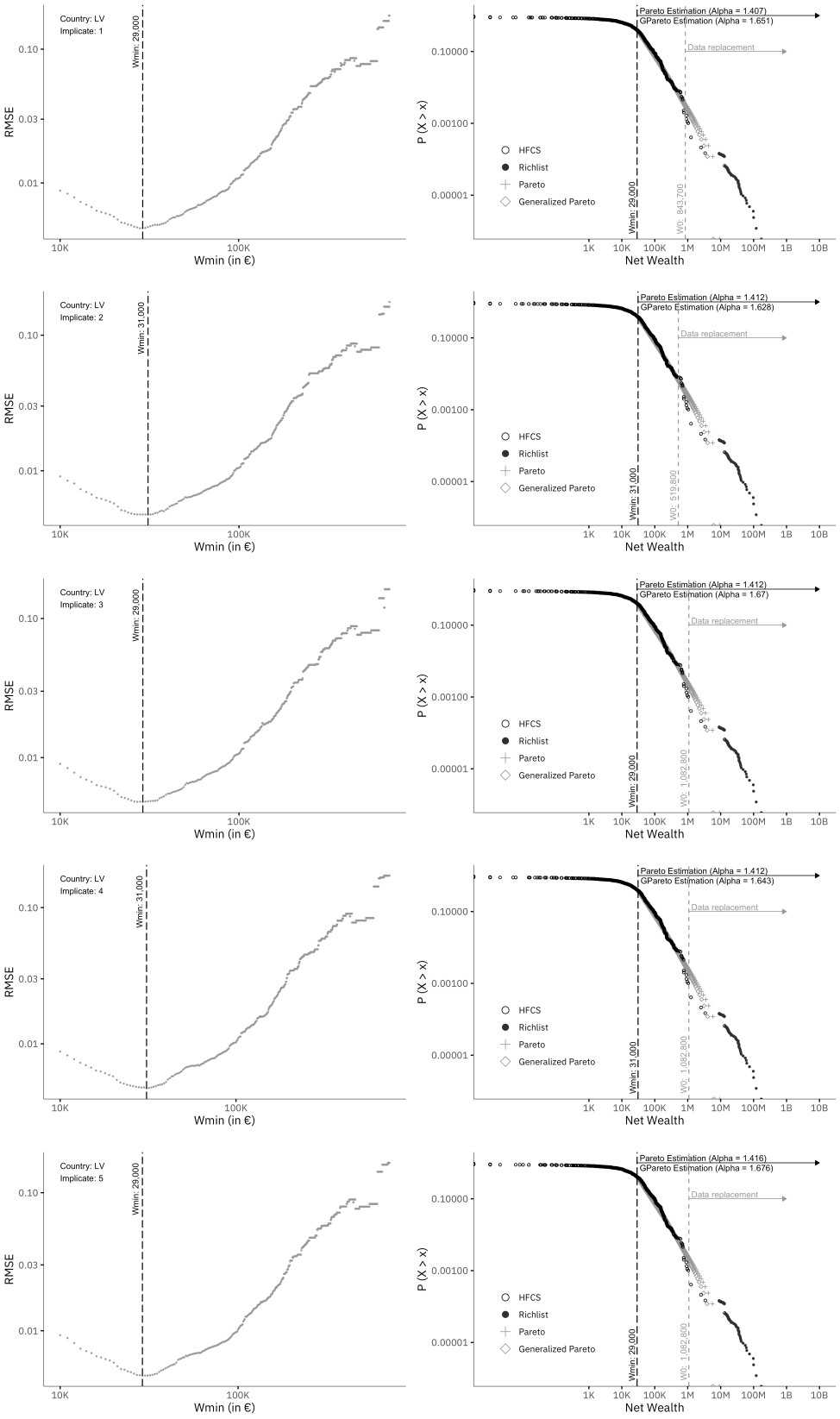


Figure B.13: CCDF and estimated parameters: NL

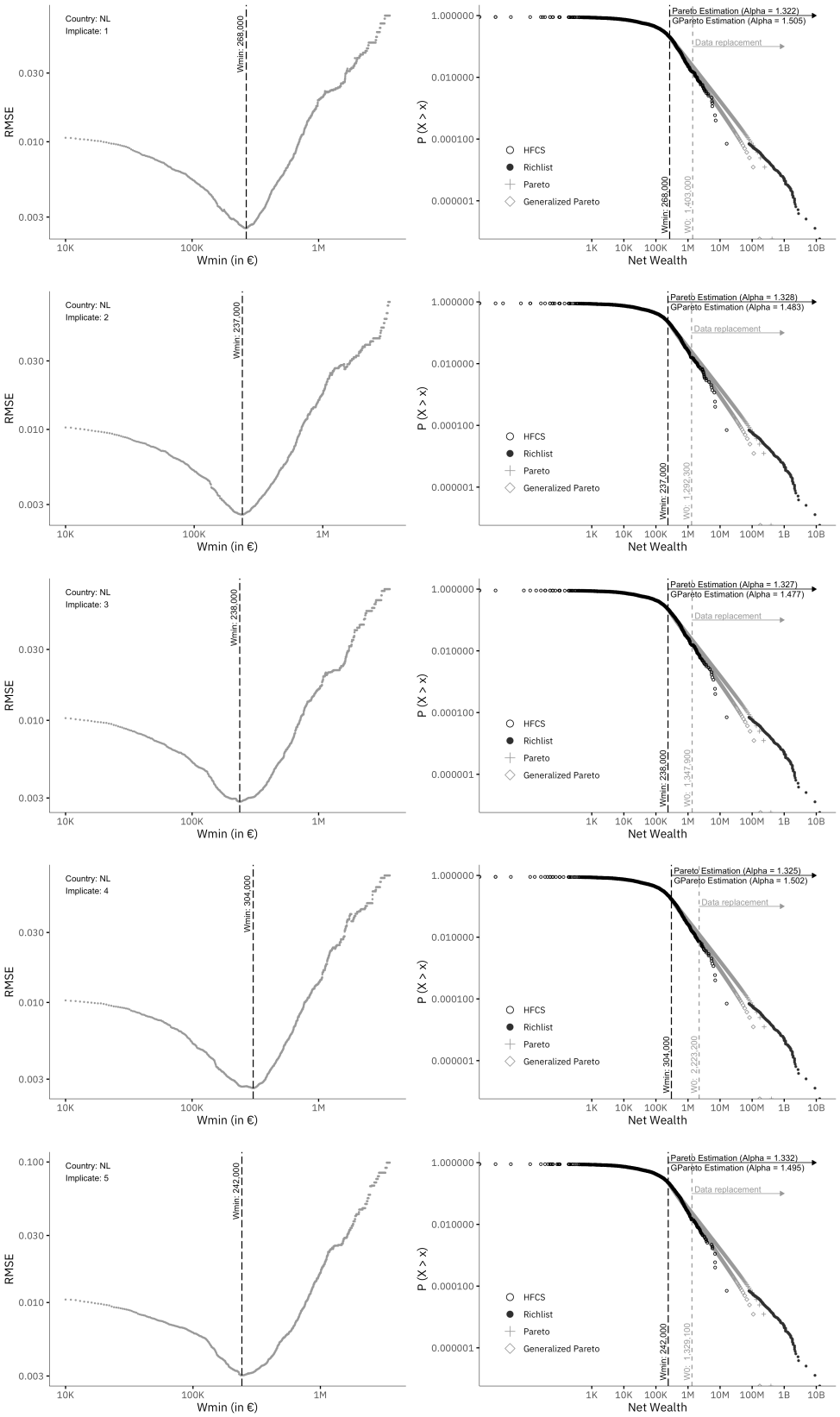


Figure B.14: CCDF and estimated parameters: PL

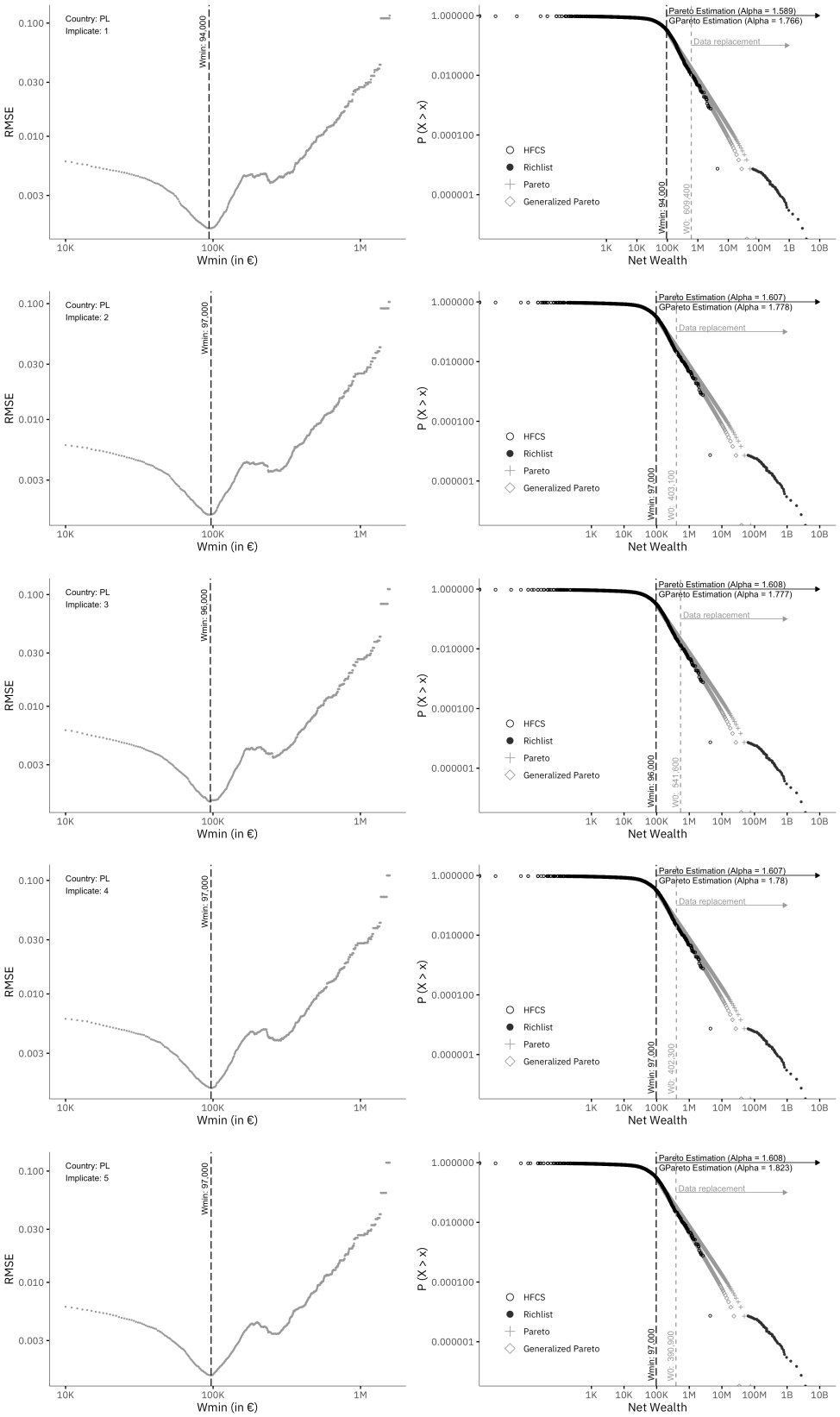


Figure B.15: CCDF and estimated parameters: PT

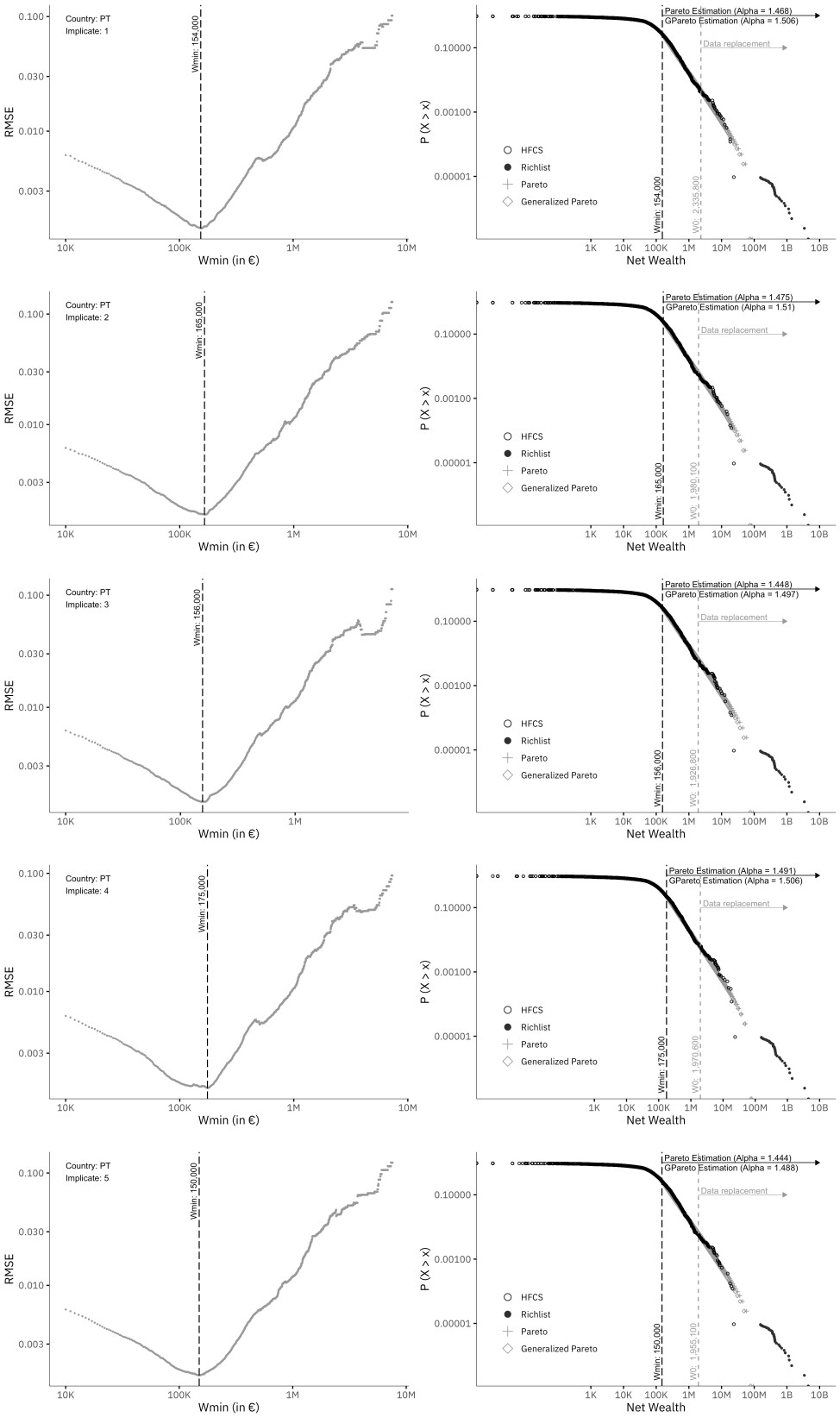


Figure B.16: CCDF and estimated parameters: SI