The missing rich:

Analysing thick tailed wealth and income survey data

London Inequality Workshop, 11 March 2022

Dr. Rafael Wildauer

Senior Lecturer in Economics

Deputy Director Institute of Political Economy, Governance, Finance and Accountability

UNIVERSITY of GREENWICH

Outline

Analysing thick tailed wealth and income survey data

1) Measuring the top tail with survey data: Advantages and Challenges
2) Methods: Using Pareto distributions to model the tail
3) Can we defend the Pareto null Hypothesis?
4) Results and applications

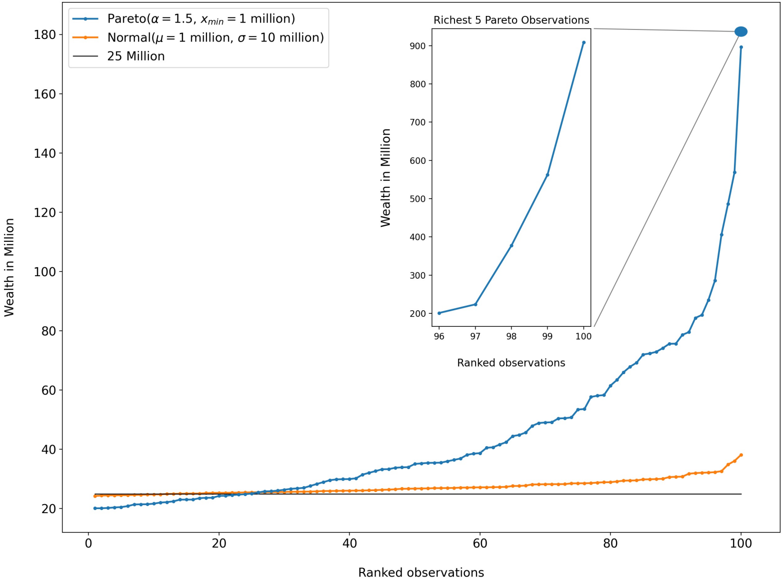
Measuring the top tail with survey data: Advantages and Challenges

What's a thick tail and why bother?

- A thick (or fat) tailed distribution produces more extreme values (billionaires) compared to thin tails
- For formal treatment see Beirlant et al. (2004, p. 49) and Langousis et al. (2016) using the concept of slowly varying functions[1]
- Why bother? Distributions without tick tails cannot explain observed inequalities (billionaires)
- Best illustrated with example:

[1] defined as any function $l(x)$ which is positive for large x, which satisfies $\frac{l(xu)}{l(x)} \rightarrow 1$ for any $u > 0$ as x tends to infinity.

Example: compare richest 100 observations from a sample of 10,000



Data sources for top wealth and incomes

1) **Survey data:** Cross section (*Household Finance and Consumption Survey, HFCS*; *Survey of Consumer Finances, SCF*) or panel data (*Wealth and Asset Survey, WAS*)
2) **Tax data:** Sample of anonymized tax data, or tabulated data
3) **Rich list data:** Compiled by journalists (*Forbes' List of Billionaires*)

Measuring the top tail

Table 1: Comparing data sources for top income and wealth

	survey data	tax data	rich list data
availability			
time period covered			
consistency across time and space			
additional information			
captures top wealth/income			

Measuring the top tail

Table 1: Comparing data sources for top income and wealth

	survey data	tax data	rich list data
availability	high for income and wealth	high for income, poor for wealth, very long time series	poor for income, high for wealth,
time period covered			
consistency across time and space			
additional information			
captures top wealth/income			

Measuring the top tail

Table 1: Comparing data sources for top income and wealth

	survey data	tax data	rich list data
availability	high for income and wealth	high for income, poor for wealth, very long time series	poor for income, high for wealth,
time period covered	US: 1989, EU: 2010	100 years plus	~1980s max
consistency across time and space			
additional information			
captures top wealth/income			

Measuring the top tail

Table 1: Comparing data sources for top income and wealth

	survey data	**tax data**	**rich list data**
availability	high for income and wealth	high for income, poor for wealth, very long time series	poor for income, high for wealth,
time period covered	US: 1989, EU: 2010	100 years plus	~1980s max
consistency across time and space	HFCS harmonised for EU22	fiscal income and wealth taxes differ	high for Forbes global list, low for others
additional information			
captures top wealth/income			

Measuring the top tail

Table 1: Comparing data sources for top income and wealth

	survey data	tax data	rich list data
availability	high for income and wealth	high for income, poor for wealth, very long time series	poor for income, high for wealth,
time period covered	US: 1989, EU: 2010	100 years plus	~1980s max
consistency across time and space	HFCS harmonised for EU22	fiscal income and wealth taxes differ	high for Forbes global list, low for others
additional information	household balance sheet, demographic info, other	poor, often confidential	limited (age, sector, country of residence) BUT name
captures top wealth/income			

Measuring the top tail

Table 1: Comparing data sources for top income and wealth

	survey data	**tax data**	**rich list data**
availability	high for income and wealth	high for income, poor for wealth, very long time series	poor for income, high for wealth,
time period covered	US: 1989, EU: 2010	100 years plus	~1980s max
consistency across time and space	HFCS harmonised for EU22	fiscal income and wealth taxes differ	high for Forbes global list, low for others
additional information	household balance sheet, demographic info, other	poor, often confidential	limited (age, sector, country of residence) BUT name
captures top wealth/income	ranges from poor to very good	income: good but tax evasion	by definition the very top

Measuring the top tail: Using survey data

- Survey data has attractive properties (availability, additional information, consistency) ...
- BUT is often plagued by two fundamental problems:
- **non-observation:** studying thick tails requires atypically large samples
  - example: top 0.1% in EU22 are roughly 220,000 households with networth above €10 million
  - a sample of 5000 households would only contain 5 individuals from this group (none from the top 0.01% and beyond)
- **differential non-response:** richer households less willing to participate
  - so in the context of the previous example this would mean we have a good chance not to observe anybody from the top 0.1% and beyond

Methods: Using Pareto distributions to model the tail

Using Pareto distributions to model the tail

- Fitting Pareto distributions can deal with **non-observation** and **differential non-response** - sometimes
- Sometimes because not all methods can solve both problems, some require additional data and fundamentally all rely on the null hypothesis that the tail follows a Pareto distribution
- So let's look at some of the key methodological choices and challenges

Which Pareto distribution?

- **Type I**
  - $CCDF(x) = Pr[X > x] = \left[\frac{x}{\sigma}\right]^{-\alpha}$
  - scale invariant: top $100 \cdot T$% of **tail population** own $T^{\frac{\alpha-1}{\alpha}}$% of tail wealth/income
  - for example: top 10% own $0.1^{\frac{1.5-1}{1.5}} = 46\%$ and the top 10% within the top 10% (i.e. the top 1% overall) own 46% of that etc.
  - **scale invariance means the pattern and degree of inequality stays the same within the tail**
- **Type II** and **Generalized Pareto Distribution**
  - $CCDF(x) = Pr[X > x] = \left[1 + \frac{(x-\mu)}{\sigma}\right]^{-\alpha}$
  - expression for share of top T% bit more involved
  - **crucially inequality within the tail not constant (scale variant)**

Determining the scale paramter (tail cut-off)

- Pareto distribution models the tail not entire range of wealth/income
- Where does the tail start?
- Several approaches:
  1) **Eyeballing** using log(rank) - log(wealth) graphs
  2) **Impose** at 'usual' cutoff points (e.g. 10%, 5%, 1%)
  3) **Statistical tests:** using Cramer von Mises type tests (Clauset et al. 2009) or RMSE (Langousis et al. 2016, Disslbacher et al. 2020)

Simple Type I Models

- Standard approach has been around for decades
- run the following OLS regression to obtain estimate for α:

$$ln(rank_i) = c - \alpha ln(wealth_i) + \epsilon_i \tag{1}$$

- Maximum Likelihood estimator with analytical solution exists as well but less robust
- Simple Type I Models yield unbiased (median) top wealth estimates in case of non-observation (Eckerstorfer et al. 2016) ...
- but they only partially alleviate differential non-response
- current standard: incorporate Gabaix's (2009) bias correction for $rank_i$. See Vermeulen (2018) and Wildauer & Kapeller (2021) for more intuitive presentation

Adding Rich Lists

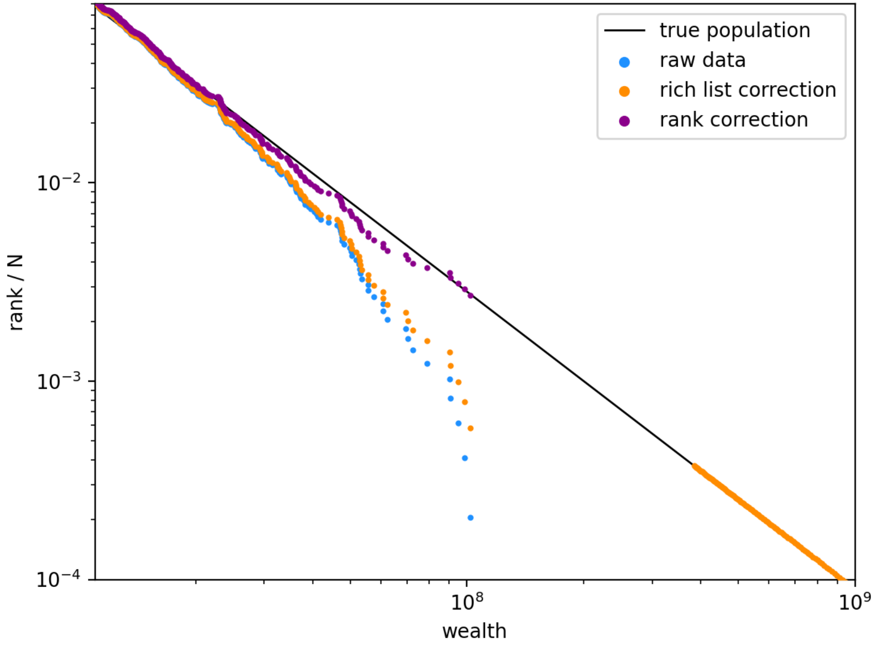
- Vermeulen (2018) introduced idea to add rich list data to survey data to tackle differential non-response
- Several incarnations:
  - Vermeulen (2018) fits type I via OLS
  - Langousis et al. (2016) fit Generalized Pareto via OLS (see section 2.2)
  - Heck et al. (2020) fit type II via Elemental Percentile Method (Castillo & Hadi 1997)
- Works well and tackles non-observation and differential non-response - in principle
- Fundamentally depends on the quality, availability and potentially consistency of rich lists

Rank Correction Approach

- Just published an alternative approach aimed at tackling non-observation and differential non-response
- Key feature: does not require rich lists (unknown quality, not available for some countries, very short on consistent basis)
- Idea is to impose a correction factor (u) on ranks which in some cases is interpreted as missing households at the top

$$ln(rank_i + u) = c - \alpha ln(wealth_i) + \epsilon_i \tag{2}$$

- Thus we called it rank correction approach (Wildauer & Kapeller 2022)

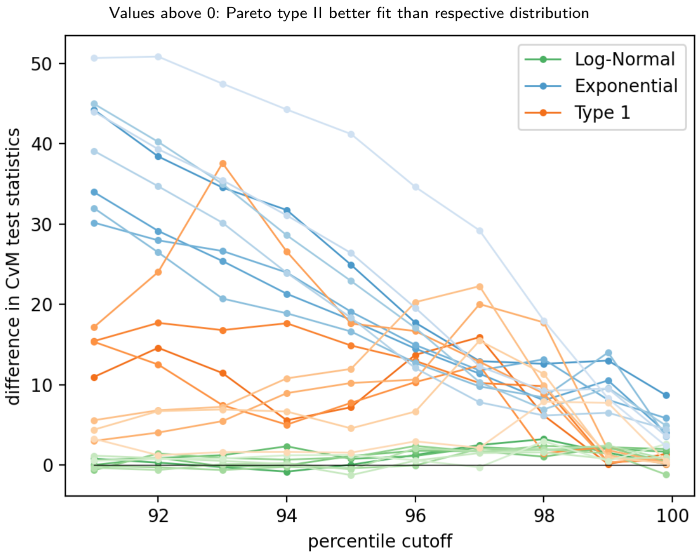

true population
raw data
rich list correction
rank correction
10^{-2}
10^{-3}
10^{-4}
rank / N
10^{8}
10^{9}
wealth

Is the Pareto null hypothesis supported by the data?

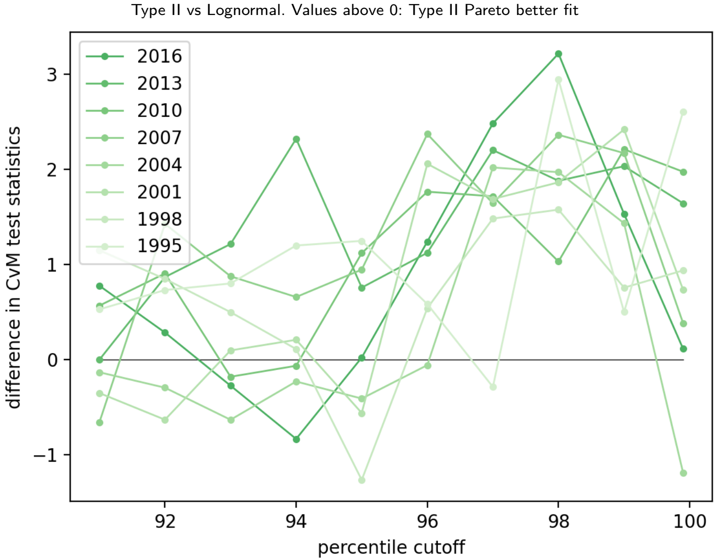
Can we defend the Pareto Null Hypothesis?

- Thick-tails of Pareto make crucial difference especially compared to e.g. lognormal
- Especially when used in theoretical models
- Two challenges when it comes to testing Pareto hypothesis:
  1) Differential non-response leads to 'spurious' rejection of Pareto null
  2) Since we need to estimate the parameters under null and sampling procedure is not public, can't compute standard p-values
- Some preliminary results where we (Ines Heck + myself) address these by:
  1) Use data which deals best with differential non-response: *Survey of Consumer Finances*
  2) Directly compare fit of pairs of distributions

Results from the Survey of Consumer Finances

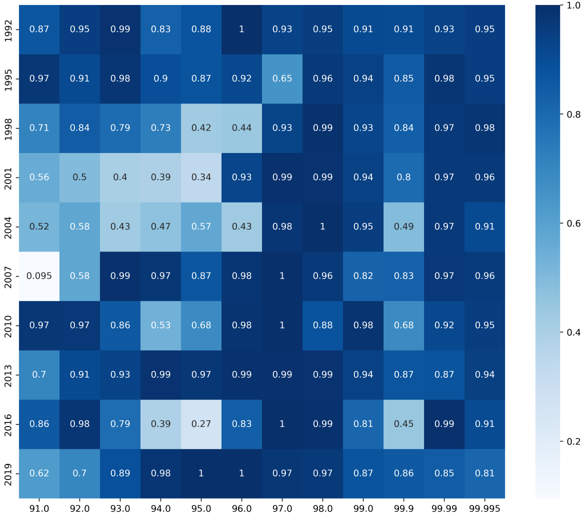


Results from the Survey of Consumer Finances



Is this statistically significant?

Type II vs Lognormal. Each cell based on 1000 replicate samples. How often is Type II fit better than LogN fit.



Pareto Hypothesis

- Alive and well especially for top 3% - 4%
- We will see later, that using these methods mostly adds wealth in top 1%
- Extension to income and other countries

Results from fitted Pareto tails and Applications

What's it all for?

- We have reviewed the methods
- we are confident about our Pareto null ...
- How can we use this apparatus? I will show:
  1) Estimating income and wealth distributional statistics. Piketty et al. (2022) see that as main goal of inequality research to inform political debate.
  2) Estimate revenues of (wealth) taxes
- Furthermore:
  3) Contribute to construction of Distributional National Accounts (DINA) as in (Piketty et al. 2018)
  4) Provide basis to incorporate thick tails into theoretical macro models

Distribution of Wealth in the EU

- We know a lot about US, UK, France ...
- much less about EU as a whole.
- In Kapeller et al. (2021) we make two contributions:
    1. Estimate wealth distribution for the EU22[2] (90.7% of EU27 GDP) by fitting Pareto tails
    2. Based on that calculate revenues for four wealth tax designs

[2] EU27 minus Bulgaria, Czechia, Denmark, Romania and Sweden.

Who is who?

- Detailed distribution tabulations for all 22 countries in ▸ Online Appendix
- For now let's focus on the EU22 distribution:
  - The poorest 20% of the population: $\leq$ €$7,000$
  - The poorest 50% (median): $\leq$ €$90,000$
  - The richest 10%: $\geq$ €$490,000$
  - The richest 3%: $\geq$ €$1,039,000$
  - The richest 1%: $\geq$ €$2,153,000$
- Keep in mind net wealth: house worth €$700,000$ with mortgage of €$500,000$ means net wealth of €$200,000$

Who owns how much?

- the richest 1% of households hold 32% of total wealth in the EU22
- some individual countries:
  - Italy: 27%
  - Poland: 33%
  - Germany: 38%
- how does that compare?
  - South Korea: 25% (2015)
  - China: 30% (2015)
  - USA: 35% (2017)
  - Russia: 43% (2015)
- Europe is much more unequal than we like to think

Comparing our results to other data sources

Table 3: Assessing the model fit

German top wealth shares	Raw survey*	Survey + Pareto*	Schröder et al 2020*
Top 1%	18.6%	37.7%	35.3%
Top 5%	40.8%	55.2%	54.9%
Top 10%	55.4%	66.3%	67.3%
French top wealth shares	**Raw survey***	**Survey + Pareto***	**Garbinti et al 2020***
Top 1%	17.1%	27.5%	23.4%
Top 5%	35.5%	43.9%	43.1%
Top 10%	49.2%	55.9%	55.3%
	Raw survey**	**Survey + Pareto****	**Krenek and Schratzenstaller 2018****
Total wealth EU22	35,713	43,629	49,599
	Raw survey	**Survey + Pareto**	**National rich lists**
Billionaires in the EU22	0	461	431

*% of total wealth holdings, **€bn. Source: raw survey estimates are from the HFCS's third wave and the survey + pareto results are based on the authors' calculations (eg. Table 2).

Table 5: Wealth Tax Designs

	Model I "flat tax"	Model II "mildly progressive"	Model III "strongly progressive"	Model IV "wealth cap"	
Approach	Flat rate	Progressive rate – slowing growth of inequality	Progressive rate – reducing inequality	Progressive rate – introducing a wealth cap	
% of population exempt	97%	97%	99%	59%	
Tax brackets		**Tax rates**		**Tax brackets**	**Tax rates**
from €1 million €1 million ≈ top 3% or 5.4 million households	2%	1%		0.5 times av. wealth	0.1%
from €2 million €2 million ≈ top 1% or 1.9 million households		2%	2%	2 times av. wealth	1%
from €5 million € 5 million ≈ top 0.3% or 550,000 households		3%	3%	5 times av. wealth	2%
from €10 million €10 million ≈ top 0.1% or 220,000 households			5%	10 times av. wealth	5%
from €50 million €50 million ≈ top 0.01% or 23,000 households			7%	100 times av. wealth	10%
from €100 million €100 million ≈ top 0.005% or 9,000 households			8%	1,000 times av. wealth	60%
from €500 million €500 million ≈ top 0.001% or 1,200 households			10%	10,000 times av. wealth	90%

Average wealth in the EU22 is €260,000 (based on Pareto tail amended data). The tax brackets for model IV therefore start at €130,000 (0.5 times average); €520,000 (2 times the average); €1.3 million (5 times the average); €2.6 million (10 times the average); €26 million (100 times the average); €260 million (1,000 times the average) and €2.6 billion (10,000 times the average).

Revenue estimation

		Survey data + Pareto tail	Survey data + Pareto tail + evasion effects
model I: flat tax	€ bn.	271	192
	% GDP	2.3%	1.6%
model II: mildly progressive	€ bn.	316	224
	% GDP	2.7%	1.9%
model III: strongly progressive	€ bn.	505	357
	% GDP	4.3%	3.0%
model IV: wealth cap	€ bn.	1,837	1,281
	% GDP	15.5%	10.8%

Thank you!

r.wildauer@gre.ac.uk

@RafaelWildauer

References I

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. & DeWaal, D. (2004), *Statistics of Extremes*, John Wiley & Sons.

Castillo, E. & Hadi, A. S. (1997), 'Fitting the generalized pareto distribution to data', *Journal of the American Statistical Association* **92**(440), 1609–1620.

Clauset, A., Shalizi, C. R. & Newman, M. E. J. (2009), 'Power-law distributions in empirical data', *SIAM Review* **51**(4), 661–703.

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).

References II

Eckerstorfer, P., Halak, J., Kapeller, J., Schuetz, 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. (2009), 'Power laws in economics and finance', *Annual Review of Economics* **1**(1), 255–294.

Heck, I., Kapeller, J. & Wildauer, R. (2020), 'Vermögenskonzentration in Österreich – ein update auf basis des hfcs 2017', *AK Wien Working Paper Series* .

Kapeller, J., Leitch, S. & Wildauer, R. (2021), 'A european wealth tax for a fair and green recovery', *Greenwich Papers in Political Economy* (81).

References III

Langousis, A., Mamalakis, A., Puliga, M. & Deidda, R. (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.

Piketty, T., Saez, E. & Zucman, G. (2018), 'Distributional national accounts: Methods and estimates for the united states', *Quarterly Journal of Economics* **133**(2), 553–609.

Piketty, T., Saez, E. & Zucman, G. (2022), 'Twenty years and counting: Thoughts about measuring the upper tail', *Journal of Econoimc Inequality* .

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

References IV

Wildauer, R. & Kapeller, J. (2021), 'A comment on fitting pareto tails to complex survey data', *Journal of Income Distribution* .

Wildauer, R. & Kapeller, J. (2022), 'Tracing the invisible rich: A new approach to modelling paretotails in survey data', *Labour Economics* .

Appendix

Accounting for tax evasion

- Based on the literature we assume the following proportion of the tax base is lost due to evasion:
- real estate 20%, financial wealth 24%, directly held companies 13% and other assets 100%
- in addition we model strong evasion as: real estate 20%, financial wealth 48%, directly held companies 26% and other assets 100%