Utilizing Pareto Interpolation to Compute Top-Income Shares from Censored Tax Data

Fabian Blasch

Specialization Course - Economics of Distribution

02/26/2020

Contents

1 Introduction

Inequality of wealth and income is a topic of great interest in economics. Unfortunately, however, the data availability often does not allow for a close examination of the underlying distribution of wealth and income. In the case of the wealth distribution this mostly stems from the fact that wealthy people have no interest to share information on their belongings. Fortunately, in the case of wage income there is a source of reliable information in the form of wage income tax statistics. Said data can be used as a source to examine the concentration of wage income at the top end of the distribution. However, in order to obtain top-income shares one needs to interpolate censored gaps in the reported data.

This short research note aims to use similar methods as Altzinger, Berka, Humer, and Moser (2012) to obtain top-income shares from 2010 until 2018. First off, a concise summary of the related literature is included. Secondly, the data structure is shortly discussed, then the theory behind the methods is displayed and in the end the results of the extension until 2018 are presented.

2 Related Literature

Vilfredo Pareto already examined income distributions in 1896 (Pareto, 1896). He was of the opinion that the income distribution is a law bound to human nature and argued that independent from changes in the economic system the distributions of income would always follow a Pareto distribution. Further research then showed that even though the close relative of the exponential distribution was a good fit, other distributions exist that better approximate a nations wage income distribution. What remains, however, is the ability to accurately estimate top income share as Altzinger et al., 2012, A. Atkinson and Piketty (2007), and Piketty and Saez (2003) displayed in their research. Since most developed states collect wage income tax, the utilization of tax filings for inequality calculations is a frequently applied procedure. One of the most prominent papers examines the changes in wage and partly in capital income in the United States of America between 1913 and 1998 (Piketty & Saez, 2003). Their main findings are that top wage income shares were stagnant previous to World War II and then dropped significantly during the war. It then took until 1970 to get to pre-WWII levels and from then onwards the top-income shares increased steadily. Besides just examining the wage income time-series of the U.S, the authors were also interested in whether this dynamic could be representative for other developed countries' long run inequality development. Examinations of wage income inequality by Piketty (2003) and A. B. Atkinson, 2001 show that the development of top wage income shares in France and Great Britain is quite similar for the same time span, with a few distinct differences like the intensity of the shock to top-income shares that was caused by WWII. In the case of France and GB the beginning of the war had a higher magnitude impact on top-income shares. Unfortunately the time-series available for Austria is too short to evaluate the magnitude of the effect of the second world war, however, Altzinger et al., 2012 also find that inequality in wage income started to increase from 1970 onwards.

3 Data

Zeile	Stufen der Bruttobezüge in 1.000 EUR			Bruttobezüge insgesamt		Durchschnittlicher Jahresnettobezug	darunter		Steuerfreie Bezüge gem. § 68		Bezüge gem. § 67 Abs. 1 und 2 vor Abzug der Sozialversicherungsbeiträge	
							Vollzeitbeschäftigung	Teilzeitbeschäftigung				
				Fälle	1.000 EUR	EUR	Fälle	Fälle	Fälle	1.000 EUR	Fälle	1.000 EUR
1	0	bis unter	2	564.855	478.877	782	136.726	185.581	47.410	2.984	396.120	50.173
2	2	bis unter	4	300.245	881.056	2.682	85.526	108.957	47.215	5.979	266.474	102.014
3	4	bis unter	6	279.717	1.387.732	4.527	50.007	109.057	43.050	8.176	268.393	175.927
4	6	bis unter	8	247.614	1.737.848	6.253	67.344	66.104	45.766	10.707	242.402	222.026
5	8	bis unter	10	268.805	2.423.201	8.005	73.350	65.141	50.294	14.581	265.922	307.020
6	10	bis unter	12	393.303	4.316.190	9.932	72.816	66.809	55.571	18.593	391.254	518.313
7	12	bis unter	15	433.705	5.851.958	11.934	105.229	105.282	92.138	37.499	431.690	787.064
8	15	bis unter	18	423.441	6.985.820	14.216	123.137	96.544	106.226	53.330	422.053	945.655
9	18	bis unter	20	273.151	5.187.335	15.769	100.839	53.394	79.248	47.916	272.504	709.026
10	20	bis unter	25	653.650	14.707.286	17.891	294.632	93.052	215.320	168.285	652.366	2.011.230
11	25	bis unter	30	607.956	16.686.992	20.754	349.701	58.179	249.509	245.732	607.031	2.272.543
12	30	bis unter	35	501.137	16.221.488	23.655	322.296	34.475	238.378	309.017	500.559	2.190.979
13	35	bis unter	40	356.800	13.321.204	26.537	248.478	20.924	185.147	297.538	356.403	1.779.822
14	40	bis unter	50	422.582	18.780.062	30.460	325.953	21.666	238.381	435.103	422.089	2.469.618
15	50	bis unter	70	368.803	21.427.966	38.123	308.630	8.807	218.386	355.889	368.259	2.770.651
16	70	bis unter	100	160.051	13.065.752	52.494	140.108	-	101.305	158.460	159.728	1.646.894
17	100	bis unter	150	55.966	6.592.556	74.857	50.758	-	35.796	63.565	55.753	785.429
18	150	bis unter	200	11.776	2.003.574	108.515	10.926	-	7.175	11.339	11.699	223.394
19	200	und mehr		9.654	3.179.647	206.148	9.085	-	5.850	8.541	9.576	317.284
20	**Insgesamt**			**6.333.211**	**155.236.545**	**18.008**	**2.875.541**	**1.093.972**	**2.062.165**	**2.253.233**	**6.100.275**	**20.285.062**

Figure 1: Censored Wage Income Tax Data (2010)

Figure 1[1] (Fischer, 2010) displays an example for the censored income wage table that is scraped from the yearly released "Statistik der Lohnsteuer" utilizing the R package "tabulizer" Leeper (2018). The brackets are more narrow towards the bottom of the distribution since more people are part of those income groups. For each bracket the amount of people as well as the total gross wage earned is reported. To be precise only people who are employed and are eligible to pay into the tax system are recorded, this includes people in retirement and also people that earn less than the tax exemption amount. Even though the monetary contribution of said people into the system is

[1] An R list object containing the required data points for the calculations of this research note can be found here.

practically zero their income will still be captured by the gross income displayed in figure 1. Additionally, it is also crucial to mention that taxable income is not only considered as wages paid to an employee but also awarded bonuses, separation benefits, factored leave compensation, as well as taxable benefits received as material compensation such as company owned cars, free gym memberships, free meals and every other non-monetary remuneration. The remaining information to the right of gross income ("Bruttobezüge Insgesamt"), is not of particular importance to this work since it is not required to calculate top-gross-income shares. The PDFs that hold these tables can be found on the Statistics Austria Homepage.

4 Theory

Besides data availability another issue renders calculating top-income shares across time quite difficult. Nominal wages are always subject to small and steady increases over time. This makes it hard to accurately examine income concentration across time since people move upwards through the income brackets which are constant (Altzinger et al., 2012). Figure 2 displays exactly this problem. When plotting the distribution with the upper bound of each bracket in a weighted empirical distribution function, one can clearly tell that people move upwards through the brackets when observing that the year for each bracket that is the furthest to the bottom is almost always the latest in time. In other words, the people that earn equal or less than the upper bound decrease for later points in time which means that the people that earn more increase[2].

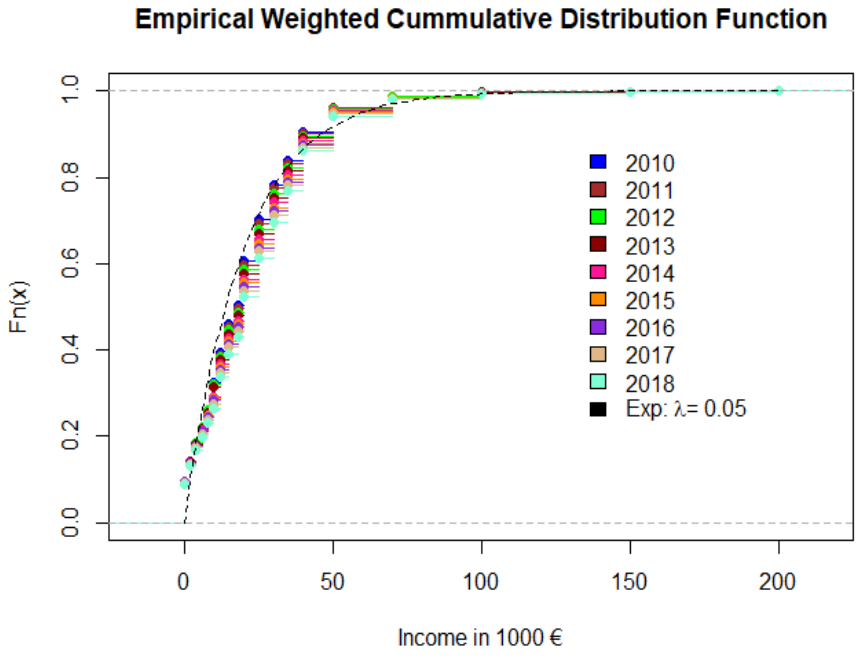

Figure 2: Weighted Empirical CDF 2010-1018

[2]Please note that displaying the distribution in this way is not an accurate depiction of the actual distribution but just a graph that is used to illustrate the point of increasing nominal wages.

Fortunately, when utilizing Pareto interpolation no adjustments for wage increases are necessary since the figures used to compute the top-income shares are always relative to the total population and to the total income earned.

Comparing actually realised top-income shares to the ones obtained by utilizing Pareto interpolation, Altzinger et al. (2012) showed that the Top-10%, Top-5% and Top-1% offer sufficiently accurate estimates for the time period between 1964 and 2009. Since I do not have access to data for a similar comparison I have to rely on the fact that the distribution of the past 45 years allows for the assumption that the top 20% are still sufficiently accurately Pareto distributed.

Assuming a Pareto distribution for the top end of the income distribution one can derive Van der Wijks Law (Cowell, 2011),

$$\hat{\alpha} = \frac{\bar{y}}{\bar{y} - \tilde{y}}. \tag{1}$$

Where $\bar{y}$ represents the average income above the upper bound of an income bracket $\tilde{y}$. When recalling the data structure visible in figure 1, one can see that we have information on the people within each bracket as well as the gross income. It follows that one can make use of the properties of the mean which can be quickly derived from the properties of the sum operator. The mean of the income above a threshold can be calculated as the weighted mean of the means of income brackets above the threshold $\tilde{y}$.

The obtained alphas can then be used to compute top-income shares using Atkinsons formula (A. Atkinson & Piketty, 2007),

$$S_i = \left(\frac{H_i}{H_j}\right)^{\frac{\hat{\alpha}-1}{\hat{\alpha}}} \times S_j. \tag{2}$$

In the equation depicted above S_i and H_i represent the income share and population share of interest. H_j is the closest population share to the one of interest that can be extracted from the data without interpolation and S_j is the matching income share.

5 Results

5.1 Income Concentration from 1964-2009

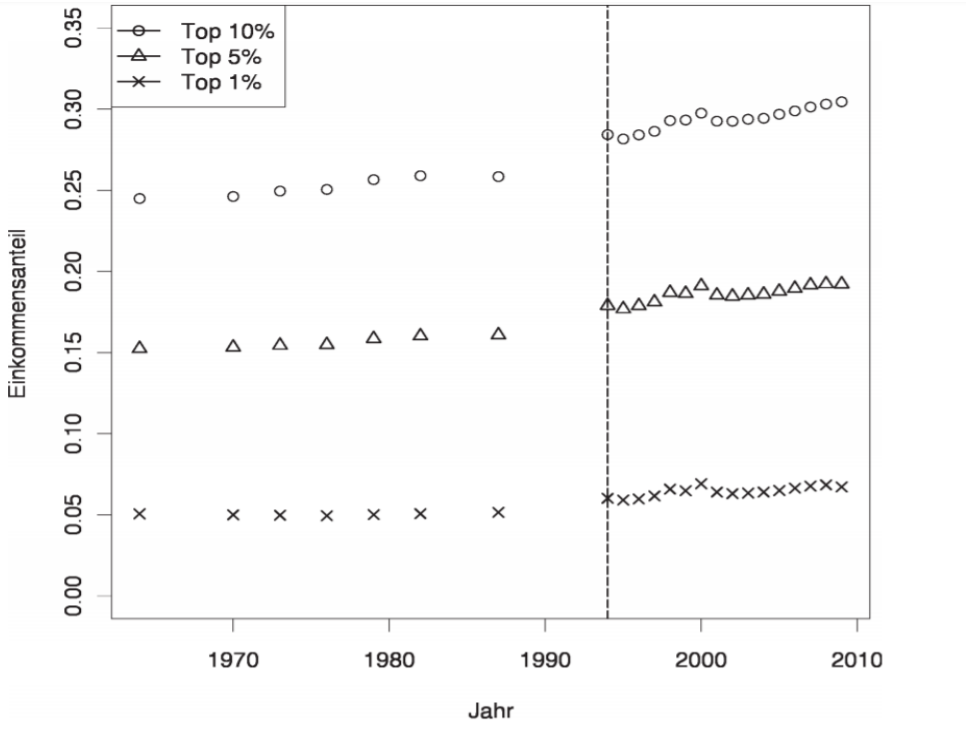

Figure 3: Top-income Shares 1964-2009

Altzinger et al. (2012) demonstrated that from around 1975 onwards income concentration increased, the graph that they provide displays the share of the Top-10%, Top-5% and Top-1% of income. The authors argue that one of the main reasons for the increase in income concentration at the top of the distribution is due to increase in productivity which affect the income of high earners more significantly in comparison to people at the lower end of the distribution.

5.2 Income Concentration from 2010-2018

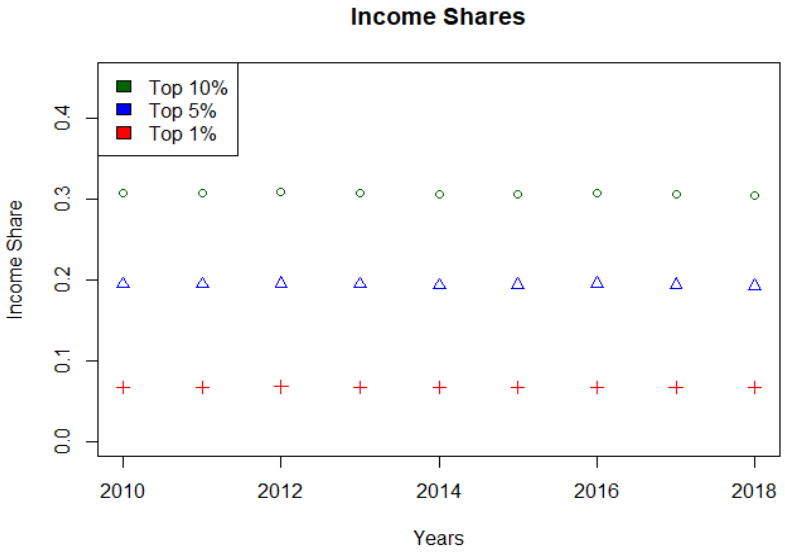

Figure 4: Top-income Shares 2010-2018

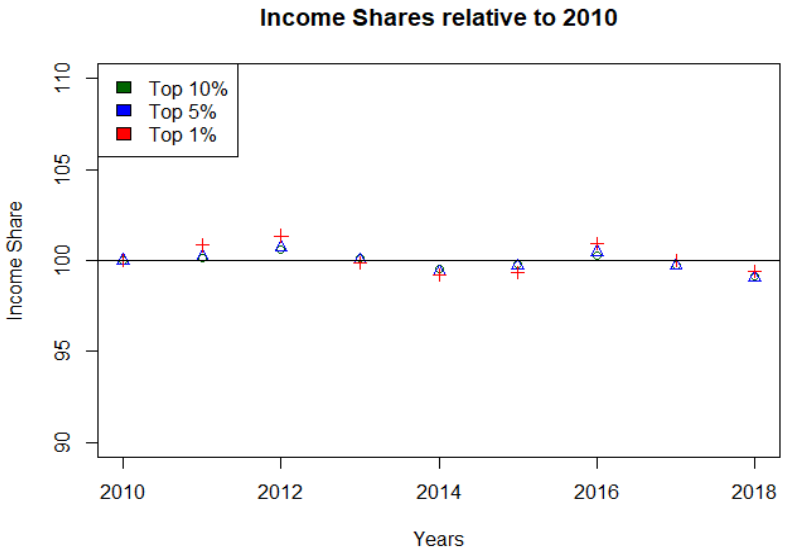

Figure 5: Top-income Shares 2010-2018

As observable in figure 4 and 5 the income concentration did not change significantly between 2010 and 2018. Figure 4 depicts that the income shares for the top ten percent consistently hovered around 31% of total income, whereas the top-5% and top-1% earned about 20% and 7% respectively. When considering that figure 4 is not optimal to detect subtle trends, figure 5 shows that there is no visibly detectable trend when plotting the change in income concentration relative to 2010.

6 Conclusion

This short research note extends a part of the work by Altzinger et al. (2012) until 2018. The main results are that income inequality seems to be stagnant when examining wage income data. However this does not mean that there was no change in inequality in Austria. Firstly, when examining wage income data, it is important to keep in mind that an increase in unemployment can ceteris paribus be considered as an increase in inequality. However, in the case of tax statistics, someone who does not file taxes will not be part of the income distribution. Additionally, it is also important to mention that the majority of inequality is assumed to result from difference in wealth and not income. Accordingly, this short research note is to be interpreted as a small part of a much bigger picture.

Additionally, when writing about Pareto interpolation it is crucial to mention that Blanchet, Fournier, and Piketty (2017) present a new interpolation approach. They describe it as generalized Pareto interpolation, which can also be used to obtain top-income shares. Comparing it to the constant Pareto coefficient interpolation commonly applied in the literature and also in this reseach note, they provide evidence that this newly developed method allows for more accurate approximations, since it does not rely on a constant Pareto coefficient. However, since the main purpose of this short research note was to have a smoothly interpretable extension of the work by Altzinger et al. (2012) a change of method was not considered. Nevertheless, the application of this

recently discovered approach seems like a promising endeavour for future wage income inequality research.

References

Altzinger, W., Berka, C., Humer, S., & Moser, M. (2012). Die langfristige entwicklung der einkommenskonzentration in österreich, 1957-2009 teil ii: Methodik und ergebnisse. *Wirtschaft und Gesellschaft*, *38*(1), 77–102.

Atkinson, A. B. (2001). Top Incomes in the United Kingdom over the Twentieth Century. *Nuffied College*.

Atkinson, A., & Piketty, T. (Eds.). (2007). *Top incomes over the twentieth century: A contrast between continental european and english-speaking countries*. Oxford University Press.

Blanchet, T., Fournier, J., & Piketty, T. (2017). Generalized pareto curves: Theory and applications.

Cowell, F. (2011). *Measuring inequality* (3rd ed.). Oxford University Press.

Fischer, F. (2010). Statistik der Lohnsteuer. *Statistik Austria*.

Leeper, T. J. (2018). *Tabulizer: Bindings for tabula pdf table extractor library*. R package version 0.2.2.

Pareto, V. (1896). Cours d'Economie Politique. Tome Premier. *The Economic Journal*, *6*(22), 249–253.

Piketty, T. (2003). Income inequality in france, 1901–1998. *Journal of political economy*, *111*(5), 1004–1042.

Piketty, T., & Saez, E. (2003). Income Inequality in the United States, 1913–1998*. *The Quarterly Journal of Economics*, *118*(1), 1–41.