A Pareto upper tail for capital income distribution

Bogdan Oancea, Dan Pirjol, Tudorel Andrei

PII: S0378-4371(17)30926-3
DOI: https://doi.org/10.1016/j.physa.2017.09.034
Reference: PHYSA 18639

To appear in: *Physica A*

Received date : 3 June 2017

Please cite this article as: B. Oancea, D. Pirjol, T. Andrei, A Pareto upper tail for capital income distribution, *Physica A* (2017), https://doi.org/10.1016/j.physa.2017.09.034

Cover letter and Highlights

We give the first determination of the capital income distribution in Romania using individual income tax data.

The capital income inequality dominates the total income inequality.

The tail distribution of the capital incomes is a Pareto distribution with Pareto exponent 1.4.

A Pareto upper tail for capital income distribution

Bogdan Oancea[*], Dan Pirjol[†], Tudorel Andrei[‡]

June 3, 2017

Abstract

We present a study of the capital income distribution and of its contribution to the total income (capital income share) using individual tax income data in Romania, for 2013 and 2014. Using a parametric representation we show that the capital income is Pareto distributed in the upper tail, with a Pareto coefficient $\alpha \sim 1.44$ which is much smaller than the corresponding coefficient for wage- and non-wage-income (excluding capital income), of $\alpha \sim 2.53$. Including the capital income contribution has the effect of increasing the overall inequality measures.

Key words: capital income distribution; Pareto tail; income inequality

1 Introduction

The related subjects of income and wealth inequality have received considerable attention recently, both as topics of public debate and for their role in economic and social theory [32, 2]. The total income can be represented largely as the sum of the labor income and of the capital income, and their

[*] The University of Bucharest, Romania, email:bogdan.oancea@faa.unibuc.ro

[†] National Institute for Nuclear Physics and Technology, Romania, email: dpirjol@gmail.com

[‡] The Bucharest Academy of Economic Studies, 6 Piata Romana, Sectorul 1, 010374 Bucharest, Romania, email:andrei.tudorel@csie.ase.ro

ratio is a measure of the relative weights of the two components in the GDP (the labour-capital split). The labor-capital split and its role in economic growth is a subject of intense debate nowadays [32, 24]. There is also an active interest in the study of the capital income and of its role in the theories of income distribution [32, 5]. This is due especially to the fact that the upper tail of the income distribution is dominated by capital income, and this observation motivates the increased interest in the study of the capital income and its distribution.

As several authors noted [26, 32], the inequality in income distribution is significantly greater for the capital income than for other types of income. It is also worth mentioning that the capital income distribution is also important for its determinant role in the personal income inequality, as shown in [5, 15, 39]. The authors of these studies present evidence that there is a strong link between the aggregate role of capital in the economy and the distribution of income.

In a recent paper [29] we presented a first study of the income distribution in Romania, using tax income data at individual level. This study included only the contributions of wages and social redistribution income (pensions, unemployment and social benefits). In this paper we complete the study of the income distribution by including also the capital income. We present an analysis of the capital income shares to the total income, and study the distribution of the capital income. Furthermore, we expand the data coverage by presenting data for income distribution in Romania for 2013-2014.

The first studies of the income distribution are traced back to the work of the Italian economist Vilfredo Pareto [31] who noticed that the upper tail of the income distribution appears to be well described by a power law. This is known as Pareto distribution, and the exponent of the power law is known as the Pareto coefficient. Later studies have shown a dependence of the Pareto coefficient on the time period and country studied.

Several distributions appearing in economics and social sciences appear to be well reproduced by Pareto distributions in their upper tail. For example, this was observed for the size distribution of cities [18], firm sizes [4, 41] and stock market movements [42], [20]. A comprehensive introduction into the various applications of the power laws to different areas of economics, as well as the a survey of the mechanisms responsible for their emergence is given in [17]. In the area of income distribution there is a general accepted idea that the upper tail follows a Pareto distribution [30], [3], [34], [35], [6], [16], [23], [22], [7, 8], [1], [13], [10], [11], [21], [44], [43].

The main results of our study are that the capital income distribution has a greater inequality than other types of income, and it is well described by a Pareto distribution in its upper tail. We present a distributional analysis of the capital income, showing that it can be decomposed into three distinct regions, and determine the types of income which dominate the contribution in each region.

The paper is organized as follows. In Section 2 we give a descriptive analysis of our data sets, giving results for the quantile distributions of the various types of incomes considered, and their inequality measures. In Section 3 we present a distributional analysis of the capital income data series, using a parametric representation which includes a Pareto distribution in the large incomes region. Section 4 discusses the conclusions of the study of the paper.

2 Income Data

We used for this analysis individual tax income data for 2013 and 2014. Three main categories of income have been aggregated: wages (A), non-wage (B), and capital (C). A detailed list of the types of income included in the study is given in the Appendix A. The main contributors to the capital income are interest income, dividends, and income from real estate sales.

We show in Table 1 the summary statistics of the partial and total incomes of each type A, B, C for 2013 and 2014, determined from an analysis of the gross personal tax income data. The income data is quoted on an annualized basis. The numbers of tax payers receiving income of each type are denoted $N_{A,B,C}$. Because each given person can receive income of several types, the total number of tax payers is different from the sum $\sum_{i=A,B,C} N_i$.

Table 1 lists the average incomes of each type $\bar{X}_i = X_i/N_i$, their standard deviation, the numbers of tax payers, and the Gini coefficients for the partial and total income. We compute averages relative to the number of persons receiving each type of income.

One can compare with the results for 2013 in [29]. In this paper the average wage income was quoted as 19,413 RON and the average non-wage income was 3,025 RON. The average wage is consistent with the result in Table 1, while the difference noted for the non-wages result is larger due to the inclusion of other types of income. In [29] this income category included only i) pensions income, ii) unemployment benefits and iii) social benefits.

The capital income contributes a significant proportion to the total income. The weights of the three types of income A, B, C in the total income are $(56.6\%, 24.8\%, 18.6\%)$ for 2013 and $(56.0\%, 24.4\%, 19.5\%)$ for 2014. We note that the capital income has a large standard deviation, which suggests a much broader distribution than that of the other types of income. This can be seen also in the Gini coefficient of this income, which is larger than that of other income types. The Gini coefficients obtained in our case (0.92 and 0.93) are greater than the values for other countries: for example in [15] it is reported that the average value for this coefficient during the period 1980-2003 is 0.75 for U.K., 0.78 for U.S. and 0.81 for Germany.

The large value of the inequality coefficient for the capital income explains the increase of the Gini coefficient of the total income, after including the capital income contribution. Additional measures of income inequality are presented in Table 7.

Figure 1 presents the Lorenz curves for the total and capital income. This graphical representation illustrates clearly that the capital income is much more inequally distributed than the total income. The Lorenz curves are tabulated in numerical form in Table 6.

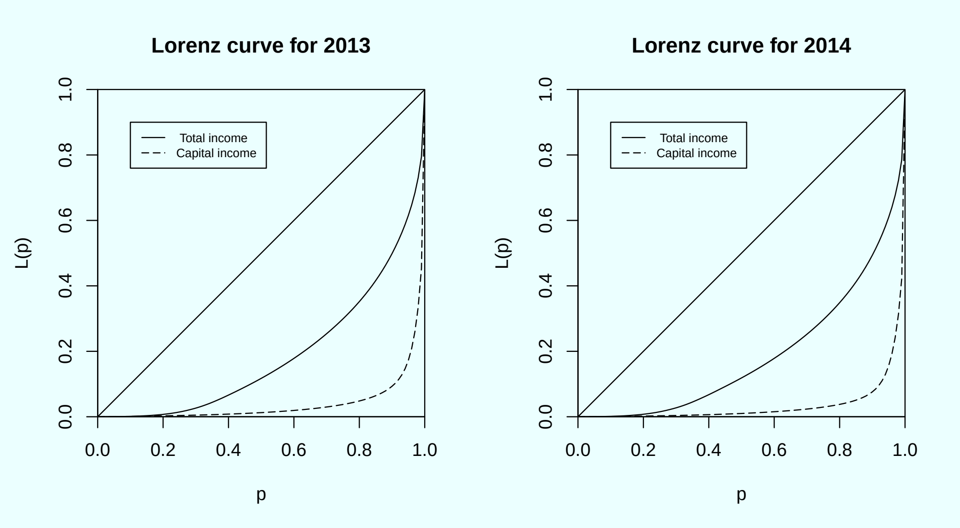

Figure 1: Lorenz curve for income distribution for 2013 and 2014. The Lorenz curve of the total income is shown as the solid curve, and the Lorenz curve of the capital income is shown as the dashed curve.

Commonly used measures of inequality are the ratios of the top 10%/top

Table 1: Summary data for the total income for 2013 and 2014: mean and standard deviation of the yearly income, the number of tax payers, and the Gini coefficient. The numbers of tax payers receiving each type of income are denoted $N_{A,B,C}$. Due to overlap between different types of income, their sum is different from the total number of tax payers N.

Parameter	2013	2014
Average Income		
Wages	19,411.8	20,623.1
Non-wages	12,037.3	12,017.2
Capital	11,464.1	13,752.1
Total	19,646.0	20,826.4
Standard Deviation of Income		
Wages	32,005	33,361
Non-wages	48,037	49,356
Capital	1,236,169	1,730,903
Total	698,923	942,826
Number of taxpayers		
N_A	6,217,274	6,302,494
N_B	4,397,170	4,719,291
N_C	3,455,928	3,298,395
N	10,853,791	11,142,053
Gini coefficient		
Wages	0.533	0.530
Non-wages	0.550	0.558
Capital	0.921	0.932
Total	0.625	0.627

90% and top 1%/top 99% income shares[1]. Table 2 presents the numerical values of these two ratios for all three types of income separately, and for the total income. From this table one can see again that this ratio is much

[1] We define them as the ratio of the total income in the 0-10% decile to the total income in the 90%-100% decile, and analogous for the top 1%/99% income shares. Another definition found in the literature uses the ratios of the average incomes in the respective deciles.

larger for the capital income than for the other types of income. The results also point to a slight increase of the inequality of the capital income in 2014 compared with the previous year.

Table 2: The values of the two ratios top 10% / top 90% and top 1% / top 99% income shares, separately by type of income, and for the total income.

	Wages (A)	Non-wages (B)	Capital income (C)	Total Income
2013				
top 10%/90%	83.21	189.61	806.39	382.92
top 1%/99%	2591.24	2329.77	5410.87	4263.01
2014				
top 10%/90%	78.91	208.4	1005	387.26
top 1%/99%	2487.23	2795.66	6827.21	4640.23

Finally, we present also the relative contributions of the capital income to the total income, separated by deciles of the total income distribution. Figure 2 shows, for each decile of the total income, the capital income shares.

3 Capital income distribution

We study in this Section the details of the distribution of the capital income. We start by introducing a few notations. Denote $p_i(x)$ the probability distribution function (pdf) of the type i of income: $i = A, B, C$. The number of persons with income of type i in the range $[x, x + dx]$ is $N_i p_i(x)dx$ where N_i is the number of people paying type-i income. The total income of type i paid is $X_i = \int_0^\infty p_i(x)xdx$. The distribution functions are normalized as $\int_0^\infty p_i(x)dx = 1$. We also denote the corresponding cumulative distribution functions $F_i(x) = \int_0^x p_i(y)dy$.

The densities $p_A(x), p_B(x)$ of the wage (A) and non-wage income (B) have been studied in [29]. The density $p_A(x)$ was found to be well approximated by an exponential distribution in the small incomes region, and by a power law (Pareto distribution) with Pareto coefficient $\alpha = 2.53$ in the high incomes region. The non-wage income (B) has a smaller contribution to the total

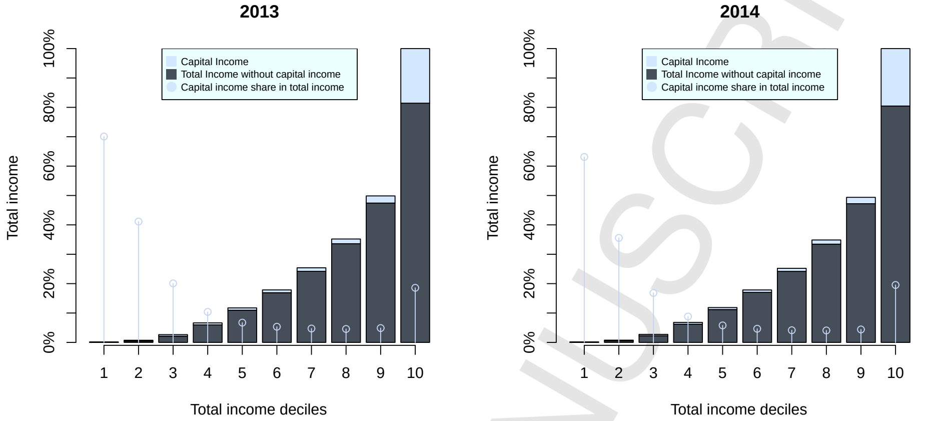

Figure 2: The relative contribution to the capital income to the total income (capital income shares), divided by deciles of the total income. The graphical representation follows that in Schlenker, Schmid [39].

income, such that the distribution of the total income (A+B) is dominated by that of the wage income (A).

In this paper we study the capital income distribution $p_C(x)$. The plots of the empirical distribution $p_C(x)$ of the capital income for 2013 and 2014 are presented in Fig. 3. We also show the logarithm of the complementary cumulative distribution function $\log(1 - F_C(x))$ vs $\log x$ in Figure 4.

While the plots of the densities of the capital income do not show much detail, the log-log plot of the complementary distribution function $\bar{F}_C(x) = 1 - F_C(x)$ shows the presence of two regions of low-middle and large incomes, with distinct qualitative behavior. In each region the plot is approximatively a straight line, and they are separated by a threshold at $x_T \simeq 120,000$ RON.

Visual inspection of these plots suggests thus that the long upper tail of the distribution $p_C(x)$ could have a Pareto form. The region of middle and low incomes could also be interpreted as a power law since the log-log plot shows an approximate straight line in this region, but this observation could be misleading since other distributions such as for example the log-normal distribution can show a straight line in a log-log plot of the complementary cumulative distribution function.

We will perform here a more careful analysis of the functional dependence, which will confirm the presence of a power law in both regions. In order to construct a parameterization for the entire capital income distribution we proceed with the analysis of the $\log x$ data series.

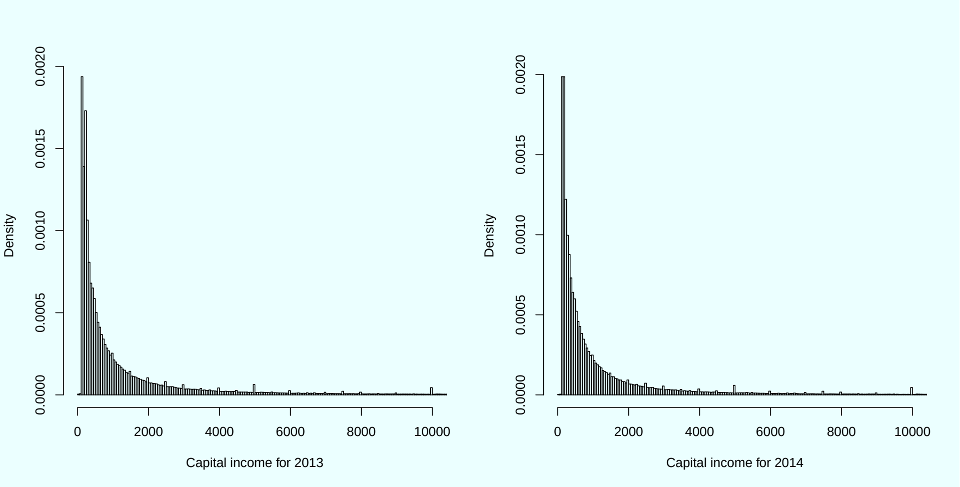

Figure 3: Plot of the empirical distribution of the capital income $p_C(x)$ for 2013 and 2014.

Figure 5 shows the density function of $\log x$ for 2013 and 2014. Both densities start to have values significantly greater then zero at $\log x = 4.6$ and they have a prominent peak at $\log x = 5.45$ for 2013 data and at $\log x = 5.05$ for 2014 data series. The 2013 data series, besides the peak at $\log x = 5.45$ have a second but smaller peak at $\log x = 4.95$. Both functions have a similar shape that can be decomposed into three distinct regions plus the above mentioned peaks.

We are thus led to adopt a piece-wise fit for the distribution of $y := \log x$, consisting of three regions:

- Low incomes region $x_1 \leq y \leq x_2$: the density $q_C(y)$ is approximated as a Gamma distribution.

- Middle incomes region $x_2 \leq y \leq x_3$: the density $q_C(y)$ is approximated as an exponential distribution.

- High incomes region $y \geq x_3$: the density $q_C(y)$ is approximated as a shifted exponential distribution.

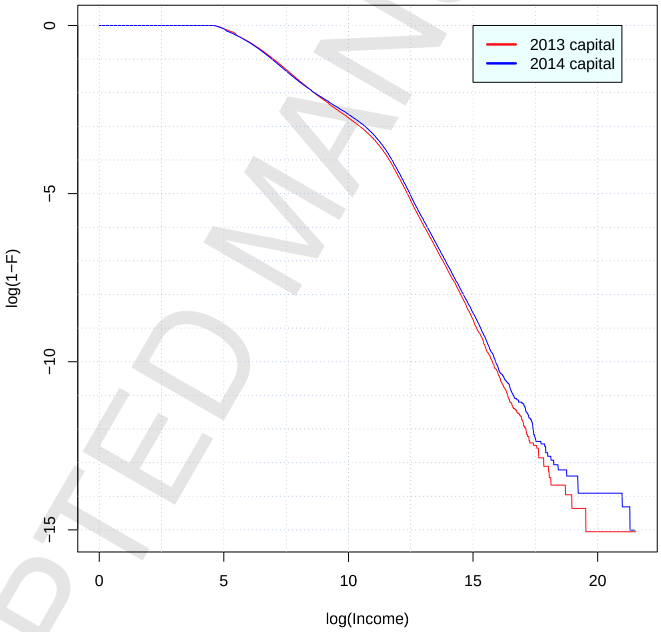

Figure 4: Log-log plot of the capital income distribution for 2013 and 2014.

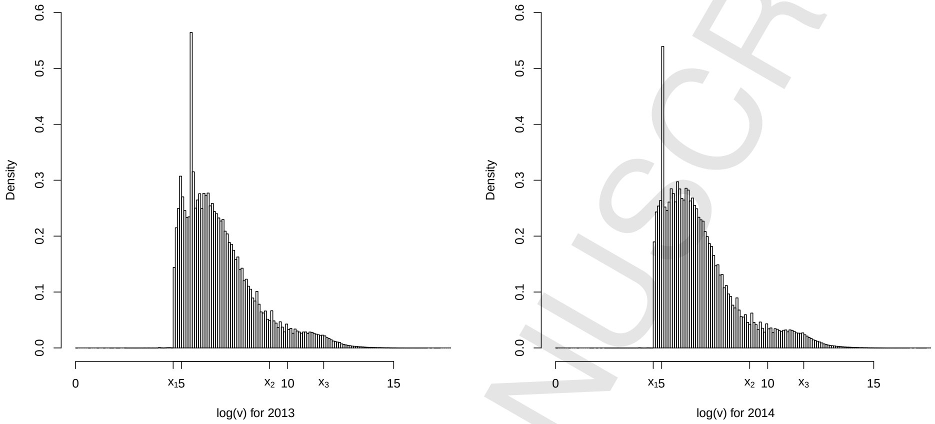

Figure 5: Density of $y = \log v$, the log of the capital income v distribution for 2013 and 2014.

In addition, the contributions of the peaks seen in Fig. 5 are modelled as step functions of constant height c. We note that $y = \log x > 0$ is positive definite, as the minimum capital income is $x = 1$.

This piece-wise representation has also an economic interpretation, as each region receives contributions from different types of capital income. The region of low capital incomes is dominated by interest income, dividends, stock transactions income, rents and lease income. These account for 86.7% of the total capital income in this region.

The middle- and large-incomes regions are dominated by income from real estate transactions. These are divided into two groups, of low- and high-value transactions, with distinct functional distribution. These account for 70% of the capital income in the middle-income region, and for about 67% of the income in the high-income region.

3.1 Parametric representation of the capital income

We will denote the distribution of $y = \log x$ as $q_C(y)$. This is related to the distribution function of the capital income $p_C(x)$ as $q_C(y) = xp_C(x) = e^y p_C(e^y)$.

The piece-wise parametric representation described above for the distribution $q_C(y)$ is written explicitly as

$$
\begin{aligned}
&(1) & y \le x_1 &: \; q_C(y) = 0 \\
&(2) & x_1 < y < xd_1 &: \; q_C(y) = \frac{b_1^{a_1}}{\Gamma(a_1)} \cdot y^{a_1-1} \cdot e^{-b_1 \cdot y} \\
&(3) & xd_1 \le y \le xd_2 &: \; q_C(y) = c_1 \\
&(4) & xd_2 < y < xd_3 &: \; q_C(y) = \frac{b_1^{a_1}}{\Gamma(a_1)} \cdot y^{a_1-1} \cdot e^{-b_1 \cdot y} \\
&(5) & xd_3 \le y \le xd_4 &: \; q_C(y) = c_2 \\
&(6) & xd_4 < y \le x_2 &: \; q_C(y) = \frac{b_1^{a_1}}{\Gamma(a_1)} \cdot y^{a_1-1} e^{-b_1 \cdot y} \\
&(7) & x_2 < y \le x_3 &: \; q_C(y) = \lambda_1 \cdot e^{-\lambda_1 \cdot y} \\
&(8) & y > x_3 &: \; q_C(y) = \lambda_2 \cdot e^{-\lambda_2 \cdot (y-a_2)}
\end{aligned}
$$

The two peaks in the low incomes region have been represented as step functions. Their boundaries for the 2013 data are $xd_1 = 4.8$ and $xd_2 = 5.1$ for the first peak, and $xd_3 = 5.4$ and $xd_4 = 5.5$ for the second and most prominent peak. The peak heights are $c_1 = 0.285$ and $c_2 = 0.565$. A similar representation is used for the 2014 data, except that there is only one peak of height $c_1 = 0.55$ in the region $[xd_1, xd_2]$, with $xd_1 = 5.0$ and $xd_2 = 5.1$.

The parameters of the probability density functions were estimated by the Maximum Likelihood Estimation (MLE) method. The boundaries x_2, x_3 were estimated using a goodness of fit approach, using an algorithm that is presented in the Appendix B, see Algorithm 1.

The numerical values of these parameters corresponding to the best fit are shown in Table 3. Figure 6 shows the empirical density function $q_C(y)$ and the result of the fit (blue curve) for the density of $y = \log x$ for 2013 and 2014. The fit quality is seen to be very good.

3.2 Discussion of the parametric distribution of the capital income

In the previous section we obtained a parametric description of the density $q_C(y)$ of $y := \log x$ with x the capital income. In this section we study the implication of the results for the distribution of the capital income $p_C(x)$. We also present the results of an alternative determination of the distribution in the upper tail of the capital income distribution.

Table 3: Parameter values for the best fit to the distribution of $\log x$, with x the capital income for 2013 and 2014. The values of the parameters for the second peak, appearing only in the distribution $q_C(y)$ for 2013, are $xd_3 = 5.4$, $xd_4 = 5.5$.

Parameter	2013	2014	Parameter	2013	2014
x_1	4.6	4.6	a_1	17.75081	19.6427
xd_1	4.8	5.0	b_1	2.78872	3.13746
xd_2	5.1	5.1	λ_1	0.15174	0.079479
x_2	9.45	9.15	λ_2	1.44091	1.43023
x_3	11.7	11.7	a_2	8.90577	9.02909
c_1	0.285	0.565	c_2	0.55	−

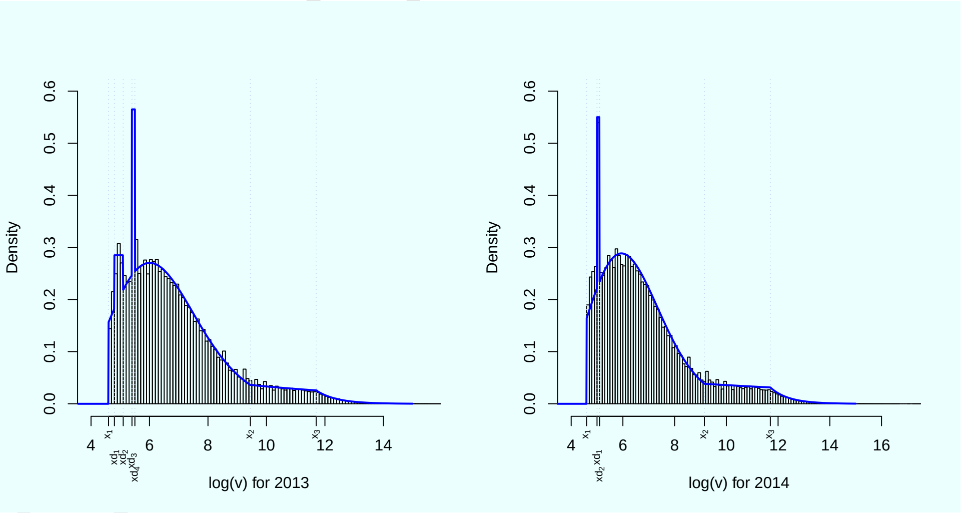

Figure 6: Plots of $q_C(y)$, the best fit to the distribution of $y = \log v$ with v the capital income.

3.2.1 Upper tail of the capital income

We start with the region of large capital incomes. This corresponds to the upper tail region $\log x > x_3$ which is described by a shifted exponential distribution with the rate λ_2. This translates into a power law distribution for $p_C(x)$, with the same exponent λ_2. The complementary cumulative distribution function is

$$\bar{F}_C(x) = \int_x^\infty \lambda_2 e^{-\lambda_2(y-a_2)} dy = x^{-\lambda_2} \cdot e^{\lambda_2 \cdot a_2} \,. \tag{9}$$

This gives a power law dependence in x in the upper tail of the capital income distribution, for both time periods under consideration. This dependence is typically denoted as a Pareto law with probability distribution function

$$f_{\text{Pareto}}(x) = cx^{-\alpha-1}, \qquad x > x_{\text{min}} \,, \tag{10}$$

where and c, x_{min} are constants and α is called the Pareto coefficient. This probability density function has the complementary cumulative distribution function

$$\bar{F}_{\text{Pareto}}(x) = \int_x^\infty f_{\text{Pareto}}(u) du = \left(\frac{x_{min}}{x}\right)^{-\alpha} \tag{11}$$

which has the same form as Eq. (9) with the identification $\alpha = \lambda_2$.

From Eq. (9) we obtain:

$$\log(1 - F_C(x)) = \lambda_2 \cdot a_2 - \lambda_2 \log x \,, \tag{12}$$

which corresponds to the linear part of the log-log plot shown in Figure 4 for the high income region.

Using the estimated values for λ_2 from Table 3 we get that the high income region of the capital income distribution is well described by a power law with the Pareto coefficients $\alpha_{2013} = 1.44$ for 2013 and $\alpha_{2014} = 1.43$ for 2014.

We checked the values of Pareto coefficients by computing them also using the procedure described in [9] and [46]. This is also based on maximum likelihood estimation. In the case of integer discrete data $\{v_i\}, i = 1, n$ which is the case of the income data, the approximate value of the Pareto coefficient α is given in Annex B.4 of [46]:

$$\hat{\alpha} \approx n \left[\sum_{i=1}^{n} \log \frac{x_i}{x_{\text{min}} - \frac{1}{2}} \right]^{-1} \,. \tag{13}$$

We used the poweRlaw package [19] to estimate the following parameters of the power law for the capital income in the higher incomes region: i) the Pareto coefficient α, ii) the threshold value $x_{\min}$ marking the lower boundary of the region for the Pareto law, and iii) the Kolmogorov-Smirnov statistics. The numerical results for these parameters are shown in Table 4. As mentioned, these parameters are related to those of the parametric distribution in the upper tail as $\alpha = \lambda_2$. Also one can identify $v_{\min} = e^{x_3}$. The values of the Pareto coefficients obtained by the method of [19] agree well with those determined by the MLE method described in Appendix B and listed in Table 3. Also the threshold values $x_{\min}$ where the Pareto distribution starts, agree well with $e^{x_3} = 120, 572$, where we used $x_3 = 11.7$ from Table 3.

We also estimated the uncertainty in determining the $v_{\min}$ and α parameters using the bootstrap procedure described in [9]. We used again the poweRlaw package [19] with $B = 2500$ bootstrap iterations and obtained the results presented in Figures 7 to 10. The goodness of fit for the Pareto region of the capital income was tested with the bootstrapping procedure described in [9]. The p-values obtained for our income data series are presented in Table 4, and show a very good fit of the data by the Pareto distribution.

Table 4: The parameters of the Pareto distribution for the capital income distribution in the upper tail region, for 2013 and 2014. D denotes the Kolmogorov-Smirnov statistics for the goodness of fit for each year.

Parameter	2013	2014
$x_{\min}$	138,000	132,000
α	2.435	2.428
D	0.006	0.005
p-value	0.93	0.92

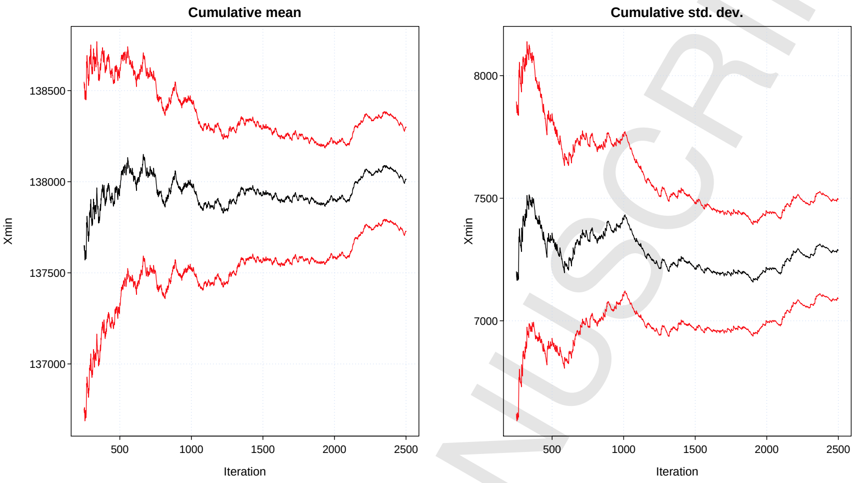

Figure 7: Uncertainty on $x_{\min}$ for 2013, high income region. Red lines denote the 95% confidence interval.

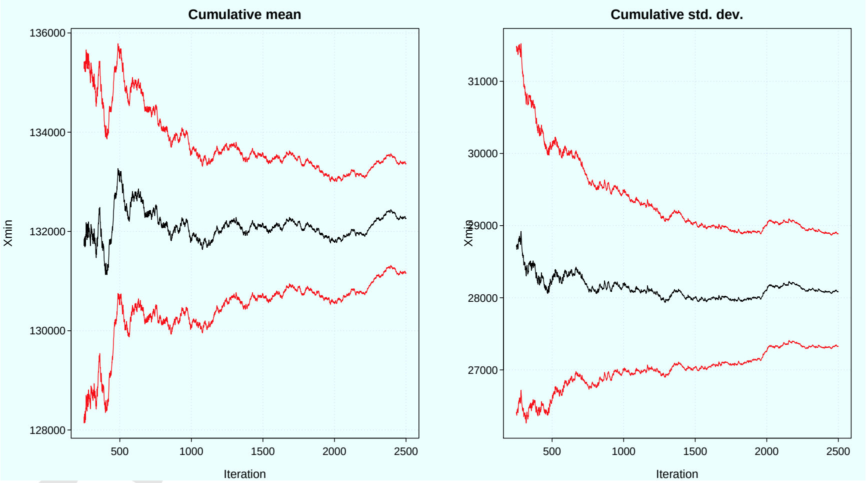

Figure 8: Uncertainty on $x_{\min}$ for 2014, high income region. Red lines denote the 95% confidence interval.

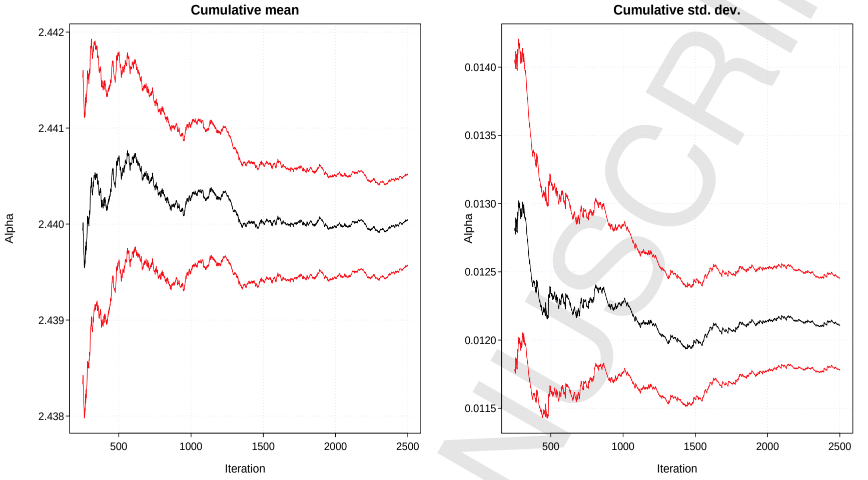

Figure 9: Uncertainty on α for 2013, high income region. Red lines denote the 95% confidence interval.

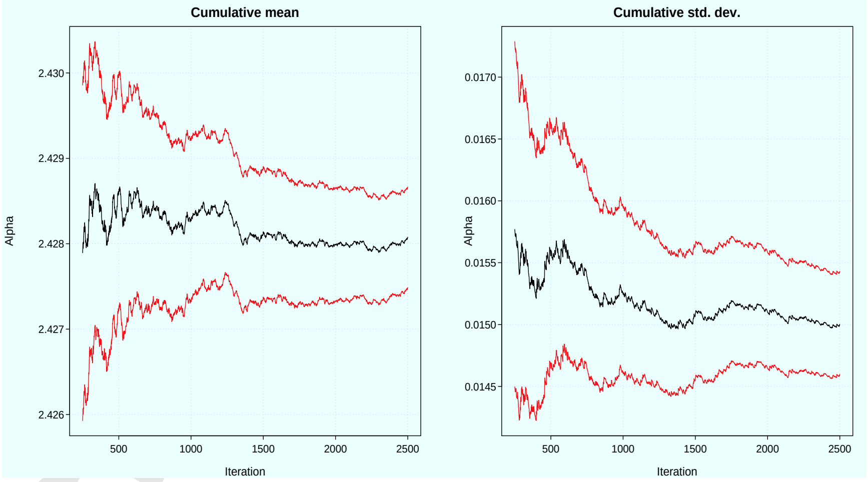

Figure 10: Uncertainty in computing α for 2014, high income region. Red lines denotes the 95% confidence interval.

In [29] the total wage (A) plus non-wage (B) income distribution for 2013 was found to be well represented for large incomes $x > 100,000$ RON by a similar Pareto distribution with exponent $\alpha = 2.53$ and scale parameter $x_{\text{min}} = 140,859$ RON. The Pareto exponent of the capital income distribution in the large incomes region $\alpha_{2013} = 1.44$ is appreciably smaller than that of the $A+B$ income distribution $\alpha = 2.53$ found in [29]. This explains partially why the capital income has a larger inequality. The magnitude of the $x_{\min}$ coefficient is also partly resposible for the difference in inequality.

We computed the length of the Pareto tail of the capital income distribution, defined as the ratio of the number of the capital income earners with income larger than $x_{\min}$ to the total number of capital income earners. This is 1.41% for 2013 and 1.73% for 2014. These numbers compare well with other studies [40] that showed a length of approximative 1% for the Pareto tail for Japan and [11] where a Pareto tail of 1% to 3% length is reported for U.K., U.S. and Germany.

We also determined that the dominant type of income for the Pareto tail comes from real estate transfers: approximatively 67% of the total income greater than the threshold shown in Table 4, while the other types of capital income are dominant in the lower and middle ranges of incomes.

3.2.2 Middle capital incomes region

We consider next the region $x_2 < \log x \leq x_3$ of middle capital income, shown as the almost flat region in Fig. 6. In this region the complementary cumulative distribution function of the capital income has the form

$$\begin{aligned} \bar{F}_C(x) &= e^{-\lambda_2 \cdot (x_3 - a_2)} + e^{-\lambda_1 \cdot \log x} - e^{-\lambda_1 \cdot x_3} = \\ &= e^{-\lambda_1 \cdot \log x} + A = x^{-\lambda_1} + A \end{aligned} \tag{14}$$

where we defined $A = e^{-\lambda_2 \cdot (x_3 - a_2)} - e^{-\lambda_1 \cdot x_3}$. In this region the capital income distribution has again Pareto form up to a constant term, with Pareto coefficient $\alpha = \lambda_1$.

3.2.3 Small capital incomes region

Finally, we consider the region $x_1 < \log x \leq x_2$ of small capital incomes. In this region the distribution of $\log x$ is well described by a Gamma distribution $\Gamma(a_1, b_1)$. In addition there is one peak (for 2014) and two peaks (for 2013), which are modeled as step functions.

The contribution to the capital income in the peaks observed in this region ($\log x \in [5.0, 5.1]$) comes mostly from interest income. For 2013 the interest income contribution is 74.4% and for 2014 it is 87.3%.

In this region the complementary cumulative distribution function of x is computed as follows. For simplicity of calculations we ignore the peak area that only add a constant value to the cumulative distribution function.

$$\begin{aligned} 1 - F_C(x) &= e^{-\lambda_2 \cdot (x_3 - a_2)} + e^{-\lambda_1 \cdot x_2} - e^{-\lambda_1 \cdot x_3} \\ &+ \frac{\Gamma(a_1, b_1 \cdot \log x)}{\Gamma(a_1)} - \frac{\Gamma(a_1, b_1 \cdot x_2)}{\Gamma(a_1)} \\ &= B + \frac{\Gamma(a_1, b_1 \cdot \log x)}{\Gamma(a_1)} \end{aligned} \tag{15}$$

where B is a constant given by $B = e^{-\lambda_2 \cdot (x_3 - a_2)} + e^{-\lambda_1 \cdot x_2} - e^{-\lambda_1 \cdot x_3} - \frac{\Gamma(a_1, b_1 \cdot x_2)}{\Gamma(a_1)}$ and $\Gamma(a_1, b_1 \cdot x_2)$ is the upper incomplete Gamma function.

Figure 11 shows again the log-log plot of the complementary cumulative distribution function for both years, together with the fitted lines given by Eqs. (12), (14) and (15) with the parameters from Table 3.

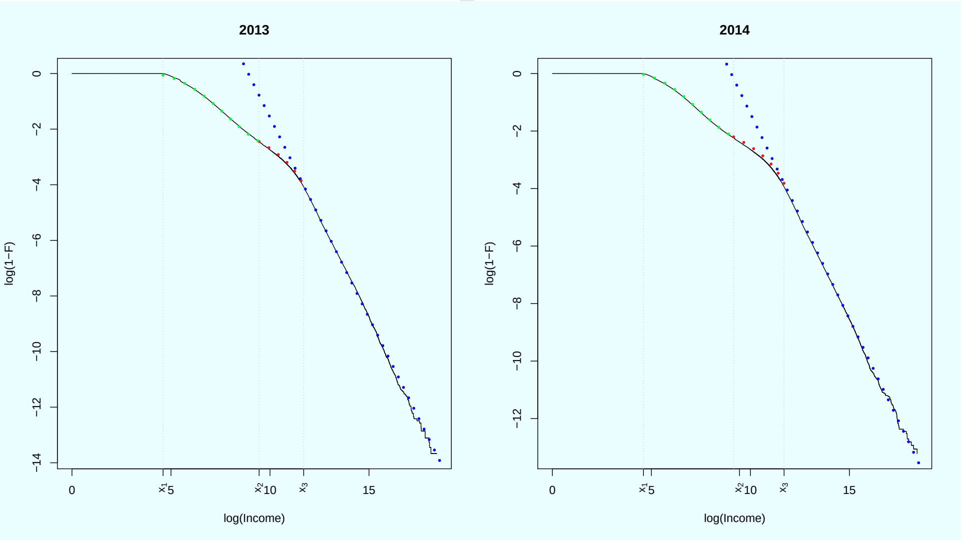

Figure 11: Plot of $\log \bar{F}_C(x)$, the complementary cumulative distribution function of the capital income x vs $\log x$, where the blue dotted line is given by Eq. (12), the red dotted line by Eq. (14) and the green dotted line by Eq. (15).

4 Conclusions

We present in this paper the first study of the capital income in Romania, using individual tax income data for 2013 and 2014. This completes a study of the income distribution performed in [29] which included only the contributions from wages and social redistribution income. We give results for the capital income shares, the inequality measures for the capital income, and its effect on the distribution and inequality of the total income. The details of the distribution of the capital income are studied using a parametric representation, which reduces in the high incomes region to a Pareto distribution.

Previous studies of the income distribution and inequality in Romania used either partial data, covering a single region [12], or used income survey data [27, 28, 38]. To our knowledge the present paper in the first such study taking into account capital income data.

Although income survey data are also available, tax records are known to be more reliable. This is especially true for capital incomes, since apart from rents (which are declared in a personal income statement), the other capital incomes are taxed at the source. For example, interest income is declared by the bank where the interest is paid, and real estate sales are declared by the public notary recording the transaction.

Our results show that the capital income is much more unequally distributed than the wages- and non-wage income. As a result, including its contribution increases the income inequality for the total income, as measured by a wide range of inequality measures, see Table 7. For example, the Gini index for capital income is much larger than that for the total income, in both years considered: while the Gini index for capital income is 0.92 for 2013 and 0.93 for 2014 it is only 0.53 for the total income. The capital income has a greater concentration than the labor income, the Herfindahl concentration index for the capital income being 0.0033 and 0.0048 for 2013 and 2014 while for the total income it is 0.0001167 and 0.000184 which agrees with other studies in this area [26].

We present results for the capital income shares, which shows the breakdown of the total income into wage- and capital-income. The results agree with a general pattern of income shares distribution observed in other EU countries, in a study covering the period 2005-2011 [39].

The distribution of the capital income has a complex structure, and our study reveals the presence of three distinct regions with qualitatively different behavior. This reflects the variety of the different types of capital income. The interest income, dividends, rent income and lease income dominate the distribution in the low-incomes region. Income from real estate transactions dominates the distribution of the capital income in the middle- and high-income region. In each of these regions the distribution follows a Pareto law. The upper tail has a Pareto coefficient $\alpha \sim 1.4$ which is much smaller than the corresponding coefficient observed for the wage income $\alpha_A = 2.53$ [29], which agrees with a higher inequality for the capital income distribution noted from the inequality indices.

A Appendix: Breakdown by types of income

Each person can receive income of three types:

A: Wage income

B: Non-wage income

C: Capital income

We list in Table 5 the details of the income sources of each type.

We also list in Table 7 the values of the inequality measures for the various contributions to the total income, and for the total income.

B Estimation of the capital income distribution

We describe in this Appendix the details of the Maximum Likelihood Estimation method used for determining the parameters entering the functional representation of the capital income distribution introduced in Sec. 3.1. The procedure used is shown in symbolic form in Algorithm 1.

The distance D between the probability distributions is measured through the Kolmogorov-Smirnov statistics which is given by:

$$D = \max_x \mid S(x) - P(x) \mid \ . \tag{16}$$

where $S(x)$ is the cumulative distribution function of the empirical data and $P(x)$ is the cumulative distribution function of the fitted distribution.

Table 5: The breakdown of the different income sources used for this study.

Type A: Wages	Type B: Non-wages
Wage income from domestic labor	Income from agricultural labor
Wage income from abroad	Free-lance and commercial activities income
	Intellectual property rights income
	Gambling income
	Pensions
	Unemployment benefits, social benefits
	Commercial mandate income
	Commission income
Type C: Capital income	
Stock and bonds transfers	
Rental income	
Leasing income	
Interest income	
Dividends	
Income from foreign currency transactions	
Real-estate transfers	

The values of $x_{2\,\inf}$, $x_{3\,\inf}$, $x_{2\,\sup}$, $x_{3\,\sup}$, and s for both 2013 and 2014 used in Algorithm 1 are given in Table B.

Step 3 of the Algorithm 1 involves solving a constrained minimization problem. The function that has to be minimized is the log-likelihood of the distribution described by Eqs. (1)-(8). This function is given below in explicit form for 2014 (the form of the log-likelihood function for 2013 is similar with the exception that it adds the contribution from the rectangular area of the

Table 6: Tabulation of the Lorenz curve $L(p)$ of the total and capital income, 2013 and 2014.

p	2013		2014	
	Total income	Capital income	Total income	Capital income
0.05	0.000409	0.0005002	0.0003734	0.0004070
0.10	0.001308	0.0011142	0.0013053	0.0009067
0.15	0.003309	0.0018528	0.0034143	0.0014841
0.20	0.007287	0.0027614	0.0076538	0.0021475
0.25	0.014535	0.0038020	0.0152053	0.0029511
0.30	0.026435	0.0049754	0.0273001	0.0039153
0.35	0.043980	0.0063881	0.0449886	0.0050683
0.40	0.066281	0.0080978	0.0675307	0.0064465
0.45	0.090793	0.0101521	0.0924016	0.0081017
0.50	0.117288	0.0126287	0.1186232	0.0100823
0.55	0.146337	0.0156423	0.1471196	0.0124731
0.60	0.178366	0.0193517	0.1784502	0.0153874
0.65	0.213965	0.0239647	0.2131903	0.0189912
0.70	0.253826	0.0298354	0.2521299	0.0235474
0.75	0.299136	0.0375472	0.2965608	0.0295122
0.80	0.351880	0.0481822	0.3485781	0.0378071
0.85	0.416018	0.0642589	0.4118694	0.0507043
0.90	0.498558	0.0931939	0.4936895	0.0760989
0.95	0.615540	0.1725850	0.6095569	0.1560156
0.96	0.647045	0.2086092	0.6406299	0.1923755
0.97	0.684159	0.2600757	0.6771740	0.2426601
0.98	0.730725	0.3353214	0.7226878	0.3140789
0.99	0.795239	0.4534462	0.7858049	0.4234765
0.999	0.910667	0.7104452	0.8994391	0.6612755

Table 7: Inequality measures for the components of the total income, and for the total income. We define them as in Sec. 2 of [29].

2013	Total	Wages (A)	Capital (C)	Non-wage (B)
Gini	0.62	0.53	0.92	0.55
Ricci-Schutz	0.45	0.39	0.81	0.37
Atkinson($p = 0.5$)	0.37	0.25	0.78	0.31
Theil	1.07	0.56	3.47	0.81
2014	Total	Wages (A)	Capital (C)	Non-wage (B)
Gini	0.62	0.53	0.93	0.56
Ricci-Schutz	0.45	0.39	0.83	0.37
Atkinson($p = 0.5$)	0.36	0.24	0.81	0.32
Theil	1.18	0.55	3.96	0.82

Table 8: Search ranges for x_2 and x_3, and the step s used to search the best fit values.

2013			2014		
x_{2_inf}	x_{2_sup}	s	x_{3_inf}	x_{3_sup}	s
8.5	10	0.05	8	9.5	0.05
x_{3_inf}	x_{3_sup}	s	x_{3_inf}	x_{3_sup}	s
11	12	0.05	11	12	0.05

Algorithm 1 The algorithm for estimating the parameters x_2 and x_3 for the distribution of $\log v$.

```
for $x_2 = x_{2\_inf}$ to $x_{2\_sup}$ step $s$ do
  for $x_3 = x_{3\_inf}$ to $x_{3\_sup}$ step $s$ do
    Compute the parameters $a_1$, $b_1$, $\lambda_1$, $a_2$, and $\lambda_2$ solving a constrained optimization problem for the log likelihood function
    Compute $D$, the Kolomogov-Smirnov distance between the fitted distribution and empirical distribution and save it to a list $ks$ together with the values of the parameters
  end for
end for
Iterate over $ks$ and find the minimum value of $D$
Save the index of this value $i$
From $ks(i)$ extract the values of the parameters of the distribution $a_1$, $b1$, $\lambda_1$, $a_2$, and $\lambda_2$
Extract $x_2$ and $x_3$ that give the minimum $D$
```

first peak).

$$
\begin{aligned}
-ll(y) = {} & n_1 \cdot \log(\Gamma(a_1)) - n_1 \cdot a_1 \cdot \log(b_1) - (a_1 - 1) \cdot \sum_{x_1 < y_i < xd_1} \log(y_i) \\
& + b_1 \cdot \sum_{x_1 < y_i < xd_1} (y_i) - c_2 \cdot n_2 + n_3 \cdot \log(\Gamma(a_1)) - n_3 \cdot a_1 \cdot \log(b_1) - \\
& (a_1 - 1) \cdot \sum_{xd_2 < y_i \le x_2} \log(y_i) + b_1 \cdot \sum_{xd_2 < y_i \le x_2} (y_i) + \lambda_1 \cdot \sum_{x_2 < y_i \le x_3} (y_i) - n_4 \cdot \log(\lambda_1) + \\
& \lambda_2 \cdot \sum_{y_i > x_3} (y_i - a_2) - n_5 \cdot \log(\lambda_2)
\end{aligned} \tag{17}
$$

In the following we will refer to the analysis of the 2014 data, the computations for 2013 being completely similar. In Eq. (17), n_1 denotes the number of data points with $x_1 < y_i < xd_1$, n_2 the number of data points with $xd_1 \le y_i \le xd_2$, n_3 the number of data points with $xd_2 \le y_i \le x_2$, n_4 the number of data points with $x_2 < y_i \le x_3$, and n_5 the number of data points with $x_3 < y_i$.

We imposed three equality constraints on the minimization of the log-likelihood function. These constraints ensure that the density $q_C(y)$ of $y =$

$\log v$ has no jumps at x_2 and x_3, and that it is properly normalized to 1. The later restriction can be written as

$$\begin{aligned}(18)\quad \textstyle\int_0^\infty q_C(y)dy &= \int_{x_1}^{xd_1} \frac{b_1^{a_1}}{\Gamma(a_1)} \cdot y^{a_1-1} \cdot e^{-b_1 \cdot y} dy + c_2 \cdot (xd_2 - xd_1) + \\ &\int_{xd_2}^{x_2} \frac{b_1^{a_1}}{\Gamma(a_1)} \cdot y^{a_1-1} \cdot e^{-b_1 \cdot y} dy + \\ &\int_{x_2}^{x_3} \lambda_1 \cdot e^{-\lambda_1 \cdot y} dy + \int_{x_3}^{\infty} \lambda_2 \cdot e^{-\lambda_2 \cdot (y-a_2)} = \\ &\frac{\gamma(a_1, b_1 \cdot xd_1)}{\Gamma(a_1)} - \frac{\gamma(a_1, b_1 \cdot x_1)}{\Gamma(a_1)} + c_2 \cdot (xd_2 - xd_1) + \\ &\frac{\gamma(a_1, b_1 \cdot x_2)}{\Gamma(a_1)} - \frac{\gamma(a_1, b_1 \cdot xd_2)}{\Gamma(a_1)} + \\ &e^{-\lambda_1 \cdot x_2} - e^{-\lambda_1 \cdot x_3} + e^{-\lambda_2 \cdot (x_3 - a_2)} = 1\end{aligned}$$

Here $\gamma(a,x) = \int_0^x t^{(a-1)} e^{-t} dt$ is the lower incomplete Γ function.

Considering the large size of our data set, in order to speed up the solution of the constrained optimization problem, we added a set of inequality constraints $\mathcal{C}$ on the acceptable values of the parameters, see (20). Using these constraints, the problem can be stated as follows

$$\begin{aligned}(19)\qquad & \min \quad \{-ll(y)\} \\ & s.t. \quad \frac{b_1^{a_1}}{\Gamma(a_1)} \cdot x_2^{a_1-1} e^{-b_1 \cdot x_2} = \lambda_1 \cdot e^{-\lambda_1 \cdot x_2}\,, \\ & \qquad \lambda_1 \cdot e^{-\lambda_1 \cdot x_3} = \lambda_2 \cdot e^{-\lambda_2 \cdot (x_3 - a_2)}\,, \\ & \qquad \frac{\gamma(a_1, b_1 xd_1)}{\Gamma(a_1)} - \frac{\gamma(a_1, b_1 x_1)}{\Gamma(a_1)} + c \cdot (xd_2 - xd_1) + \\ & \qquad\quad \frac{\gamma(a_1, b_1 x_2)}{\Gamma(a_1)} - \frac{\gamma(a_1, b_1 xd_2)}{\Gamma(a_1)} + \\ & \qquad\quad e^{-\lambda_1 \cdot x_2} - e^{-\lambda_1 \cdot x_3} + e^{-\lambda_2 \cdot (x_3 - a_2)} - 1 = 0\,, \\ (20)\qquad & \mathcal{C}: \qquad \{\lambda_1 > 0, \lambda_2 > 0, 1 < a_1 < 45, 1 < a_2 < 15\}\,.\end{aligned}$$

The nonlinear optimization problem with equality and inequality constraints defined by (19) was solved for both data sets using the augmented Langrangian method. The general form of the optimization problem can be written as:

$$\begin{aligned} &\min \quad f(x)\,, \\ &s.t. \quad g(x) = 0 \\ &\qquad\quad h(x) \geq 0 \end{aligned} \tag{21}$$

Using slack variables the inequalities constraints can be easily transformed into equality constraints:

$$\begin{aligned} &\min \quad f(x)\,, \\ &s.t. \quad g(x) = 0 \\ &\qquad\quad h(x) - s = 0 \\ &\qquad\quad s \geq 0. \end{aligned} \tag{22}$$

The augmented Lagrangian of the problem Eq. (22) is given by:

(23)

$$L(x, s; \lambda_1; \lambda_2; \sigma_1; \sigma_2) = f(x) - \lambda_1 \cdot g(x) - \lambda_2 \cdot (h(x) - s) + \frac{1}{2}\sigma_1 \|g(x)\|^2 + \frac{1}{2}\sigma_2 \|h(x) - s\|^2\,,$$

where $\|f(x)\|^2$ is the L^2 norm.

The constraint optimization problem (21) is now transformed into a series of unconstrained optimization problems of the form (23). A sketch of the algorithm solving the optimization problem using the augmented Lagrangian method is shown in Algorithm 2.

Algorithm 2 Augmented Lagrangian method

```
1: Choose initial values for $\lambda_1$, $\lambda_2$, $\sigma_1$, $\sigma_2$
2: repeat
3:     Compute $x_{\lambda_1,\lambda_2,\sigma_1,\sigma_2}$ given by $\text{argmin}_x\, L(x, \lambda_1, \lambda_2, \sigma_1, \sigma_2)$
4:     Update $\lambda_1$, $\lambda_2$, $\sigma_1$, $\sigma_2$
5: until Stopping criteria are satisfied
```

A complete demonstration and a description of the augmented Lagrangian method for nonlinear optimization problems with equality and inequality constraints is beyond the purpose of this paper but an interested reader could find such details in [25].

We solved the optimization problem described by Eq. (19) using the alabama R package [45], step 3 from algorithm 1 being a call of alabama::auglag function which returns the optimal solution. Given the complexity of the algorithm, and the very large size of our data sets, we implemented the `for` loops in the algorithm 1 using the `foreach` package [36] to run each estimation in parallel, and the `doSNOW` package [37] as a parallel back-end.

The numerical values of these parameters corresponding to the best fit are shown in Table 3.

References

[1] D.J. Aigner and A.S. Goldberger, Estimation of Pareto's Law from grouped observations. Journal of the American Statistical Association 65, 712-723 (1970)

[2] A.B. Atkinson, T. Piketty, E. Saez, Top incomes in the long run of history, Journal of Economic Literature 49 (1), 2-71 (2011).

[3] A.B. Atkinson, Pareto and the Upper tail of the income distribution in the UK: 1799 to the present, Economica, vol. 84 issue 334, 129-156 (2017)

[4] Robert L. Axtell, Zipf Distribution of US Firm Sizes, Science 293, 1818-1820 (2001).

[5] E. Bengtsson, D. Waldenstrom, Capital Shares and Income inequality: Evidence from the Long Run, Discussion Paper No. 9581, December 2015.

[6] J. Benhabib and A. Bisin, Skewed wealth distributions: theory and empirics. NBER Working Paper no. 21924, 2016.

[7] R. Cerqueti, M. Ausloos, Evidence of economic regularities and disparities of Italian regions from aggregated tax income size data, Physica A 421, 187-207 (2014).

[8] R. Cerqueti, M. Ausloos, Assessing the inequalities of wealth in regions: the Italian case, Qual. Quant. 49, 2307-2323 (2015).

[9] A. Clauset, C.R. Shalizi, and M.E.J. Newman, Power-law distributions in empirical data, SIAM Review 51(4), 661-703 (2009).

[10] F. Clementi and M. Gallegati, Power Law Tails in the Italian Personal Income Distribution. Physica A 350:427-438 (2005).

[11] F. Clementi and M. Gallegati, Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States, in *Econophysics of Wealth Distributions*, Editors: A. Chatterjee, S. Yarlagadda and B.K. Chakrabarti, Springer, 2005, pp. 3-14.

[12] D. Derzsy, Z. Neda and M. Santos, Income distribution patterns from a complete social security database, Physica A 391, 5611-5619 (2012).

[13] A.A. Dragulescu and V.M. Yakovenko, Exponential and Power-Law Probability Distributions of Wealth and Income in the United Kingdom and the United States. Physica A 299:213-221 (2001).

[14] C. Efthimiou, A. Wearne, Household income distribution in the USA, Eur. Phys. J. B 17, 82 (2016).

[15] A. Frassdorf, M.M. Grabka and J. Schwarze, The impact of household capital income on income inequality - a factor decomposition analysis for the UK, Germany and the USA, The Journal of Economic Inequality, 2011, 9:35.

[16] X. Gabaix, Power laws in economics and finance. Annual Review of Economics 1, 25-93 (2009).

[17] X. Gabaix, Power Laws in Economics: An Introduction, Journal of Economic Perspectives, Volume 30, Number 1, Winter 2016, 185-206.

[18] K. Giesen, A. Zimmerman and J. Suedekum, The size distribution across all cities - double Pareto lognormal strikes. Journal of Urban Economics 68, 129-137 (2010).

[19] C.S. Gillespie, Fitting Heavy Tailed Distributions: The `poweRlaw` Package Journal of Statistical Software, 64(2), 1-16 (2015)

[20] P. Gopikrishnan, M. Meyer, L.A.N. Amaral, and H.E. Stanley, Inverse Cubic Law for the Probability Distribution of Stock Price Variations, Eur. Phys. J. B: Rapid Communications 3, 139-140 (1998).

[21] Gholamreza Hajargasht, William E. Griffiths, Pareto-Lognormal distributions: Inequality, poverty, and estimation from grouped income data, Economic Modelling 33, 593-604 (2013).

[22] M. Jagielski, K. Czyzewski, R. Kutner and H.E. Stanley, Income and wealth distribution of the richest Norwegian individuals: An inequality analysis Physica A: Statistical Mechanics and its Applications 474, 330-333

[23] S.P. Jenkins, Pareto models, top incomes, and recent trends in UK income inequality. ISER Working Paper 2016-07, University of Essex

[24] Charles I. Jones, Pareto and Piketty: The Macroeconomics of Top Income and Wealth Inequality, Journal of Economic Perspectives 29(1), 29-46 (2015)

[25] K. Madsen, H.B. Nielsen and O. Tingleff, Optimization With Constraints, 2nd Edition, March 2004, IMM, Technical University of Denmark

[26] Branko Milanovic, Increasing Capital Income Share and Its Effect on Personal Income Inequality, MPRA Paper No. 67661, 2015.

[27] M. Molnar, Income polarization in Romania, Rom. J. Forecast. 14(2), 64-83 (2011).

[28] M. Molnar, Income distribution in Romania, Technical Report of the Institute of National Economy of the Romanian Academy, 2010.

[29] B. Oancea, T. Andrei and D. Pirjol, Income distribution in Romania: the exponential-Pareto distribution, Physica A 469, 486-498 (2017).

[30] V. Pareto, La courbe de la repartition de la richesse. Recueil publie par la Faculte de Droit a la occasion de l'Exposition Nationale Suisse. 1896, Lausanne: Ch. Viret-Genton.

[31] V. Pareto, Cours d'Economie Politique, Librairie Droz, 1896.

[32] T. Piketty, Capital in the Twenty-First Century, Cambridge, MA: Harvard University Press, 2014

[33] W.J. Reed, The Pareto, Zipf and other power laws. Economics Letters 74, 15-19 (2001).

[34] W.J. Reed, The Pareto law of incomes - an explanation and an extension. Physica A 319, 469-486 (2003).

[35] W.J. Reed, Murray Joergensen, The Double Pareto-Lognormal Distribution - A New Parametric Model for Size Distributions, Communications in Statistics - Theory and Methods 33 (2004).

[36] *Revolution Analytics* and Steve Weston, `foreach` Package: Provides Foreach Looping Construct for R, R package version 1.4.3, https://CRAN.R-project.org/package foreach (2015)

[37] *Revolution Analytics* and Stephen Weston, `doSNOW` Package: Foreach Parallel Adaptor for the `snow` Package, R package version 1.0.14, https://CRAN.R-project.org/package=doSNOW (2015)

[38] S. Rose and C. Viju, Income inequality in post-communist Central and Eastern European countries, East J. Eur. Studies 5(1), 5 (2014).

[39] E. Schlenker and K.D. Schmid, Capital income shares and income inequality in the European Union, Working Paper ECINEQ WP 2014 329, Society for the Study of Economic Inequality, 2014.

[40] W. Souma, Universal structure of the personal income distribution, Fractals 9(4), 463-470 (2001)

[41] M.H.R. Stanley, S.V. Buldyrev, S. Havlin, R. Mantegna, M.A. Salinger, and H.E. Stanley, Zipf plots and the size distribution of firms, Economics Lett. 49, 453-457 (1995).

[42] H.E. Stanley, L. A.N. Amaral, P. Gopikrishnan, and V. Plerou, Scale Invariance and Universality of Economic Fluctuations, Physica A 283, 31-41 (2000).

[43] A.A. Toda, Income dynamics with a stationary double Pareto distribution. Phys. Rev. E83, 046122 (2011).

[44] A.A. Toda, The double power law in income distribution: Explanations and evidence, Journal of Economic Behavior and Organization 84, 364-381 (2012).

[45] Ravi Varadhan, alabama: Constrained Nonlinear Optimization R package version 2015.3-1, 2015, https://CRAN.R-project.org/package=alabama

[46] Y. Virkar and A. Clauset, Power-law distributions in binned empirical data. Annals of Applied Statistics 8(1), 89 - 119 (2014).