

PARIS SCHOOL OF ECONOMICS

MASTER ANALYSE ET POLITIQUE ÉCONOMIQUES

MASTER'S THESIS

---

**Generalized Pareto curves:  
Theory and application using income and inheritance  
tabulations for France 1901-2012**

---

Juliette FOURNIER

September 2015

*Supervisor:*

Thomas PIKETTY

*Referee:*

Facundo ALVAREDO

**JEL codes:** C14; C46; D31.

**Keywords:** Income distribution, Wealth distribution, Pareto law, Generalized Pareto curves, Copulas.

## Abstract

We develop a new nonparametric method to estimate shares of income and wealth accruing to the different deciles and percentiles of the distribution. Whereas methods usually employed in the literature are parametric, inasmuch as they are typically based on the assumption that the top of the income distribution follows a Pareto distribution, we are able to relax any assumption on the shape of the distribution. Namely, we evaluate non-parametrically the distribution after estimating the empirical "generalized Pareto curve". It is defined as the curve of inverted Pareto coefficients  $b(p)$ , where  $p$  is the percentile rank and  $b(p)$  is the ratio between the average income above percentile  $p$  and the income threshold at percentile  $p$  (i.e.  $b(p) = \mathbb{E}[y|F(y) \geq p]/F^{-1}(p)$ ).

We exploit income tax tabulations from 1915 to 2012 in France to generate new series that we can compare to existing WTID series for top shares. We find that old and new series are almost equal throughout the period. This confirms that the Paretian form fits indeed relatively well the top of the distribution. However, the Pareto hypothesis is only valid locally, whereas the method elaborated here allows derivation of estimates for the whole income distribution. In particular, we provide computer codes that can be used to simulate reliable synthetic micro-files from income tabulations. Another potential application is the homogenization of series obtained for individual-based tax systems and for household-based tax systems, and for the income concept. Finally, we provide a preliminary application to French 1901-2000 inheritance tabulations and the distribution of wealth.

## **Acknowledgements**

I would like to thank Thomas Piketty for his insightful comments and suggestions, his support, and the time he devoted to supervise this work.

I am also grateful to Facundo Alvaredo for having accepted to be the referee for this master's thesis.

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
<b>2</b>	<b>Existing literature on the Pareto law in income and wealth distributions</b>	<b>11</b>
2.1	Previous attempts to estimate and generalize Pareto distributions . . . . .	11
2.1.1	The Pareto distribution . . . . .	11
2.1.2	Common estimation methods . . . . .	13
2.1.3	Usual representations of the income distribution . . . . .	19
2.2	Theoretical models yielding Pareto distributions . . . . .	27
2.2.1	Earnings distribution . . . . .	27
2.2.2	Accumulation models for wealth distribution . . . . .	32
<b>3</b>	<b>Generalized Pareto curves: theory and evidence</b>	<b>40</b>
3.1	Theory . . . . .	40
3.1.1	The income distribution . . . . .	40
3.1.2	Pareto curve and quantile function . . . . .	41
3.1.3	Lorenz curve . . . . .	44
3.2	Evidence using micro-files for France 2006 . . . . .	46
3.2.1	The Pareto curve . . . . .	46
3.2.2	Asymptotic decline of the Pareto curve for finite populations . . . . .	51
3.2.3	Estimations using tabulations of the income tax . . . . .	52
<b>4</b>	<b>Application to French income and inheritance tax tabulations 1901-2012</b>	<b>61</b>
4.1	Application to income tabulations for France 1915-2012 . . . . .	61
4.1.1	The income tax in France . . . . .	61
4.1.2	Corrections . . . . .	69
4.1.3	Estimations . . . . .	75
4.2	Application to inheritance tax tabulations 1902-1994 . . . . .	79
4.2.1	The inheritance tax in France . . . . .	79
4.2.2	Estimations . . . . .	81
<b>5</b>	<b>Conclusion</b>	<b>85</b>
<b>A</b>	<b>Pareto curves of usual parametric distributions</b>	<b>87</b>
<b>B</b>	<b>Estimating the generalized Pareto curve</b>	<b>94</b>
B.1	A first try: approximation by a suited functional form . . . . .	94
B.2	Shape-preserving interpolation . . . . .	96
B.2.1	Review of basic interpolation methods . . . . .	97
B.2.2	Piecewise cubic Hermite polynomial interpolation . . . . .	102
B.3	Extrapolation . . . . .	103

B.3.1	Lower incomes . . . . .	103
B.3.2	Top of the distribution . . . . .	105
<b>C</b>	<b>Simulation of synthetic micro-files</b>	<b>106</b>
C.1	Simulation of a population using tax tabulations . . . . .	106
C.1.1	The inversion method . . . . .	106
C.1.2	Matlab code . . . . .	107
C.2	Comparison of the results with microdata . . . . .	115
<b>D</b>	<b>From households to individuals: correcting for the variations in tax units</b>	<b>118</b>
D.1	Homogenization of series across countries: the problem of the changes in tax units	118
D.1.1	The problem . . . . .	118
D.1.2	Method to correct for changes in tax units . . . . .	118
D.2	Matlab code . . . . .	119
D.2.1	Description of the code . . . . .	119
D.2.2	Code . . . . .	120
D.2.3	Evidence with microdata of France 2006 . . . . .	121
<b>E</b>	<b>New series</b>	<b>123</b>
E.1	Income distribution . . . . .	123
E.1.1	Taxable income . . . . .	123
E.1.2	Fiscal income . . . . .	127
E.1.3	Estimations for the years 2001-2012 . . . . .	133
E.2	Inheritance distribution . . . . .	134
<b>F</b>	<b>Empirical Pareto curves of the income and inheritance distributions</b>	<b>137</b>
F.1	Pareto curves of the income distribution 1915-2012 . . . . .	137
F.2	Pareto curves of the inheritance distribution 1902-1994 . . . . .	138

# List of Figures

2.1	Density of the Pareto distribution . . . . .	12
2.2	The normal distribution . . . . .	19
2.3	The lognormal distribution . . . . .	20
3.1	Frequency distribution of incomes, France 2006 . . . . .	46
3.2	Pareto curve, France 2006 . . . . .	47
3.3	Pareto curve, France 2006 - Zoom . . . . .	48
3.4	Interpolation of the Pareto curve, France 2006 . . . . .	50
3.5	Zoom on the top 0.1 percent of the Pareto curve, France 2006 . . . . .	51
3.6	Final drop - Comparison with simulation . . . . .	52
3.7	Approximation of the Lorenz curve, France 2006 . . . . .	56
3.8	Ratios of estimated values of true values for different deciles and percentiles of the population . . . . .	60
4.1	Pareto curve of the income distribution, France 1981 . . . . .	75
4.2	Comparison of new estimations of taxable income with estimations of Piketty [2001] . . . . .	77
4.3	Evolution of the top shares of income accruing to different percentiles of the population . . . . .	78
4.4	Pareto curve of the inheritance distribution, France 1943 . . . . .	81
4.5	Comparison of new estimations of the inheritance distribution with estimations of Piketty [2001] . . . . .	83
4.6	Evolution of the top shares of inheritance accruing to different percentiles of the population . . . . .	84
A.1	Pareto curves of Pareto type I and type II distributions . . . . .	88
A.2	Pareto curves of type III and type IV Pareto distributions . . . . .	89
A.3	Pareto curves of lognormal distributions . . . . .	90
A.4	Pareto curves of Champernowne and Sech <sup>2</sup> distributions . . . . .	91
A.5	Pareto curves of Gamma distributions . . . . .	92
A.6	Pareto curves of Weibull and Singh-Maddala distributions . . . . .	93
B.1	Approximation of the generalized Pareto curve, France 2006 . . . . .	95
B.2	Approximation of the generalized Pareto curve, France 2006 . . . . .	96
B.3	Linear interpolation of the generalized Pareto curve . . . . .	98
B.4	Cubic spline interpolation of the generalized Pareto curve . . . . .	98
B.5	PCHIP interpolation of the Pareto curve . . . . .	100
B.6	PCHIP interpolation of the Pareto curve - Zoom on top percentiles . . . . .	101
B.7	Extrapolation of the lower part of the Pareto curve, France 2006 . . . . .	105
C.1	Comparison of simulated population and microdata . . . . .	116
C.2	Comparison of simulated population and microdata - Lorenz curve . . . . .	117

# List of Tables

3.1	Income tax tabulation, France 2006 . . . . .	53
3.2	Tabulation corresponding to microdata - Tax scale: France 2006 . . . . .	53
3.3	Tabulation corresponding to microdata - Tax scale: France 2012 . . . . .	54
3.4	Thresholds corresponding to different deciles and percentiles of the population . .	57
3.5	Average income above different deciles and percentiles of the population . . . . .	57
3.6	Values taken by the Lorenz curve at different deciles and percentiles of the population	57
4.1	Share of taxable households and number of tax brackets by year, France 1915-1944	64
4.2	Share of taxable households and number of tax brackets by year, France 1945-1998	67
4.3	Deductibility of the IGR: corrective rates for incomes of the years 1916-1947 . . .	72
4.4	Deductibility of the IGR and schedular taxes: global corrective rates for incomes of the years 1916-1970 . . . . .	74
4.5	Category abatements: corrective rates for incomes of the years 1915-1998 . . . . .	75
4.6	Share of positive legacies and number of thresholds in the inheritance tax tabula- tions by year . . . . .	80
C.1	Input of the Matlab program - Worksheet "Tabulation" . . . . .	108
C.2	Input of the Matlab program - Worksheet "Average income" . . . . .	108
D.1	Comparison with microdata - Thresholds . . . . .	122
D.2	Comparison with microdata - Average income . . . . .	122
E.1	Estimations of taxable income 1919-1944 - Thresholds . . . . .	123
E.2	Estimations of taxable income 1945-1998 - Thresholds . . . . .	124
E.3	Estimations of taxable income 1919-1944 - Average income . . . . .	125
E.4	Estimations of taxable income 1945-1998 - Average income . . . . .	126
E.5	Estimations of fiscal income 1919-1944 - Thresholds . . . . .	127
E.6	Estimations of fiscal income 1945-1998 - Thresholds . . . . .	128
E.7	Estimations of fiscal income 1919-1944 - Average income . . . . .	129
E.8	Estimations of fiscal income 1945-1998 - Average income . . . . .	130
E.9	Estimation of fiscal income 1919-1944 - Shares . . . . .	131
E.10	Estimation of fiscal income 1945-1998 - Shares . . . . .	132
E.11	Estimations of income distribution 2001-2012 - Threshold . . . . .	133
E.12	Estimations of income distribution 2001-2012 - Average income . . . . .	133
E.13	Estimations of income distribution 2001-2012 - Share . . . . .	133
E.14	New estimations of inheritance distribution - Threshold . . . . .	134
E.15	New estimations of inheritance distribution - Average income above . . . . .	135
E.16	New estimations of inheritance distribution - Share . . . . .	136

# Section 1

## Introduction

During the past fifteen years, the renewed interest for the long-run evolution of income and wealth distribution gave rise to a flourishing literature (see Piketty [2001], Atkinson and Piketty [2007, 2010], Piketty [2013]). To a large extent, this literature follows the pioneering work of Kuznets [1953] and Atkinson and Harrison [1978] and extends it to many more countries and years. In particular, a series of studies estimated the evolution of shares accruing to top income groups over the long-run for a range of more than twenty countries (see Atkinson et al. [2011] and Alvaredo et al. [2013] for recent surveys). This work was recently extended to study the long run evolution of wealth-income ratios and top wealth shares (Piketty and Zucman [2014] and Saez and Zucman [2014]). Existing data is available online on the World Top Income Database<sup>1</sup> (WTID) [Alvaredo et al., 2015], a database that is currently being subsumed into a broader World Wealth and Income Database (W2ID) (Alvaredo, Atkinson, Piketty, Saez and Zucman, 2015). The present master thesis contributes to this literature by developing new nonparametric methods to estimate the shape of income and wealth distributions and to generate reliable synthetic micro-files.

The construction of long run top income series usually rests on income tax tabulations released by fiscal administration. Commonly, in countries where a progressive income tax is established, tax authorities publish annually tables assembling taxpayers by range of income, giving the number of taxpayers in each bracket and the total income they earned. The thresholds in these administrative statistics are those of the income tax and do not coincide with the considered groups such as the top 10% or the top 1%. To interpolate the figures of interest, some assumption has to be made on the shape of the distribution between the different tax thresholds. Typically, the top of the income distribution is said to follow a Pareto law.

Indeed, it is widely accepted that the upper tail of the income distribution is Paretian while the middle part is lognormal. Hence, the distribution of incomes is usually represented in a piecemeal fashion rather than as a whole. Distinct functional forms are adopted to fit the different parts of the distribution. Economists generally take a parametric approach to modeling the taxpaying population. This approach implicitly assumes that observed incomes in the population are realizations of a random variable following an unknown probability distribution which belongs to some parametrized family of probabilistic models. Finding the underlying distribution then

---

<sup>1</sup>Accessible at <http://topincomes.g-mond.parisschoolofeconomics.eu/>.

comes down to fit the parameters to the data observed in the tax tabulations. To evaluate how well the selected statistical model fits the data, various measures of goodness-of-fit summarize the discrepancies between observations and the values expected under the model at issue. Monte Carlo simulations allows to appraise numerically the likelihood that the sample do follow the specified model. However, such tests often reject the Pareto hypothesis (see Clauset et al. [2009] or Cho et al. [2015] for instance). As soon as the Paretian form of the distribution is taken for granted, it is used to infer the shape of the distribution and to compute the desired estimates.

On the opposite, we adopt a nonparametric approach in this dissertation. We develop a method that allows to compute statistical quantities without deploying any statistical model (such as a lognormal distribution or a Pareto distribution). We do not assume any more the existence of an underlying distribution under which the incomes of the observed population would be drawn. We only consider the "true" empirical distribution of a large number of individual incomes, which we approximate by a more convenient and manageable continuous probability distribution function. Starting with available data, we determine this empirical distribution function.

In nonparametric statistics, it is common to assess the shape of the probability distribution function ruling a sample of independent and identically distributed observations using kernel density estimations. These methods work whenever we can directly observe a set of draws from this probability distribution. In the issue in question, we do not observe a sample of individual incomes but only the data in the tax tabulations. By interpolating appropriately the data in these tables, we will be able to determine the shape of the income distribution.

We obtain an internally consistent representation of the whole taxpaying population. We suggest four main empirical applications of this work.

First, we are able to generate more precise estimations of the shares accruing to the most affluent percentiles of the population than standard methods. However, resulting series remain very close to the old ones. The hypothesis of a Paretian upper part is still very satisfactory. There is no doubt that the corrections made are negligible with respect to the inherent discrepancies of income tax tabulations-based estimations (namely, the collection of tax statistics through an administrative process which is thus not tailored to economists' needs, exemptions, tax avoidance and tax evasion).

Second, our method allows to compute estimations of shares on the lower and the middle parts of the income distribution for which usual methods were silent. Indeed, the Pareto hypothesis restricted the scope of studies to the upper tail. Here, once the fraction of the population filing tax returns is large enough, we are able to measure the share earned by the top 60% for instance.

Third, we can simulate numerically a sample of taxpayers whose incomes follow the same distribution as the incomes of the true population corresponding to the tax tabulations.

Fourth, we tackle the empirical issue of the comparability of series across countries which define different tax units (the household or the individual). Even within one country, changes in the tax legislation may hinder the creation of homogenous series.

We start by reviewing the methods previously used to estimate top income shares. They were based on a number of statistical models that aim at generalizing the lognormal and the

Pareto distributions. Such representations of the income distribution have theoretical grounds. A growing literature tries to explain the occurrence of Pareto upper tails and to analyze the economic mechanisms at work. We will survey this literature at the end of this second section.

We then turn to generalized Pareto curves. We first outline our nonparametric method from a theoretical point of view. We subsequently supply evidence of its accuracy and describe its potential applications using micro-files provided by French tax authorities for 2006.

In a last part, we apply our new technique using tabulations for France for incomes from 1915 and for inheritance from 1902.

## Section 2

# Existing literature on the Pareto law in income and wealth distributions

### 2.1 Previous attempts to estimate and generalize Pareto distributions

It is broadly agreed among economists that the upper tails of income and wealth distributions are well-described by the so-called Pareto law. This functional form was named after the Italian economist Vilfredo Pareto who found that this specific form fitted well to the income distributions which he was examining.

Let us first remind some basic facts about the Pareto distribution. In the rest of the section, we will first discuss some methods often used to make estimates based on fiscal sources. Then, we will describe past attempts to generalize this form.

#### 2.1.1 The Pareto distribution

##### 2.1.1.1 Presentation

In the 1890s, Vilfredo Pareto<sup>1</sup> looked into income distributions using data from England, Italian cities, German states, Paris and Peru [Pareto, 1896]. He took advantage of the newly implemented tax systems to collect data from tax tabulations. His pioneering empirical work helped provide statistical grounds to the politically and intellectually passionate debates on income distribution.

An engineer by training, Pareto had the idea to plot on a double-logarithmic scale the number of incomes above a certain threshold against the respective threshold. He claimed that the resulting graphs were parallel straight lines which translated into his well-known curve:

$$\log N = A - a \log x \tag{2.1}$$

where  $N$  is the number of households which income is greater than  $x$ ,  $A$  is a parameter and  $a$  is

---

<sup>1</sup>For an detailed reference on Pareto's work, see [Persky, 1992].

the opposite of the slope.

Taking exponential of both sides, this equation is equivalent to:

$$N(x) = Cx^{-a} \quad (2.2)$$

with  $C = e^A$ . This property characterizes Pareto distributions<sup>2</sup>.

### 2.1.1.2 Definition and basic facts

Formally, a random variable  $Y$  is said to follow a Pareto distribution if its survival function is given by:

$$\bar{F}(y) = \mathbb{P}(Y > y) = \begin{cases} \left(\frac{k}{y}\right)^a & \text{if } y \geq k, \\ 1 & \text{otherwise.} \end{cases} \quad (2.3)$$

where  $k > 0$  is the scale parameter and  $a > 1$  is a parameter determining the shape of the distribution.

Equivalently, the cumulative distribution function (CDF) of a Paretian random variable is:

$$F(y) = 1 - \left(\frac{k}{y}\right)^a \quad (2.4)$$

and its density (PDF) writes:

$$f(y) = \frac{ak^a}{y^{1+a}}. \quad (2.5)$$

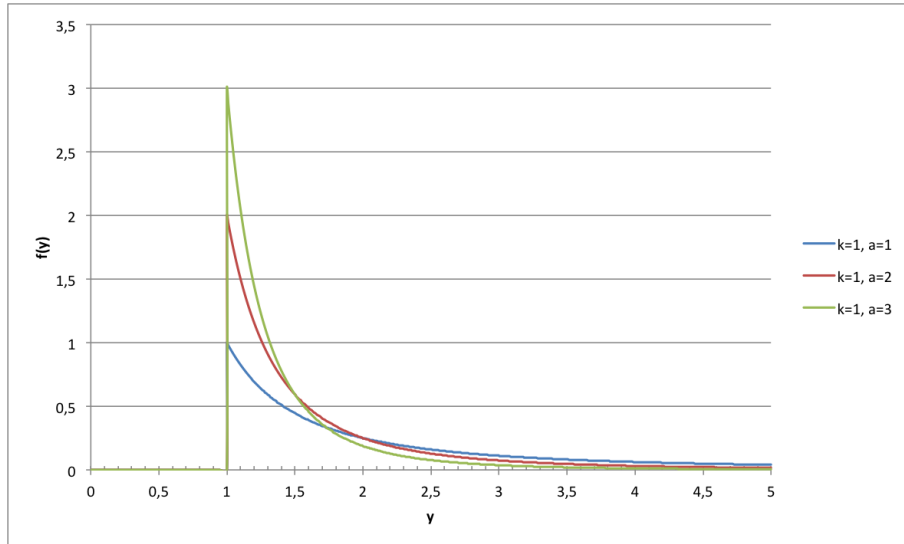


Figure 2.1: Density of the Pareto distribution

As noticed by Pareto, when the associated survival function is plotted employing logarithmic

---

<sup>2</sup>Which are sometimes called *power laws*.

scales on both axes, the graph turns out to be a straight line.

$$\log(\bar{F}(y)) = \log(1 - F(y)) = a \log(k) - a \log(y) \quad (2.6)$$

Pareto stressed that this asymmetric distribution was fundamentally different from a normal curve. It is skewed (and even one-sided) and heavy-tailed. Its density begins at a minimum income  $k$  and decreases monotonically afterwards.

A key property of the Pareto distribution is often referred to as the Van der Wijk's law [der Wijk, 1939]. For any level of income  $y$ , the average income of the subgroup that earns at least  $y$  is simply  $b \times y$  where  $b = \frac{a}{a-1}$ . To follow the notations in [Atkinson et al., 2011], the constant of proportionality  $b$  is called the (inverted) Pareto coefficient. More specifically,

$$y^*(y) = \mathbb{E}[Y|Y > y] = \frac{\int_{z>y} z f(z) dz}{\int_{z>y} f(z) dz} = \frac{\int_{z>y} \frac{dz}{z^a}}{\int_{z>y} \frac{dz}{z^{1+a}}} = \frac{a}{a-1} y \quad (2.7)$$

so that the ratio is independent of the income  $y$ .

A high Pareto coefficient  $b$  is associated with a distribution of income that exhibits a high level of inequality.

Pareto found out that the slopes of the plots that he had drawn for different periods and different countries lied in a narrow range, so that the values of the crucial parameter  $a$  actually clustered around 1.5. He inferred from this apparent stability that some natural law was ruling the distribution of incomes and hoped that he had discovered a universal constant that resulted from underlying economic mechanisms.

This earlier dogmatic interpretation has been discarded for long, but the fact is that the Paretian functional form fits well the distribution of top incomes. Consequently, the use of the Pareto law to describe and approximate the top of the distribution is now widely spread.

## 2.1.2 Common estimation methods

While early Pareto literature centered on the estimation of the parameter  $a$  which was supposed to characterize on its own the top of the income distribution, posterior tax-based research strove to describe the distribution accurately by constructing more meaningful income share time series. The Paretian functional form then proved to have convenient properties useful for technical manipulations.

### 2.1.2.1 Tax tabulations data

Standard tabulations made available by tax authorities give the fraction of the population in each tax bracket and generally to the total amount they earn. These tables are published every year in France by fiscal administration since the creation of the income tax in 1915. Thresholds are arbitrary and vary over time and across countries. Typically, the income intervals do not coincide with the groups of taxpayers we are concerned about. We cannot directly find figures of interest such as the share of national product accruing to the top 1% of households, their

average income or the threshold for being in the top 10%. To assess these quantities, we have to extrapolate from raw data by making some assumption about the shape of the distribution.

Let's assume that we observe in administrative tabulations a set of  $\omega$  taxable income brackets

$$[\theta_1, \theta_2), [\theta_2, \theta_3), \dots, [\theta_\omega, +\infty)$$

with  $0 \leq \theta_1 < \theta_2 < \dots < \theta_\omega < +\infty$ . For each interval  $[\theta_i, \theta_{i+1})$ , the number  $n_i$  of taxpayers whose income lies between  $\theta_i$  and  $\theta_{i+1}$  is also known. Often, tabulations give  $y_i$ , the total amount of taxable income declared by these taxpayers.

From this, we can easily compute  $\mu_i$ , the mean income of people in income class  $[\theta_i, \theta_{i+1})$ , and  $b_i$ , the empirical counterpart of Pareto coefficient  $b$  which is equal to:

$$b_i = \frac{1}{\theta_i} \frac{y_i + \dots + y_\omega}{n_i + \dots + n_\omega}. \quad (2.8)$$

We denote  $N = n_1 + \dots + n_\omega$  the total number of taxpayers,  $\phi_i = n_i/N$  the relative frequency within interval  $[\theta_i, \theta_{i+1})$ , and  $p_i = \phi_1 + \dots + \phi_{i-1}$  the fractile corresponding to the threshold  $\theta_i$ .

### 2.1.2.2 The interpolation problem

Using only the information published in tax tabulations, we want to find out about the true density function  $f$  (and its associated cumulative distribution function  $F$ ) in order to derive estimates of interest.

Notice that  $p_i$  is the empirical counterpart of  $F(\theta_i)$ , and similarly we observe:

$$\phi_i = \int_{\theta_i}^{\theta_{i+1}} f(y) dy, \quad (2.9)$$

$$\mu_i = \frac{1}{\phi_i} \int_{\theta_i}^{\theta_{i+1}} y f(y) dy, \quad (2.10)$$

$$b_i = \frac{1}{\theta_i} \frac{\int_{\theta_i}^{+\infty} y f(y) dy}{\int_{\theta_i}^{+\infty} f(y) dy} = \frac{1}{(1 - F(\theta_i))\theta_i} \int_{\theta_i}^{+\infty} y f(y) dy. \quad (2.11)$$

A first approach is to calculate lower and upper bounds consistent with available data and within which the true value of the quantity we care about must lie. For instance, the total amount earned by the top 10 percent is between

$$T_L = \frac{p_{i+1} - 0.1}{p_{i+1} - p_i} \phi_i \mu_i + \sum_{j \geq i+1} \phi_j \mu_j$$

and, if we define  $\alpha_i \in [0, 1]$  such that  $\mu_i = \alpha_i \theta_i + (1 - \alpha_i) \theta_{i+1}$ ,

$$T_U = \begin{cases} \frac{p_{i+1} - 0.1}{p_{i+1} - p_i} \phi_i \theta_{i+1} + \sum_{j \geq i+1} \phi_j \mu_j & \text{if } 0.1 \geq p_i + \alpha_i \phi_i, \\ \left( \alpha_i - \frac{0.1 - p_i}{p_{i+1} - p_i} \right) \phi_i \theta_i + (1 - \alpha_i) \phi_i \theta_{i+1} + \sum_{j \geq i+1} \phi_j \mu_j & \text{otherwise,} \end{cases}$$

where  $p_i \leq 0.90 < p_{i+1}$ .

These two bounds are met when all the taxpayers in this income class all earn the same income (lower bound) or if a fraction  $\alpha_i$  earns  $\theta_i$  and the others earn  $\theta_{i+1}$  (upper bound). But this range of values is quite large and boundaries would be reached under unrealistic conditions about the distribution.

As the minimal and maximal values would be hypothetically obtained with some specific distributions, we cannot produce more precise estimates without additional assumption about the shape of the underlying distribution. In reality, the taxpayers within each income interval do not all gather near one boundary of the bracket so that the true income distribution is presumably rather smooth.

A more satisfactory way to address the estimation problem is thus to suppose a certain distribution of incomes satisfying plausible assumptions<sup>3</sup> within tax intervals. We describe below alternative estimation methods depending on whether we know the income means  $\mu_i$ , as presented in the technical appendix of [Cowell, 2009]. Most of these methodologies rely on the assumption that the top of income distribution obeys a Pareto law.

More specifically, we assume that the density function on interval  $[\theta_i, \theta_{i+1})$  is  $f_i$ , a particular functional form whose parameters are determined using information we have on interval  $i$  (namely, its boundaries  $\theta_i$  and  $\theta_{i+1}$ , its relative frequency  $\phi_i$  and possibly its mean  $\mu_i$ ).

Then, we can compute the CDF:

$$F(y) = p_i + \int_{\theta_i}^y f_i(z) dz \quad (2.12)$$

and the total income of people earning at most  $y$ :

$$s(y) = s_i + \int_{\theta_i}^y z f_i(z) dz. \quad (2.13)$$

### 2.1.2.3 Estimation when intervals means are unknown

First, let us suppose that the only information we have about each interval  $[\theta_i, \theta_{i+1})$  is its relative frequency  $\phi_i$ .

**Histogram density** The simplest form we can think of is a density function which is uniform within each tax bracket. The density on interval  $i$  is given by:

$$f_i(y) = \frac{\phi_i}{\theta_{i+1} - \theta_i}, \quad \theta_i \leq y < \theta_{i+1}. \quad (2.14)$$

**Paretian density** Another approach is the one initially developed by Pareto [1896], and subsequently used by Kuznets [1953] and then by Feenberg and Poterba [1993] with income tax tabulations of the US.

---

<sup>3</sup>A list of assumptions that should be verified by the interpolated distribution is given by Cowell and Mehta [1982].

Taking advantage of the good fit of the Paretian form with the income distribution, we assume that the density within each bracket writes:

$$f_i(y) = \frac{a_i k_i^{a_i}}{y^{a_i+1}}, \quad \theta_i \leq y < \theta_{i+1} \quad (2.15)$$

where  $a_i$  and  $k_i$  are the parameters to find.

There are two alternative ways to identify these parameters, depending on where we put  $k$ , the minimal income above which the Paretian law holds.

- A first option is to assume that the Pareto form is defined on  $[\theta_i, +\infty)$ . In particular,  $k = \theta_i$ . Then, we have from the linearity of the Pareto diagram (2.6):

$$\log(1 - \phi_i) = a_i \log(\theta_i) - a_i \log(\theta_{i+1}) \quad (2.16)$$

which leads to:

$$a_i = \frac{\log(1 - \phi_i)}{\log\left(\frac{\theta_i}{\theta_{i+1}}\right)}. \quad (2.17)$$

- Another possibility is to assume that the Paretian form is valid on the whole distribution of incomes, and to assess from this hypothesis the value of  $k$ . Again, we make use of the linearity of the Pareto diagram (2.6) to get the two equations:

$$\log(1 - p_i) = a_i \log(k) - a_i \log(\theta_i) \quad (2.18)$$

and

$$\log(1 - p_{i+1}) = a_i \log(k) - a_i \log(\theta_{i+1}). \quad (2.19)$$

Subtracting (2.19) to (2.18), we get the estimate of  $a_i$ :

$$a_i = \frac{\log((1 - p_i)/(1 - p_{i+1}))}{\log(\theta_{i+1}/\theta_i)} \quad (2.20)$$

and then the estimation of  $k$ :

$$k = \theta_i (1 - p_i)^{1/a_i}. \quad (2.21)$$

This is the method that has been initially used by Pareto [1896], and then for the US by Kuznets [1953] and Feenberg and Poterba [1993].

The formula  $1 - F(y) = \left(\frac{k}{y}\right)^a$  then allows to extrapolate the whole distribution of the top incomes. For instance, to calculate the threshold and average income of the top 0.5%, first choose the bracket  $i$  the closer to  $p = 0.995$ .

$$P99.5 = \frac{k_i}{0.005^{1/a_i}} = \left(\frac{1 - p_i}{0.005}\right)^{(b_i-1)/b_i} \theta_i$$

$$P99.5 - 100 = b_i \cdot P99.5 = b_i \left(\frac{1 - p_i}{0.005}\right)^{(b_i-1)/b_i} \theta_i$$

We can deduce the threshold above which one household is in the top-earning 0.5%:

$$P99.5 = \frac{k}{0.005^{1/a}}$$

and the average taxable income that those households earn:

$$P99.5 - 100 = \frac{a}{a-1} P99.5 = \frac{a}{a-1} \frac{k}{0.005^{1/a}}.$$

Finally, the share of total taxable income earned by the top 0.5% of taxpayers is:

$$\frac{0.005 \times P99.5 - 100}{\bar{y}}$$

where  $\bar{y}$  is the average taxable income in the whole population.

#### 2.1.2.4 Estimation when intervals means are known

**Polynomial interpolation** A first idea is to consider a polynomial interpolation  $f_i$  of the density over each interval  $[\theta_i, \theta_{i+1})$ . The conditions on the boundaries and the means of the brackets boil down to the systems:

$$\begin{cases} \phi_i &= \int_{\theta_i}^{\theta_{i+1}} f_i(y) dy \\ \mu_i &= \frac{\int_{\theta_i}^{\theta_{i+1}} y f_i(y) dy}{\int_{\theta_i}^{\theta_{i+1}} f_i(y) dy} \end{cases} \quad (2.22)$$

If the degree of the polynomials  $f_i$  is higher than 2, the density is likely to exhibit turning points or even to become negative within the intervals. So in practice, the use of polynomials to approximate the density functions is limited to polynomials of order 1 or 2.

For instance, the straight line density is given by the formula<sup>4</sup>:

$$f_i(y) = b_i + c_i y, \quad \theta_i \leq y < \theta_{i+1} \quad (2.23)$$

where

$$b_i = \frac{12\mu_i - 6(\theta_{i+1} - \theta_i)}{(\theta_{i+1} - \theta_i)^3} \phi_i \quad (2.24)$$

and

$$c_i = \frac{\phi_i}{\theta_{i+1} - \theta_i} - \frac{1}{2}(\theta_{i+1} + \theta_i)b_i. \quad (2.25)$$

**Split histogram density** Another simple approach is to approximate the density function with an histogram. Now, to meet the two conditions on the relative frequency and the mean in

---

<sup>4</sup>See [Cowell, 2009].

the interval, we have to split each the bracket in two.

$$f_i(y) = \begin{cases} \frac{\phi_i}{\theta_{i+1}-\theta_i} \frac{\theta_{i+1}-\mu_i}{\mu_i-\theta_i} & \text{if } \theta_i \leq y < \mu_i, \\ \frac{\phi_i}{\theta_{i+1}-\theta_i} \frac{\mu_i-\theta_i}{\theta_{i+1}-\mu_i} & \text{if } \mu_i \leq y < \theta_{i+1}. \end{cases} \quad (2.26)$$

**Piecewise Paretian interpolation** Again, we can assume that the distribution is Paretian within each bracket and define:

$$f_i(y) = \frac{a_i k_i^{a_i}}{y^{a_i+1}}, \quad \theta_i \leq y < \theta_{i+1}. \quad (2.27)$$

- The easiest way is to use the  $b_i$  corresponding to the thresholds  $\theta_i$  given in the tabulations and to compute for each threshold:

$$a_i = \frac{b_i}{b_i - 1} \quad (2.28)$$

and:

$$k_i = \theta_i (1 - p_i)^{1/a_i}. \quad (2.29)$$

This is the method developed by Piketty [2001] for France.

- A second option is to solve the following system of equations. If we compute the two conditions of the system (2.22), we get:

$$a_i k_i^{a_i} = \frac{a_i \phi_i}{\theta_i^{-a_i} - \theta_{i+1}^{-a_i}} \quad (2.30)$$

where  $a_i$  is the root of the following equation<sup>5</sup>:

$$\mu_i = \frac{a_i}{a_i - 1} \frac{\theta_i^{1-a_i} - \theta_{i+1}^{1-a_i}}{\theta_i^{-a_i} - \theta_{i+1}^{-a_i}}. \quad (2.31)$$

### 2.1.2.5 Discussion of the methods

Histogram and polynomial density are the more elementary methods to interpolate the income distribution. However, the choice of such functional forms to fit the data has no theoretical support.

On the other hand, the usual assumption that the distribution of incomes among the rich follows (at least locally) a Pareto law is a rationale for the Paretian interpolation methods. Conversely, their accuracy relies crucially on the extent to which the Pareto hypothesis is satisfied. This approach has been validated by Feenberg and Poterba [1993] and Piketty [2001] by comparing results with micro data.

But the Paretian shape of the income distribution is questioned. As far back as 1905, Lorenz [1905] argued that "logarithm curves are more or less treacherous". Indeed, the use of logarithmic

---

<sup>5</sup>Which can be solved analytically.

scales in the Pareto diagram compresses the data and conceals irregularities. The response of Johnson [1937] to critics attacking the poor fit of the Pareto distribution is not fully convincing.

Actually, this assumption is never globally verified. Even locally, coefficients  $b_i$  fluctuate and the Pareto hypothesis is only approximately satisfied so that the obtained estimations remain imprecise. While common methods employed to analyze Pareto-like distributions merely give estimates of parameters, Clauset et al. [2009] suggest a rigorous statistical framework to discern whether the data do exhibit a Pareto behavior. When they apply their method to wealth distribution, they find no evidence that it obeys a Pareto law at all.

Furthermore, piecewise Paretian interpolation is quite unsatisfactory as the implied interpolated distribution is not plausible: the corresponding density is not continuous as the intervals are considered independently.

### 2.1.3 Usual representations of the income distribution

The most common continuous probability law is the normal (or Gaussian) distribution whose density can be expressed on  $\mathbb{R}$  as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad (2.32)$$

where  $\mu$  and  $\sigma^2$  are respectively the mean and the variance.

Its predominance is justified on a theoretical basis by the central limit theorem which states that the mean of a large number of independent and identically distributed random variables satisfying quite general conditions is approximately normally distributed, regardless of the law shared by these variables.

Consequently, the normal law is often employed to approximate the distribution of random variables whose underlying law is unknown and which are expected to result from many independent processes.

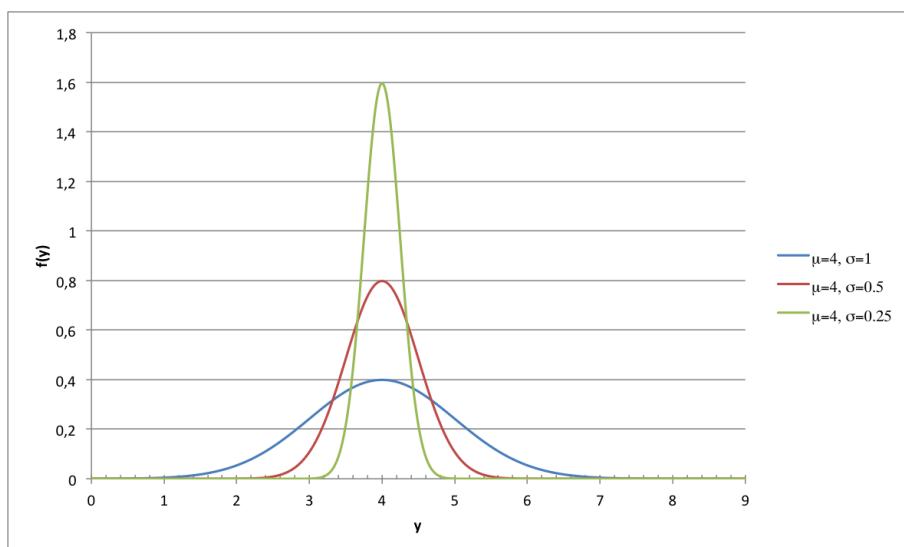


Figure 2.2: The normal distribution

Yet, the Gaussian functional form poorly fits income distributions. A first reason is that a normal distribution takes negative values. Even restricting attention to the positive part of the curve, it appears that the Gaussian shape does not look like the income distribution. The normal distribution is symmetrical about its mean (so that the mean, the median and the mode are indeed equal), and its density is nearly zero as soon as  $x$  is more than a few standard deviations away from the mean. On the contrary, the typical income distribution is positively skewed and heavy-tailed for top incomes. Concretely, if we set reasonable values for the means and the standard deviations, such as  $\mu = 30000$  €,  $\sigma = 15000$  €, there is the same absolute difference between the mean and the 10th percentile ( $p_{10} = 10777$  €) and between the mean and the 90th percentile ( $p_{90} = 49223$  €). The top thresholds are excessively low ( $p_{99} = 64895$  €,  $p_{99.9} = 76353$  €,  $p_{99.99} = 85785$  €) and the Pareto coefficient lies between 1.04 and 1.14 on the top 10 percent. The predicted shares for top-earning individuals are also widely underestimated ( $s_{90} = 18.8\%$ ,  $s_{99} = 2.1\%$ ,  $s_{99.9} = 0.03\%$ ).

### 2.1.3.1 Lognormal distribution

A more promising candidate to fit the shape of the income distribution is the lognormal law. By definition, a random variable  $Y$  follows a lognormal law if  $X = \log Y$  is normally distributed. One can check that the density of  $Y$  is then given by:

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right), \quad \forall y > 0. \quad (2.33)$$

The median of the lognormal distribution is  $e^\mu$ , its mean is  $e^{\mu+\sigma^2/2}$ , its mode is  $e^{\mu-\sigma^2}$  and its variance is  $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$ .

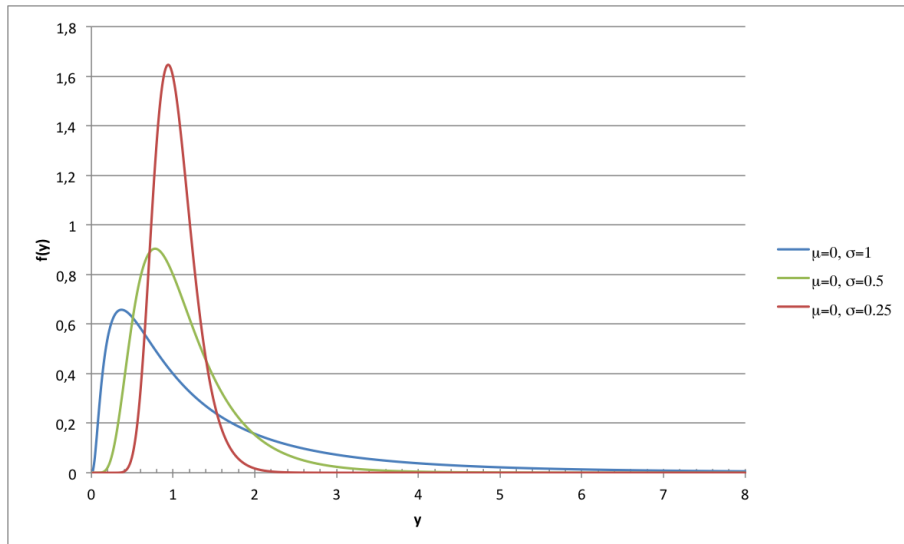


Figure 2.3: The lognormal distribution

The lognormal distribution appears to be well-suited to represent the distribution of incomes. Firstly, it takes only positive values. But an even more attractive feature is that it is positively

skewed and heavy-tailed<sup>6</sup>, just as the income distribution. Its mean is greater than its median which in turn is greater than its mode.

This functional form allows for a higher dispersion than the normal distribution. Indeed, the ratio of the 90th percentile to the median is now the same as the ratio of the median to the 10th percentile while the absolute differences of these percentiles were equal for the normal law. But to obtain a gap not too wide,  $\sigma$  has to be very low, which generates then an upper tail not spread out enough beyond the 10th percentile. More specifically, with a median  $\mu = 30000$  €,  $\sigma$  has to be close to 2 to get realistic values for  $p_{10}$  and  $p_{90}$ . This induces a Pareto coefficient ranging from 1.20 to 1.45 between  $p_{90}$  and  $p_{99.99}$  and top shares respectively equal to  $s_{90} = 27.8\%$ ,  $s_{99} = 5.1\%$  and  $s_{99.99} = 0.1\%$ . Actually, inequality is much more pronounced in the top 10 percent.

### 2.1.3.2 Pareto distribution

The Paretian functional form introduced in section 2.1.1 works well to describe the distribution of income and wealth among the rich. The Pareto hypothesis is usually made to represent the upper tail of the distribution.

In France, the Pareto coefficient  $b$  for top incomes was around 2.2 and 2.3 in the interwar years, and lies 1.7 and 1.8 since World War II. It has declined from WWII to the 1970s, and rises within top 10 percent since the 1970s.

### 2.1.3.3 Piecemeal distributions

As we have just seen, the middle part of the distribution (from about the 10th percentile to the 80th percentile) is well approximated by a lognormal law. The upper tail, that is, incomes above the 20th percentile, is better described by a Pareto distribution. To model the income distribution as a whole, one could take a piecemeal approach and try to "stick" a Paretian upper tail to a lognormal central part. In this way, the different parts of the income distribution would be appropriately approximated with different functional forms.

For this purpose, we have to determine some percentile  $p^*$  as the threshold above which the distribution is Paretian. Say,  $p^* = 0.9$  and  $y^*$  is the income such that  $F(y^*) = p^*$  if  $F$  is the CDF of a lognormal distribution. Therefore,  $y^*$  is determined by the parameters  $\mu$  and  $\sigma$  which are chosen for the lognormal distribution to fit the central part of empirical data. Then, the relation  $1 - p^* = (k/y^*)^a$  gives the value of  $k$  as a function of the parameter  $a$ .

A natural requirement to obtain a plausible functional form would be that the density is continuous and smooth at the point  $y^*$ . However, such a property cannot be satisfied: the continuity condition and the smoothness condition at the sticking point do not give the same value for  $a$  as a function of  $\mu$  and  $\sigma$ . Indeed, the slope of the lognormal density would imply an upper tail heavier than the continuity condition does.

---

<sup>6</sup>Formally, a distribution is said to have a heavy right tail if:

$$\lim_{x \rightarrow +\infty} e^{\lambda x} \overline{F}(x) = +\infty \quad \forall \lambda > 0, \quad (2.34)$$

where  $\overline{F}$  is the associated survival function. Examples of such probability laws include the lognormal distribution and the Pareto law.

A less stringent assumption is that the density is only required to be continuous. Then, there are three degrees of liberty, and we have several options for parametrization, depending on the choice of the three parameters.

#### 2.1.3.4 Empirical evidence

Clementi and Gallegati [2005] analyze income datasets for the United States (1980-2001), United Kingdom (1991-2001) and Germany (1990-2002). They advocate a mixture of the lognormal functional form (for the low-middle income group) and the Pareto function (for the high income group) to represent the income size distribution.

According to Aitchison and Brown [1957], the lognormal form is empirically appropriate for the distribution of earnings in homogenous occupational groups. They rely on disaggregated data for the year 1950 in Great Britain. The respective distributions of earnings in nine agricultural occupations (lorry drivers, stockmen, horsemen...) seem to be lognormally distributed.

Harrison [1979, 1981] also argues after examining British data that the Pareto distribution is less appropriate than the lognormal form to fit the upper tail of earnings distributions when data is disaggregated by occupational group. He explains the persistent validity of the Pareto distribution for the overall distribution by the significant variations of the standard deviations of logarithms of incomes  $\sigma$  across the different groups which prevent the aggregate from being lognormal.

#### 2.1.3.5 Other distributions

A range of functional forms have been suggested to fit the income distributions. They reproduce the shape of the lognormal and Paretian forms, but allow for more flexibility with a higher number of parameters.

They can be classified into several families. For an overview of these functional forms, see [Cowell, 2009] and for an inventory of their properties, see [Kleiber and Kotz, 2003].

**Three-parameter lognormal distribution** The random variable  $X$  is said to follow a three-parameter lognormal distribution if there exists a real number  $\lambda$  such that  $X = \ln(Y - \lambda)$  is normally distributed. Its PDF writes:

$$f(y) = \frac{1}{(y - \lambda)\sigma\sqrt{2\pi}} \exp\left(-\frac{[\ln(y - \lambda) - \mu]^2}{2\sigma^2}\right), \quad \forall y > \lambda. \quad (2.35)$$

According to Metcalf [1969], the lognormal distribution overcorrects for the positive skewness of the income distribution, so that the observed data exhibits negative skewness after logarithmic transformation. If the random variable  $Y$  is positively skewed and  $\ln Y$  negatively skewed, there exists a  $C > 0$  such that  $\ln(Y + C)$  has zero skewness. Therefore, Metcalf uses the three-parameter lognormal distribution to obtain the desired degree of skewness and claims that this functional form provides a good fit for the lower tail of the income distribution for US data from 1949 to 1965.

**Pareto-type distributions** The distribution previously named Pareto distribution is only one of the many functional forms suggested by Pareto himself. Its CDF could be expressed:

$$F(y) = 1 - \left[ \frac{y}{\sigma} \right]^{-\alpha}, \quad \forall y \geq \sigma, \quad (2.36)$$

with  $\sigma$  and  $\alpha$  positive parameters. Arnold [2015] uses the following typology.

**Pareto type II distribution** The CDF of a Pareto type II distribution is given by:

$$F(y) = 1 - \left[ 1 + \frac{y - \mu}{\sigma} \right]^{-\alpha}, \quad \forall y \geq \mu, \quad (2.37)$$

with  $\sigma, \alpha > 0$  and  $\mu \in \mathbb{R}$ .

**Pareto type III distribution** The CDF of a Pareto type III distribution writes:

$$F(y) = 1 - \left[ 1 + \left( \frac{y - \mu}{\sigma} \right)^{1/\gamma} \right]^{-1}, \quad \forall y \geq \mu, \quad (2.38)$$

where  $\sigma, \gamma > 0$  and  $\mu \in \mathbb{R}$ .

**Pareto type IV distribution** We can express the CDF of a Pareto type IV distribution as:

$$F(y) = 1 - \left[ 1 + \left( \frac{y - \mu}{\sigma} \right)^{1/\gamma} \right]^{-\alpha}, \quad \forall y \geq \mu, \quad (2.39)$$

with  $\sigma, \gamma, \alpha > 0$  and  $\mu \in \mathbb{R}$ .

The Pareto type II distribution corresponds to the special case  $\gamma = 1$  and the Pareto type III to the case  $\alpha = 1$ . With  $\gamma = 1$  and  $\mu = \sigma$ , we have the Pareto type I distribution.

**Champernowne distribution** Champernowne [1953] derives another distribution whose form is based on reasoning about processes of income generation. Its CDF writes:

$$F(y) = 1 - \frac{1}{\theta} \arctan \left( \frac{\sin \theta}{\cos \theta + [y/y_0]^\alpha} \right), \quad \forall y \geq 0. \quad (2.40)$$

This functional form is Paretian in the upper tail as:

$$1 - F(y) \sim_{y \rightarrow +\infty} C y^{-\alpha} \quad (2.41)$$

with  $C = \frac{1}{\theta} y_0^\alpha \sin \theta$ .

Thatcher [1968] provides empirical evidence with the distribution of earnings in Great Britain.

**Sech square distribution** Fisk [1961] analyzes a special case of the Champernowne distribution which he claims to fit reasonably well the income distributions which are homogenous

in occupation while remaining tractable. Its CDF is given by the formula:

$$F(y) = 1 - \frac{1}{1 + [y/y_0]^\alpha}, \quad \forall y \geq 0. \quad (2.42)$$

### Gamma-type distributions

**Gamma distribution** The well-known Gamma distribution has been used to approximate the income distribution by Salem and Mount [1974]. Its PDF is given by the formula:

$$f(y) = \frac{[y/y_0]^{\gamma-1} e^{-y/y_0}}{\Gamma(\gamma)}, \quad \forall y \geq 0, \quad (2.43)$$

with parameters  $\gamma \geq 0$  and  $y_0 > 0$ .

The authors show that the Gamma distribution outperforms the lognormal distribution to fit the personal income data in the United States for the years 1960 to 1969.

**Generalized Gamma distribution** Esteban [1986] speak in the Generalized Gamma distribution favor's from a theoretical perspective. This functional form has PDF:

$$f(y) = \frac{\beta [y/y_0]^{\beta\gamma-1} e^{-[y/y_0]^\beta}}{\Gamma(\gamma)}, \quad \forall y \geq 0, \quad (2.44)$$

with  $\beta > 0$ ,  $\gamma \geq 0$  and  $y_0 > 0$ .

**Weibull distribution** The CDF of the Weibull distribution can be expressed as:

$$F(y) = 1 - \exp(-[y/y_0]^\beta), \quad \forall y \geq 0, \quad (2.45)$$

where  $\beta$  and  $y_0$  are positive real numbers.

**Singh-Maddala distribution** Singh and Maddala [1976] generalize the Pareto distribution and the Weibull distribution in order to approach the entire income distribution with the following form:

$$F(y) = 1 - \frac{1}{(1 + [y/y_0]^\beta)^\alpha}, \quad \forall y \geq 0, \quad (2.46)$$

where  $\alpha$ ,  $\beta$  and  $y_0$  are positive parameters. They claim that this functional form fits remarkably well the US income data from the 1960s.

### Beta-type distributions

**Beta distribution** Thurow [1970] uses the Beta distribution to fit the size distribution of incomes. As the support of this distribution is bounded, a maximum value for the incomes has

to be determined. The PDF is:

$$f(y) = \frac{[y/y_0]^{\gamma-1}(1 - [y/y_0])^\alpha}{B(\gamma, \alpha + 1)}, \quad \forall 0 \leq y \leq y_0, \quad (2.47)$$

where  $\alpha \geq 0$ ,  $\gamma > 0$  and  $y_0 > 0$ . The Beta function  $B$  is a normalization constant defined as  $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$  for all positive  $x$  and  $y$ .

**Beta type II distribution** Slottje [1984] asserts that the Beta type II distribution provides a good approximation to empirical income data in the US for the years 1952-1980. The PDF is given by the formula:

$$f(y) = \frac{[y/y_0]^{\gamma-1}}{B(\gamma, \alpha + 1)(1 + [y/y_0])^{\alpha+\gamma+1}}, \quad \forall y \geq 0, \quad (2.48)$$

with  $\alpha \geq 0$ ,  $\gamma > 0$  and  $y_0 > 0$ .

**Generalized Beta distributions** The generalized Beta distribution of the first and the second kind are very flexible four-parameters distributions. They include the Beta, the Beta type II, the Singh-Maddala, the lognormal, the Gamma, the generalized Gamma, the Weibull, the Sech<sup>2</sup> and the exponential distributions as special or limiting cases. The interrelationships between these forms are given in [McDonald, 1984].

The generalized Beta distribution of the first kind has a bounded support. Its PDF can be expressed as:

$$f(y) = \frac{\beta[y/y_0]^{\beta\gamma-1}(1 - [y/y_0]^\beta)^\alpha}{B(\gamma, \alpha + 1)}, \quad \forall 0 \leq y \leq y_0, \quad (2.49)$$

with  $\alpha \geq 0$ ,  $\beta > 0$ ,  $\gamma > 0$  and  $y_0 > 0$ .

The PDF of the generalized Beta distribution of the second kind has a bounded support writes:

$$f(y) = \frac{\beta[y/y_0]^{\beta\gamma-1}}{B(\gamma, \alpha + 1)(1 + [y/y_0]^\beta)^{\alpha+\gamma+1}}, \quad \forall y \geq 0, \quad (2.50)$$

with  $\alpha \geq 0$ ,  $\beta > 0$ ,  $\gamma > 0$  and  $y_0 > 0$ .

**Empirical comparison** McDonald and Ransom [1979] compare the performance of the lognormal, the Gamma, the Beta, and the Singh-Maddala distributions with various estimation methods. Using family income data for 1960 and 1969 through 1975, they conclude that the Singh-Maddala distribution provides better fits than the others. Also, the Gamma distribution outdoes the lognormal distribution regardless of the estimation technique used.

McDonald [1984] discusses the alternative descriptive models for the income distribution. As mentioned above, he considers the Beta, Beta type II, Singh-Maddala, lognormal, Gamma, Weibull, Sech<sup>2</sup>, and exponential distributions as special or limiting cases of the generalized Beta distribution of the first and the second kind. The four-parameter generalized Beta function of the second kind appeared to fit the US family income data relatively better than the other distributions considered. But the three-parameter Singh-Maddala distribution did better than

all the others (except the generalized Beta distribution of the second kind), and its simple closed form makes it easy to manipulate.

## 2.2 Theoretical models yielding Pareto distributions

Pareto had the feeling that the regularities he had discovered in income distributions stemmed from underlying mechanisms. The fact that all income distributions are continuous, unimodal, highly skewed and heavy-tailed hints that some economic structure is behind their particular form. Pareto emphasized luck, social institutions and human nature as possible sources of inequality. He dismissed chance and social institutions as major determinants of the shape of the curve. The former because he had shown that if individual incomes followed from the accumulation of chance as represented by a simple binomial process, the right form could not be generated. The latter because the income distributions appeared to be similar across very different societies. He concluded that the distribution of incomes was either related to an underlying distribution of abilities or to the will of the elites to appropriate a certain share of the national resources.

The apparent empirical stability of the Pareto law gave rise to an abundant literature of generative models which attempt to explain the observed patterns in income and wealth distributions. They try to clarify why top incomes follow a Pareto distribution, and to specify which economic forces affect the inequality level embodied by the Pareto coefficient.

### 2.2.1 Earnings distribution

We first look at models dealing with wage distribution. Notably, these models do not include any accumulation process and are therefore simpler than models of wealth distribution. Economists suggested processes of individual income development as well as models depicting the hierarchical structure of society to describe the economic mechanisms involved.

Existing models attempt to make explicit the deciding factors at work. Are there some intrinsic characteristics that determine individual incomes? Is chance, or rather the accumulation of idiosyncratic shocks, sufficient to explain the spreading of the distribution? As put forward by Lydall [1959], the income distribution may be more than the mere aggregation of individual wages and may reflect a hierarchical structure prevailing within enterprises.

Models producing Pareto tails are mostly based on a few elementary mechanisms surveyed by Gabaix [2009, 2014]<sup>7</sup>. We examine below these mechanisms, detailing more carefully the shape of the income distribution predicted by each of them. Indeed, the resulting Pareto coefficient is not in general constant throughout the distribution, as it would be with a true Pareto distribution. As we shall see, the distribution is most of the time only expected to be asymptotically Paretian, with an upper tail close to power law (formally  $\mathbb{P}(Y > y) \propto y^{-a}$ ), in a sense to be defined.

As underlined by Gabaix [2009], power-law behavior (that is, the fact that the upper tail of the distribution is Paretian) is a stable property. Inheritance mechanisms ensure that, if we combine two Paretian variables, the one with the "fattest" tail will dominate. This stability property is true for the sum, the product, the maximum of random variables. Also, adding or multiplying by random variables which are not Paretian (normal, lognormal, or exponential variables for instance) preserves the Pareto exponent.

---

<sup>7</sup>Gabaix deals with the example of city sizes, but the same mechanisms work for incomes or wealth.

### 2.2.1.1 Models based on an underlying distribution of talent

Any simple static mechanism to generate labor income inequality requires two components:

1. some *heterogeneity* in the population, i.e. a certain attribute has to be unequally distributed;
2. a mechanism of *wage formation* specifying how the relevant attribute influences wage.

**Exponential growth** A first insight to understand how a Pareto distribution may emerge is analyzed by Jones [2015]: *exponential growth that occurs for an exponentially distributed amount of time generates a Pareto distribution*. Here, we assume first that people are unevenly endowed in talent or education or experience. We further hypothesize that the corresponding variable  $X$  is exponentially distributed:  $\mathbb{P}(X > x) = e^{-\delta x}$ . Secondly, wage will be assumed to grow exponentially with  $X$ :  $w(x) = e^{\mu x}$ . Under these hypotheses, one can easily show that earnings  $Y = w(X)$  obey a Pareto distribution of exponent  $a = \delta/\mu$ :

$$\mathbb{P}(Y > y) = y^{-\delta/\mu}. \quad (2.51)$$

This model dates back to Cantelli [1921].

**Matching and superstars effects** The previous model does not provide any justification for the alleged distribution of talent. It does not either clarify why wages should grow exponentially with talent.

Rosen [1981] suggests a more elaborate model to account for the skewness characterizing the top of the distribution. He argues that small differences in talent translate into large differences in revenue because the function associating talent to revenue is convex: "small differences in talent become magnified in larger earnings differences, with great magnification if the earnings-talent gradient increases sharply near the top of the scale". Imperfect substitution among sellers and consumption technologies allowing for scale economies in production jointly explain this winner-take-all phenomenon and the marked concentration of incomes at the top.

This *superstar effect* is studied by Gabaix and Landier [2008] who provide a calibrated model for the market for CEOs. Their model is of specific interest for us as it is fully calculable. Their model makes explicit the matching of CEOs and firms. Authors use extreme value theory to justify intellectually assumptions about the distribution of attributes. Extreme value theory shows that the spacing between talents takes on a quite universal form: for any "regular" distribution, rank in the upper tail is of the form

$$T'(x) = -Bx^{\beta-1} \quad (2.52)$$

with  $B$  a constant, up to a "slowly varying function"<sup>8</sup>.

---

<sup>8</sup>See [Gabaix, 2009] or [Gabaix and Landier, 2008] for details.

Gabaix and Landier [2008] find that the distribution of wages is Paretian in the upper tail. The earnings distribution is not an exact Pareto distribution any more.

**Roy’s model of multiplicative talent** According to Roy [1950], personal productivity is the product of a myriad of attributes which are i.i.d. random variables.

Wage depends linearly on talent, so that the earnings are lognormally distributed.

**O-Ring theory** In Kremer [1993]’s O-Ring theory, the production function induces complementarity between workers’ skills. The marginal product of a worker increases with the quality of other workers. In equilibrium, workers with the same skills work together. This implies that the distribution of wages is more skewed than the distribution of abilities, and is a Paretian.

### 2.2.1.2 Chance and Markov processes

Stochastic theories of the earnings distribution originate in Champernowne [1953]. Income does not depend on abilities or talent, but evolves randomly over time. Income is embodied by a Markov process. This account for the fact that chance is the only factor that affects the evolution of wage, and that the past history of its evolution does not matter.

In these models, a Paretian distribution arise eventually for a given pattern of social mobility regardless of the income distribution originally prevailing.

**Proportional random growth processes** The following mechanism is a dynamic model of the evolution of income (or alternatively wealth) over time in which the steady state distribution has a Pareto tail. Individual incomes are depicted as stochastic processes which fluctuate as time goes by. They grow proportionally according to a random variable. Their overall distribution tends to some equilibrium whose upper tail is Paretian.

Unlike the exponential growth model, this model does not assume the existence of any individual characteristic determining wage. The individual variations are purely random. But considered as a whole, the distribution of incomes in the entire population appears to have a Pareto tail. This statistical regularity comes from the overall structure of social mobility. As we shall see, it is the distribution of the growth of incomes with respect to the growth of the mean that matters for the shape of the steady state income distribution. The more dispersed this distribution is, the higher inequality.

The idea was first developed with discrete probability distributions by the statistician Yule [1925] to explain the distribution of biological species and genera. The economist Simon [1955] underlined the broad range of potential applications in sociology, biology and economics. The family of random growth models for income distributions was initiated by Champernowne [1953].

We first provide the main intuition behind this class of models as it is described by Gabaix [2009]. Then we give the result proved rigorously by Kesten [1973].

**Intuition** The population is represented by a continuum of individuals of mass 1. Each individual  $i$  earns an income  $Y_t^i$  at time  $t$ . Incomes are normalized so that the average in the

population is always equal to 1. The rationale for detrending incomes is that we want to ensure the existence of a limiting distribution.

Individual incomes raise between time  $t$  and  $t + 1$  by a gross growth rate  $\gamma_{t+1}^i$ :

$$Y_{t+1}^i = \gamma_{t+1}^i Y_t^i. \quad (2.53)$$

As incomes are normalized,  $\gamma_{t+1}^i$  must be interpreted as the growth rate compared with the average:  $\gamma_{t+1}^i$  is less than 1 if the income of individual  $i$  grows slower than the mean income and is greater than 1 if it grows faster.

The central assumption is that growth rates  $\gamma_{t+1}^i$  are identically and independently distributed under a density  $g(\gamma)$  at least in the upper tail.

We denote  $\bar{F}_t(y) = \mathbb{P}(Y_t^i > y)$  the counter-cumulative distribution function of incomes at time  $t$ . Its law of motion of  $\bar{F}_t$  is given by:

$$\begin{aligned} \forall y \geq 0, \quad \bar{F}_{t+1}(y) &= \mathbb{P}(Y_{t+1}^i > y) \\ &= \mathbb{P}(\gamma_{t+1}^i Y_t^i > y) \\ &= \mathbb{P}\left(Y_t^i > \frac{y}{\gamma_{t+1}^i}\right) \\ &= \int_0^\infty \bar{F}_t\left(\frac{y}{\gamma}\right) g(\gamma) d\gamma. \end{aligned}$$

Therefore the steady state distribution must satisfy if it exists:

$$\forall y \geq 0, \quad \bar{F}(y) = \int_0^\infty \bar{F}\left(\frac{y}{\gamma}\right) g(\gamma) d\gamma. \quad (2.54)$$

The Paretian functional form  $\bar{F}(y) = \left(\frac{k}{y}\right)^a$  (with  $k$  a constant) is a solution if  $a$  is a root of:

$$1 = \int_0^\infty \gamma^a g(\gamma) d\gamma \quad (2.55)$$

This condition is equivalent to the simple equation (2.56) called *Champernowne's equation*<sup>9</sup> by Gabaix [2009]:

$$\mathbb{E}[\gamma^a] = 1. \quad (2.56)$$

This relation relates the value of  $a$  to the dispersion of the values of  $\gamma$ . Nirei [2009] proves that if  $\mathbb{E}[\gamma] < 1$  and under a few other non-restrictive conditions on the distributions of the shocks, the exponent  $a$  is decreasing in the variance of the gross rate  $\gamma$ . This means that inequality is higher in the top of the income distribution when there are large random variations of the growth rate.

Gibrat [1931] pointed out that the distribution of incomes cannot converge to a steady state if (2.53) holds throughout the distribution. Indeed, the variance of the income distribution would write:

$$\text{Var}[\ln S_t^i] = \text{Var}[\ln S_0^i] + \text{Var}[\ln \gamma] t \quad (2.57)$$

---

<sup>9</sup>First published in [Champernowne, 1953].

and would grow to infinity.

Consequently, as suggested by Gabaix [2009], we need to deviate from pure random growth processes and to add some frictions that prevent incomes from becoming too small. For instance, a positive constant may be included in (2.53) or a reflecting barrier may enforce a lower bound for incomes. These frictions affect only low incomes, so that the Pareto exponent is unchanged.

**Kesten processes** Kesten [1973] is credited with the first rigorous study of random growth processes of the form  $Y_t = A_t Y_{t-1} + B_t$  where  $(A_t, B_t)$  are i.i.d. random variables. We reproduce below his main result predicting the asymptotic behavior of the limiting distribution.

**Theorem 2.2.1 (Kesten, 1973)**

Let for some  $a > 0$ ,

$$\mathbb{E}[|A|^a] = 1 \quad (2.58)$$

and  $\mathbb{E}[|A|^a \max(\ln(A), 0)] < +\infty$ ,  $0 < \mathbb{E}[|B|^a] < +\infty$ . Also, suppose that  $B/(1 - A)$  is not degenerate (i.e. can take more than one value), and the conditional distribution of  $\ln|A|$  given  $A \neq 0$  is non lattice (i.e. has a support that is not included in  $\lambda\mathbb{Z}$  for some  $\lambda$ ), then there are constants  $k_+$  and  $k_-$ , at least one of them positive, such that

$$x^a \mathbb{P}(Y > x) \longrightarrow k_+, \quad x^a \mathbb{P}(Y < -x) \longrightarrow k_- \quad (2.59)$$

as  $x \rightarrow +\infty$ , where  $Y$  is the solution of  $Y \stackrel{d}{=} AY + B$ . Furthermore, the solution of the recurrence equation  $Y_{t+1} = A_{t+1}Y_t + B_{t+1}$  converges in probability to  $Y$  as  $t \rightarrow +\infty$ .

**Continuous-time processes** The proofs with discrete-time processes are rather technical. The benchmark provided by the theory of stochastic calculus and the theory of stochastic differential equations make computations much easier for continuous-time processes. The density of the steady state distribution is found as the solution of an ordinary differential equation.

If individual incomes obey the stochastic differential equation

$$dY_t = \mu(Y_t)dt + \sigma(Y_t)dz_t \quad (2.60)$$

where  $z_t$  is a Brownian motion, then from Kolmogorov equation the density of the limiting distribution is a solution of

$$0 = \partial_x[\mu(x)f(x)] + \partial_{xx}\left[\frac{\sigma^2(x)}{2}f(x)\right]. \quad (2.61)$$

For instance, in the case of a random growth process, if  $\mu(X) = gX$  and  $\sigma(X) = vX$ , the solution is Paretian in the upper tail.

**Reed's model** Reed [2001] shows that if individual incomes follow a geometric Brownian

motion, and if they are observed after an exponentially distributed time  $T$  (the age of individuals), then the overall distribution obeys a *double Pareto distribution*, which has a power-law behavior both in the lower and in the upper tails.

In this model, we notice that if observed individuals had all the same age, the distribution would be lognormal (since individual trajectories follow a geometric Brownian motion). But the stacking of lognormal distributions associated with different standard deviations causes an overall Paretian behavior. The age heterogeneity conveys a fatter upper tail.

### 2.2.1.3 Lydall's model of social hierarchy

Lydall [1959] explains the overall distribution of incomes by the pyramid pattern of the social hierarchy. Each supervisor controls a fixed number of persons. The wage of a supervisor is assumed to depend on the aggregate income of the persons that he immediately supervises. These assumptions lead to a Pareto distribution of wages.

## 2.2.2 Accumulation models for wealth distribution

Explaining the shape of the wealth distribution is essentially explaining the heterogeneity in wealth accumulation across individuals. To understand this heterogeneity, we have first to figure out the individuals' motives to save, that is, to grasp why rational agents choose to allocate a part of their resources to savings rather than to direct consumption.

Piketty and Zucman [2015] give a detailed account of the related literature. We mainly follow their presentation in this section.

### 2.2.2.1 Motives for wealth amassing

The early literature paid little attention to the rationale of savings. Keynes [1936] referred to a "fundamental psychological law" which characterized the average saving behavior in the population: "men are disposed, as a rule and on the average, to increase their consumption as their income increases but not by as much as the increase in the income". In other words, the marginal propensity to save is greater than zero but less than unity.

To escape bare considerations on psychological inclinations, one has to go beyond this purely static framework. Indeed, dynamic models are needed to make out why agents do not only worry about direct consumption. Two main rationales have been underscored in the literature. First, lifecycle motives: individuals save while they are young, and dissave after retirement to maintain their consumption level. Second, dynastic altruism: individuals care about their descendants and want to leave them a bequest when they die.

Accordingly, the wealth held by individuals will primarily depend on their age and on their expected length of retirement. Within the same cohort, wealth will depend on opportunities (the lifetime resources of agents), and on preferences (dynastic altruism for bequests, risk aversion which determines precautionary savings, taste for a social status granted by wealth, etc).

**Lifecycle motive** Modigliani [1986] claims that the life-cycle hypothesis provides an apt description of wealth accumulation patterns. The key assumption is that agents smooth their consumption over their lifetime. Wealth accumulation is thus driven by lifecycle motives.

In the "stripped-down" version of this model, income is constant until retirement, and is zero thereafter. The interest rate is zero. Individuals prefer a constant consumption over life, and leave no bequest.

At the individual level, accumulated wealth follows a hump-shaped path. The length of retirement is the main parameter controlling the form of the so-called *Modigliani triangle*. Formally, if the agent earns an income  $\bar{Y}$  during adulthood, if retirement occurs at age  $N$  and if  $L$  is the age of death, the wealth holdings have the following profile:

$$W(T) = \begin{cases} \frac{L-N}{L} T \bar{Y} & \text{if } 0 \leq T \leq N, \\ \frac{N}{L} (L - T) \bar{Y} & \text{if } N \leq T \leq L. \end{cases} \quad (2.62)$$

Consumption is then equal to:

$$C(T) = \frac{N}{L} \bar{Y}. \quad (2.63)$$

It is then constant throughout life.

To assess the distribution of wealth at the aggregate level, some demographical structure has to be incorporated in the model. The overlapping-generation model, while explicitly recognizing the finite life of individuals, allows to think about the wealth distribution in a pure age war context. Individuals live for two periods. They save when they are young, and they consume all their savings when they are old. This model can encompass population growth, economic growth and positive interest rate. Wealth inequality reflects both the age distribution and the income distribution.

This range of models generate reasonable values for wealth-income ratios (between 5 and 10). However, wealth inequality merely mirrors wage inequality. Data paint a completely different picture: wealth concentration is far more pronounced than income concentration. The problem is that individuals are assumed to leave no inheritance to their descendants when they die. Consequently, cumulative effects across generations due to inheritance transmission are concealed in this framework. Inheritance does matter in the wealth accumulation process.

**Dynastic altruism** Another motive to save a part of one's resources lies in dynastic altruism. Individuals care about their offspring and want to leave them a bequest. We present here two alternative specifications that can account for this altruistic concern.

**Bequest in the utility** A first option is to fashion the instantaneous utility function in order to take into account the will to leave a bequest. The wealth-increase-in-the-utility model relates the case where the agent is only concerned about leaving a higher wealth to his successors than he received from his parents. The corresponding utility function writes:

$$U(c, \Delta) = c^{1-s} \Delta^s, \quad (2.64)$$

so that at each period the agent's program is:

$$\max U(c_t, \Delta w_t), \quad \text{subject to } c_t + \Delta w_t \leq y_t, \quad (2.65)$$

where  $c_t$ ,  $y_t$ , and  $w_t$  denote respectively consumption, income, and wealth at period  $t$ , and  $\Delta w_t = w_{t+1} - w_t$  is the wealth increase between  $t$  and  $t + 1$ .

This Cobb-Douglas specification leads to  $\Delta w_t = sy_t$ .

Besides, the bequest-in-the-utility model depicts a situation where the individual is driven by the desire to leave the largest possible bequest at the time of death. The utility function is then:

$$U(c, w) = c^{1-s} w^s \quad (2.66)$$

and the optimization problem

$$\max U(c_t, w_{t+1}), \quad \text{subject to } w_{t+1} \leq w_t + y_t - c_t, \quad (2.67)$$

leads to  $w_{t+1} = s(w_t + y_t)$ .

**Dynastic model** In the same vein, the dynastic model provides an infinite-horizon framework where each individual maximizes the dynastic utility function

$$V_T = \int_{t=T}^{+\infty} e^{-\theta t} U(c_t) dt. \quad (2.68)$$

where  $\theta$  is the (fixed) rate of time preference and  $U(c) = (1 - \gamma)c^{1-\gamma}$  is the utility function with a constant intertemporal elasticity of substitution  $1/\gamma$ .

The long-run rate of return  $r$  is determined by tastes and by the growth rate and is greater than  $g$ . Its expression is given by the so-called modified Golden Rule of capital accumulation,

$$r = \theta + \gamma g.$$

These wealth accumulation frameworks account for a pure class war situation. One drawback of such models is that they allow for no social mobility. Wealth inequality is self-sustaining. Hence these models explain why wealth inequalities perpetuate, but they do not say anything about why they appear and how they evolve.

### 2.2.2.2 Random shocks and cumulative effects

Lifecycle and dynastic models describe the reasons why people accumulate wealth. However, these models do not provide any insight into the shape of the wealth distribution and do not encompass social mobility. In this setting, wealth inequality merely mirrors income inequality and age distribution.

These models constitute a framework upon which more elaborate models of wealth distribution are based. Multiplicative random shocks models of wealth combine analysis of the individ-

ual wealth accumulation process, describing how assets accumulate over time, and heterogeneity birth-and-death process, stochastic savings etc. In the long-run, we observe both convergence of macroeconomic variables to steady-state values and of the wealth distribution to a limiting distribution. Such models are ergodic inasmuch the steady state does not depend on the initial conditions.

The individual parameters determining the amount of wealth that will be accumulated by different individuals are distributed randomly across the population. Agents also face random shocks that affect their income and their assets. Wealth inequality springs from cumulative effects: multiplicative random shocks that accumulate over generations, together with preferences and income inequality, explain the high inequality level characterizing the wealth distribution.

Idiosyncratic shocks that affect the individuals can be demographic (number of children, age at death), or can hit the rates of return of assets, the bequest tastes or labor productivities.

Wold and Whittle [1957] present a mechanism of divided inheritance which leads to a Paretian upper tail of the wealth distribution. Stiglitz [1969] studies the implication for both the distribution of income and the distribution of wealth of alternative assumptions on savings behavior, demography, inheritance policies, labor heterogeneity and taxation patterns in the context of a neoclassical growth model. He identifies equalizing forces and forces that make the distribution more unevenly distributed. Cowell [1998] provides a simple model of inheritance and relates demographical features of the population, such as the distribution of family sizes, marriage patterns, taxation or savings habits to the distribution of wealth and to its Paretian upper tail.

Benhabib and Zhu [2008] suggests a mechanism that generates a double Pareto distribution in the benchmark of an overlapping generation model. The effects of inheritance, stochastic returns on capital, uncertain lifespan and fiscal policies on wealth inequality are investigated by the authors. Benhabib et al. [2011] show that capital income risks, rather than labor income risks, drive the properties of the Paretian right tail of the wealth distribution

Nirei [2009] proves that if households undergo random investment shocks in some neoclassical growth model, the income and wealth distributions converge to a Pareto distribution. He relates the Pareto exponent to the shock variance, to the economy growth rate, and to redistribution policies.

More recently, the central role of  $\bar{r} - g$  (where  $\bar{r}$  is the net-of-tax rate of return) has been emphasized in wealth inequality. Piketty and Zucman [2015] give a detailed account of the related literature. For instance, Rodriguez [2014] derives a model where wealth obeys a Pareto type II distribution. Jones [2015] describes simple models of wealth distribution and warns that comparative statics can change depending on whether the model is considered in partial or in general equilibrium.

**A model with bequest tastes shocks** Piketty and Zucman [2015] provide a discrete-time model of a closed economy that exemplifies multiplicative random shocks models of wealth distribution. Here, the cumulative effects arise from the random shocks that affect the bequest tastes  $s_{ti}$  in a bequest-in-the-utility framework.

The stationary population is represented by a continuum of agents  $N_t = [0, 1]$ . Effective labor input grows at an exogenous productivity rate  $g$ ,

$$L_t = N_t h_t = h_0 (1 + g)^t. \quad (2.69)$$

The production function writes

$$Y_t^d = F(K_t, L_t). \quad (2.70)$$

Each agent  $i$  receives the same labor income  $y_{i,t}^L = y_t^L$ . The rate of return  $r_{i,t} = r_t$  is common to all.

Each individual  $i$  chooses  $c_{i,t}$  and  $w_{i,t+1}$  to maximize a Cobb-Douglas utility function of the form  $U(c_{i,t}, w_{i,t+1}) = c_{i,t}^{1-s_{i,t}} w_{i,t+1}^{s_{i,t}}$  with bequest taste parameter  $s_{i,t}$  under the budget constraint

$$c_{i,t} + w_{i,t+1} \leq y_t^L + (1 + r_t)w_{i,t}. \quad (2.71)$$

Random shocks  $s_{i,t}$  are i.i.d. random processes with mean  $s = \mathbb{E}[s_{i,t}] < 1$ . This specification leads to

$$w_{i,t+1} = s_{i,t}[y_t^L + (1 + r_t)w_{i,t}]. \quad (2.72)$$

At the aggregate level,  $y_t = y_t^L + r_t w_t$  and

$$w_{t+1} = s[y_t^L + (1 + r_t)w_t] = s[y_t + w_t] \quad (2.73)$$

In order to study the steady-state distribution of wealth, let us now consider the normalized individual wealth  $z_{i,t} = w_{i,t}/w_t$ . As we have  $w_{t+1} = (1 + g)w_t$  in the long run, the transition equation for wealth at the individual-level can be written:

$$z_{i,t+1} = \frac{s_{i,t}}{s} [(1 - \omega) + \omega z_{i,t}] \quad (2.74)$$

where we have set

$$\omega = s \frac{1 + r}{1 + g} < 1. \quad (2.75)$$

The theory of Kesten processes ensures that the distribution  $\Psi_t(z)$  of relative wealth will converge toward a unique steady-state distribution  $\Psi(z)$  with a Pareto shape and a Pareto coefficient that depends on the variance of taste shocks  $s_{i,t}$  and on the  $\omega$  coefficient. If  $\omega_{i,t} = \omega \frac{s_{i,t}}{s}$ , the Pareto coefficient  $a$  is such that the Champervorne's equation is satisfied:

$$\mathbb{E}[\omega_{i,t}^a] = 1. \quad (2.76)$$

Then,  $\omega = \mathbb{E}[\omega_{i,t}] < 1$ .

In the special case when  $s_{i,t} = 0$  with probability  $1 - p$  and  $s_{i,t} = s/p > 0$  with probability  $p$  (or equivalently  $\omega_{i,t} = 0$  with probability  $1 - p$  and  $\omega_{i,t} = \omega/p$ ), we have:

$$p \left( \frac{\omega}{p} \right)^a = 1. \quad (2.77)$$

The Pareto coefficients are then given by the formulae:

$$a = \frac{\log(1/p)}{\log(\omega/p)} > 1 \quad (2.78)$$

and

$$b = \frac{a}{a-1} = \frac{\log(1/p)}{\log(1/\omega)} > 1. \quad (2.79)$$

An increase in  $\omega$  triggers a fall in  $a$  and a rise in  $b$ . The higher  $\omega$ , the more marked the concentration of wealth. At the first order,  $\omega$  is an increasing function of  $r - g$ . Thus, as underlined by Piketty and Zucman [2015], when  $\omega$  rises, "the multiplicative wealth inequality effect becomes larger as compared to the equalizing labor income effect".

**A model with random shocks affecting the rates of return** Nirei [2009] develops a wealth accumulation model in a neoclassical growth framework where shocks affect the rates of returns of agents. In other words, there is an uninsurable and undiversifiable investment risk.

The population is formed of a continuum of infinitely-living individuals  $i \in [0, 1]$ . They can be interpreted as dynasties.

Each agent is endowed with one unit of labor and an initial capital  $k_{i,0}$ . Nirei [2009] specifies a "backyard" production technology which takes the form of a Cobb-Douglas production function:

$$y_{i,t} = k_{i,t}^\alpha (a_{i,t} l_{i,t})^{1-\alpha} \quad (2.80)$$

where  $l_{i,t}$  is the labor employed by  $i$  and  $k_{i,t}$  is the capital owned by  $i$ . The labor-augmenting productivity  $a_{i,t}$  is an i.i.d. random variable across households and across periods with a common trend  $\gamma > 1$ :

$$a_{i,t} = \gamma^t \varepsilon_{i,t} \quad (2.81)$$

where the  $\varepsilon_{i,t}$  are temporary productivity shocks, with  $\mathbb{E}[\varepsilon_{i,t}] = 1$ .

Crucially, an individual does not have means to insure against the productivity shocks  $\varepsilon_{i,t}$  except for his own savings.

Each household supplies one unit of labor inelastically. At the individual level, the capital accumulation follows:

$$k_{i,t+1} = (1 - \delta)k_{i,t} + s(\pi_{i,t} + w_t) \quad (2.82)$$

where  $s$  is a constant savings rate and  $\pi_{i,t} + w_t$  is the income of household  $i$ .  $\pi_{i,t}$  denotes the profit from the production:

$$\pi_{i,t} = \max_{l_{i,t}, y_{i,t}} y_{i,t} - w_t l_{i,t}. \quad (2.83)$$

The Cobb-Douglas specification of the production function leads to  $\pi_{i,t} = \alpha y_{i,t}$  and  $w_t l_{i,t} = (1 - \alpha)y_{i,t}$ .

At the aggregate level, we have:

$$\int_0^1 l_{i,t} di = 1, \quad (2.84)$$

aggregate output writes:

$$Y_t = \int_0^1 y_{i,t} di \quad (2.85)$$

and capital is:

$$K_t = \int_0^1 k_{i,t} di. \quad (2.86)$$

Therefore, wage is given by  $w_t = (1 - \alpha)Y_t$  and the equation of motion for the aggregate capital in the Solow model is:

$$K_{t+1} = (1 - \delta)K_t + sY_t. \quad (2.87)$$

To study the steady-state distribution, we are interested in detrended aggregate capital  $X_t = \frac{K_t}{\gamma^t}$ . Its (deterministic) law of motion writes:

$$\gamma X_{t+1} = (1 - \delta)X_t + s\eta X_t^\alpha \quad (2.88)$$

where:

$$\eta = \mathbb{E} \left[ \varepsilon_{i,t}^{(1-\alpha)/\alpha} \right]^\alpha \quad (2.89)$$

At the steady state, detrended aggregate capital is equal to:

$$\bar{X} = \left( \frac{s\eta}{\gamma - 1 + \delta} \right)^{1/(1-\alpha)}. \quad (2.90)$$

The dynamics of the detrended individual capital is characterized by the equation of motion:

$$x_{i,t+1} = g_{i,t}x_{i,t} + z. \quad (2.91)$$

That is, detrended individual capital is a Kesten process, with the return of detrended capital  $g_{i,t}$  given by:

$$g_{i,t} = \frac{1 - \delta}{\gamma} + \frac{\alpha(\gamma - 1 + \delta)}{\gamma} \frac{\varepsilon_{i,t}^{(1-\alpha)/\alpha}}{\mathbb{E} \left[ \varepsilon_{i,t}^{(1-\alpha)/\alpha} \right]} \quad (2.92)$$

and savings from detrended labor income  $z$  which can be expressed as:

$$z = \frac{s\eta(1 - \alpha)}{\gamma} \bar{X}^\alpha = \frac{s\eta(1 - \alpha)}{\gamma} \left( \frac{s\eta}{\gamma - 1 + \delta} \right)^{\alpha/(1-\alpha)}. \quad (2.93)$$

Again, well-known results about the convergence of Kesten processes guarantee that individuals' detrended capital  $x_{i,t}$  has a stationary distribution whose tail follows a Pareto distribution:

$$\mathbb{P}(x_{i,t} > x) \propto x^{-a}. \quad (2.94)$$

The Pareto exponent  $a$  is determined by Champernowne's equation:

$$\mathbb{E} [g_{i,t}^a] = 1. \quad (2.95)$$

Consider the special case where the returns shocks  $\log \varepsilon_{i,t}$  follow a normal distribution with mean  $-\sigma^2/2$  and variance  $\sigma^2$ . Then Nirei [2009] proves that there exists  $\bar{\sigma}$  such that, for all  $\sigma > \bar{\sigma}$ , the Pareto exponent  $a$  is uniquely determined by equation (2.95). This coefficient satisfies  $a > 1$  and the stationary distribution has a finite mean. Moreover,  $a$  is decreasing in  $\sigma$ . Intuitively, the higher the variance of multiplicative random shocks, the more pronounced wealth inequality.

## Section 3

# Generalized Pareto curves: theory and evidence

### 3.1 Theory

The standard methodology to compute estimates about the income distribution consists in assuming a Pareto shape, at least at the top. But if a Pareto distribution may be fitted to two points observed in the data, it does not exactly coincide with the other points. Even for top incomes, the Pareto coefficients  $b_i$  vary slightly and discrepancies affect the precision of the results.

In this part, we present a method that allows to relax the Pareto hypothesis. We take a nonparametric approach. Instead of trying to make a preconceived functional form fit the data by adjusting a set of parameters, we start from the observed Pareto curve. To obtain curves that are comparable among countries and homogenous over time, we consider the Pareto coefficients as a function of percentiles of the population.

The idea is to approximate the Pareto curve  $b(p)$  using the values  $b_1, \dots, b_\omega$  that we find in the administrative tabulations. Then, we are able to extrapolate the shape of the whole distribution of taxable income.

In this way, we can build sharper estimators that are not restricted to the top of the distribution. Besides, we can generate simulations of the population of taxpayers.

#### 3.1.1 The income distribution

Let  $F$  be the cumulative distribution function (CDF) of the income distribution (which is no longer assumed to be Paretian), and  $f = F'$  the associated probability density (PDF). Let us remind that a cumulative distribution function is a non-decreasing and right-continuous function such that:

$$\lim_{y \rightarrow -\infty} F(y) = 0, \quad \lim_{y \rightarrow +\infty} F(y) = 1.$$

As we are considering an income distribution, we assume that  $F(0) = 0$ .

Its inverse  $Q$  is called the quantile function defined by:

$$Q(p) = \inf\{y \in \mathbb{R} : p \leq F(y)\}. \quad (3.1)$$

$Q$  is a non-decreasing and left-continuous function.

In the following, we will assume that the quantile function  $Q$  is continuous, which means that there is no gap between values in the domain of the CDF.

The (inverted) Pareto coefficient can be expressed for any income  $y$ :

$$\tilde{b}(y) = \frac{1}{(1 - F(y))y} \int_y^{+\infty} z f(z) dz. \quad (3.2)$$

With  $p = F(y)$ , we can express  $b$  in percentiles:

$$b(p) = \frac{1}{(1 - p)Q(p)} \int_p^1 Q(r) dr$$

for all  $p \in [0, 1]$ .  $b(p)$  is finite if and only if  $Q(p)$  is positive.

$Q$  is an increasing function, thus we can define  $p_{min} \geq 0$ :

$$p_{min} = \inf\{r \in [0, 1] : Q(r) > 0\}.$$

As  $Q$  is continuous,  $Q(p_{min}) = 0$ . For all  $p \in [0, p_{min}]$ ,  $Q(p) = 0$ .

Intuitively,  $p_{min}$  is the fraction of the population with zero income (or zero wealth).

### 3.1.2 Pareto curve and quantile function

#### Definition 3.1.1 (*Pareto curve*)

A Pareto curve is a continuous function  $b : [0, 1] \rightarrow [0, +\infty]$ , such that:

(i)  $\{p \in [0, 1] : b(p) = +\infty\}$  is either empty or a closed interval including 0 that will be denoted  $[0, p_{min}]$  if we define:

$$p_{min} = \inf\{p \in [0, 1] : b(p) < +\infty\};$$

(ii)  $b$  is differentiable on  $(p_{min}, 1]$ ;

(iii) for all  $p \in [0, 1]$ ,  $b(p) \geq 1$ ;

(iv) for all  $p \in (p_{min}, 1]$ ,  $1 - b(p) + (1 - p)b'(p) \leq 0$ .

In particular, for each cumulative distribution function  $F$  we can define on  $[0, 1]$  the associated Pareto curve by:

$$b(p) = \begin{cases} \infty & \text{if } Q(p) = 0, \\ \frac{1}{(1-p)Q(p)} \int_p^1 Q(r) dr & \text{otherwise.} \end{cases} \quad (3.3)$$

We easily check that it satisfies conditions (i), (ii), (iii) and (iv).

The proposition below states that the reverse is true: for each Pareto curve, there is a associated cumulative distribution function which is unique up to scalar multiplication.

**Proposition 3.1.1**

*For each Pareto curve  $b : [0, 1] \rightarrow [0, +\infty]$ , there exists a cumulative distribution function  $F$  with domain in  $\mathbb{R}_+$  such that:*

$$b(p) = \begin{cases} \infty & \text{if } Q(p) = 0, \\ \frac{1}{(1-p)Q(p)} \int_p^1 Q(r)dr & \text{otherwise,} \end{cases} \quad (3.4)$$

*where  $Q = F^{-1}$  is the corresponding quantile function.*

*Such a distribution is unique up to scalar multiplication.*

PROOF:

**Unicity** Let  $b$  be a Pareto curve and  $F$  be a CDF such that (3.4) holds.

We are going to express the quantile function  $Q = F^{-1}$  with  $b$ . For any percentile  $p$ ,  $y = Q(p)$  is the corresponding income.  $p_{min}$  is defined as above.

For all  $p > p_{min}$ :

$$(1-p)Q(p)b(p) = \int_p^1 Q(r)dr$$

Differentiating, we get for  $p_{min} < p \leq 1$  :

$$(1-p)b(p)Q'(p) + [(1-p)b'(p) - b(p)]Q(p) = -Q(p)$$

and then:

$$\frac{Q'(p)}{Q(p)} = -\frac{b'(p)}{b(p)} + \frac{1}{1-p} - \frac{1}{(1-p)b(p)}.$$

We set  $p^* > p_{min}$ . For all  $p > p_{min}$ , we get by integrating:

$$\ln \left( \frac{Q(p)}{Q(p^*)} \right) = -\ln \left( \frac{b(p)}{b(p^*)} \right) - \ln \left( \frac{1-p}{1-p^*} \right) - \int_{p^*}^p \frac{1}{(1-q)b(q)} dq$$

Finally,

$$Q(p) = \begin{cases} 0 & \text{if } 0 \leq p \leq p_{min}, \\ y^* \frac{(1-p^*)b(p^*)}{(1-p)b(p)} \exp \left( - \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right) & \text{if } p > p_{min} \end{cases} \quad (3.5)$$

where  $y^* = Q(p^*)$ .

Therefore,  $Q$  is uniquely defined up to scalar multiplication, i.e. if we normalize wage mean to 1,  $Q$  is unique.

**Existence** Let  $b$  be a Pareto curve. Let  $p_{min} = \inf\{p \in [0, 1] : b(p) < +\infty\}$ . We define the function  $Q$  by:

$$Q(p) = \begin{cases} 0 & \text{if } p \leq p_{min}, \\ \frac{(1-p^*)b(p^*)}{(1-p)b(p)} \exp \left( - \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right) & \text{otherwise,} \end{cases} \quad (3.6)$$

for some  $p^* \in [0, 1)$  such that  $b(p^*) < +\infty$ . For  $p \in (p_{min}, 1]$ ,

$$\begin{aligned} ((1-p)b(p)Q(p))' &= \left( (1-p^*)b(p^*) \exp \left( - \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right) \right)' \\ &= - \frac{(1-p^*)b(p^*)}{(1-p)b(p)} \exp \left( - \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right) \\ &= -Q(p). \end{aligned}$$

As  $(1-p)b(p)Q(p)$  is equal to zero when  $p = 1$ , we find:

$$(1-p)b(p)Q(p) = \int_p^1 Q(r)dr,$$

so that (3.4) is satisfied. □

## Remarks

### 1. Continuity of $Q$ in $p_{min}$

We can check that if  $p > p_{min}$  :

$$Q(p) = y^* \frac{(1-p^*)b(p^*)}{(1-p)b(p)} \exp \left( - \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right) \xrightarrow{p \rightarrow p_{min}} 0$$

since  $b(p) \rightarrow +\infty$  when  $p \rightarrow p_{min}$ .

### 2. Special case of a Pareto distribution

Let's assume that the Pareto coefficient is constant:

$$b(p) \equiv b.$$

The quantile function writes:

$$\begin{aligned} Q(p) &= y^* \exp \left( - \int_{p^*}^p \frac{1-b}{b(1-q)} dq \right) \\ &= y^* \exp \left( \frac{1-b}{b} \ln \left( \frac{1-p}{1-p^*} \right) \right) \\ &= y^* \left( \frac{1-p}{1-p^*} \right)^{\frac{1-b}{b}} \end{aligned}$$

With the transform  $y = Q(p)$ , we get:

$$y = y^* \left( \frac{1-F(y)}{1-F(y^*)} \right)^{\frac{1-b}{b}},$$

$$\frac{1-F(y)}{1-F(y^*)} = \left( \frac{y^*}{y} \right)^{\frac{b}{b-1}}.$$

As  $F(y^*) \rightarrow F(y_{min}) = 0$  when  $y$  goes to  $y_{min}$ , we find as expected:

$$F(y) = 1 - \left( \frac{y_{min}}{y} \right)^{\frac{b}{b-1}}.$$

3. Necessity of condition (iv) in the definition of a Pareto curve

For any function continuous  $b : [0, 1] \rightarrow [0, +\infty]$  verifying (i) and (ii), we can define the function  $Q : [0, 1] \rightarrow \mathbb{R}_+$ :

$$Q(p) = \begin{cases} 0 & \text{if } p \leq p_{min}, \\ \frac{(1-p^*)b(p^*)}{(1-p)b(p)} \exp \left( - \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right) & \text{otherwise.} \end{cases} \quad (3.7)$$

For  $Q$  to be a quantile function, it has to be non-decreasing.

The quantity  $(1-p)b(p)Q(p) \exp \left( \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right)$  is constant over  $(p_{min}, 1)$ . So  $p \mapsto (1-p)b(p) \exp \left( \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right)$  has to decrease. But we have:

$$(1-p)b(p) \exp \left( \int_{p^*}^p \frac{1}{(1-q)b(q)} dq \right) = \exp \left( \int_{p^*}^p \frac{1-b(q) + (1-q)b'(q)}{(1-q)b(q)} dq \right)$$

so that the condition becomes:

$$\forall p \in (p_{min}, 1), \quad 1 - b(p) + (1-p)b'(p) \leq 0, \quad (3.8)$$

which gives (iv).

4. We can check that condition (iii) is in fact a consequence of (iv).

### 3.1.3 Lorenz curve

For all  $p \in [0, 1]$ , the share of total income accruing to the bottom  $p$  percentile of taxpayers is by definition:

$$L(p) = \frac{\int_0^{Q(p)} y f(y) dy}{\int_0^{+\infty} y f(y) dy} = \frac{\int_0^p Q(r) dr}{\int_0^1 Q(r) dr} = 1 - \frac{(1-p)Q(p)b(p)}{\bar{y}} \quad (3.9)$$

where  $\bar{y}$  is the average income of the population.

#### Definition 3.1.2

A Lorenz curve is a continuous function  $L : [0, 1] \rightarrow [0, 1]$  such that:

- (i)  $L(0) = 0$  and  $L(1) = 1$ ;
- (ii)  $L$  is increasing and convex.

For any distribution with CDF  $F$  and quantile function  $Q$ , the function  $L$  defined by (3.9) is a Lorenz curve.

Consider an income distribution with associated Pareto curve  $b$  and Lorenz curve  $L$ . We can derive the following equality from 3.5:

$$\forall p \in [0, 1], \quad b(p) = \frac{1 - L(p)}{(1 - p)L'(p)}. \quad (3.10)$$

More generally, any Lorenz curve  $L$  uniquely defines a Pareto curve  $b$  by formula (3.10).

As a result of the main proposition above, we have the following property.

**Proposition 3.1.2 (Lorenz curve)**

Let  $b : [0, 1] \rightarrow [0, +\infty]$  be a Pareto curve, and let  $Q$  be a CDF associated to  $b$ . Then there exists a unique Lorenz curve  $L$  that gives the share of total income earned by the bottom  $p$  percentile of taxpayers.

This Lorenz curve is given by:

$$L(p) = \begin{cases} 0 & \text{if } p \leq p_{\min}, \\ 1 - \frac{y^*(1-p^*)b(p^*) \exp\left(-\int_{p^*}^p \frac{1}{(1-q)b(q)} dq\right)}{\bar{y}} & \text{if } p > p_{\min}. \end{cases} \quad (3.11)$$

where  $p^* \in (p_{\min}, 1]$ ,  $y^* = Q(p^*)$  and  $\bar{y} = \int_0^1 Q(r) dr$ .

The Lorenz curve  $L$  is uniquely defined by  $b$ , and in particular it does not depend on the chosen representation of  $Q$ .

We now assume that a Lorenz curve is given. Then we can find an associated distribution that is unique up to scalar multiplication.

**Proposition 3.1.3**

Let  $L : [0, 1] \rightarrow [0, 1]$  be a Lorenz curve. Then, there exists a CDF  $F$  with domain in  $\mathbb{R}_+$  such that:

$$L(p) = \frac{\int_0^p Q(r) dr}{\int_0^1 Q(r) dr} \quad (3.12)$$

where  $Q = F^{-1}$  is the corresponding quantile function.

Such a distribution is unique up to scalar multiplication.

**Remark** The convexity of the Lorenz curve is equivalent to the fact that the CDF is increasing, and to condition (iv) for the Pareto curve.

## 3.2 Evidence using micro-files for France 2006

We now turn to the implementation the method based on the theoretical approach developed in previous section using the microdata for France 2006.

For this year, tax authorities provided internal computer files<sup>1</sup> that allow to reconstitute the distribution of incomes. Theses files, issued each year from 1988 by French administration, consist of anonymized weighted observations of tax returns. They are exhaustive for the top of the distribution.

Through this sample, we are able to see what an income Pareto curve looks like in practice. We can experiment different interpolation methods to fit its shape and set up the our new approximation technique. We can also test its robustness by comparing true values and predicted values and by comparing the true shape of the whole distribution with micro-simulations.

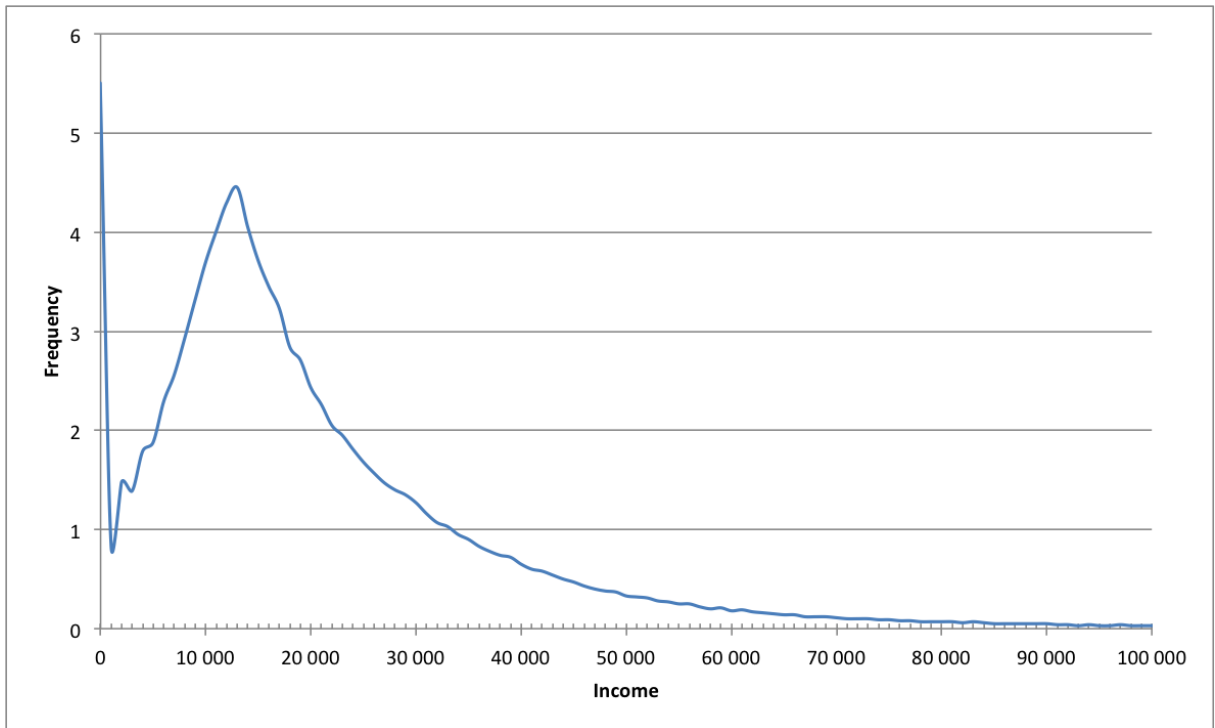


Figure 3.1: Frequency distribution of incomes, France 2006

Frequency distribution of the *revenu fiscal de référence*, in euros. Reading: 2.43% of the population earned between 20,000 and 21,000 euros in 2006. Source: Micro-files provided by tax authorities (821,815 observations).

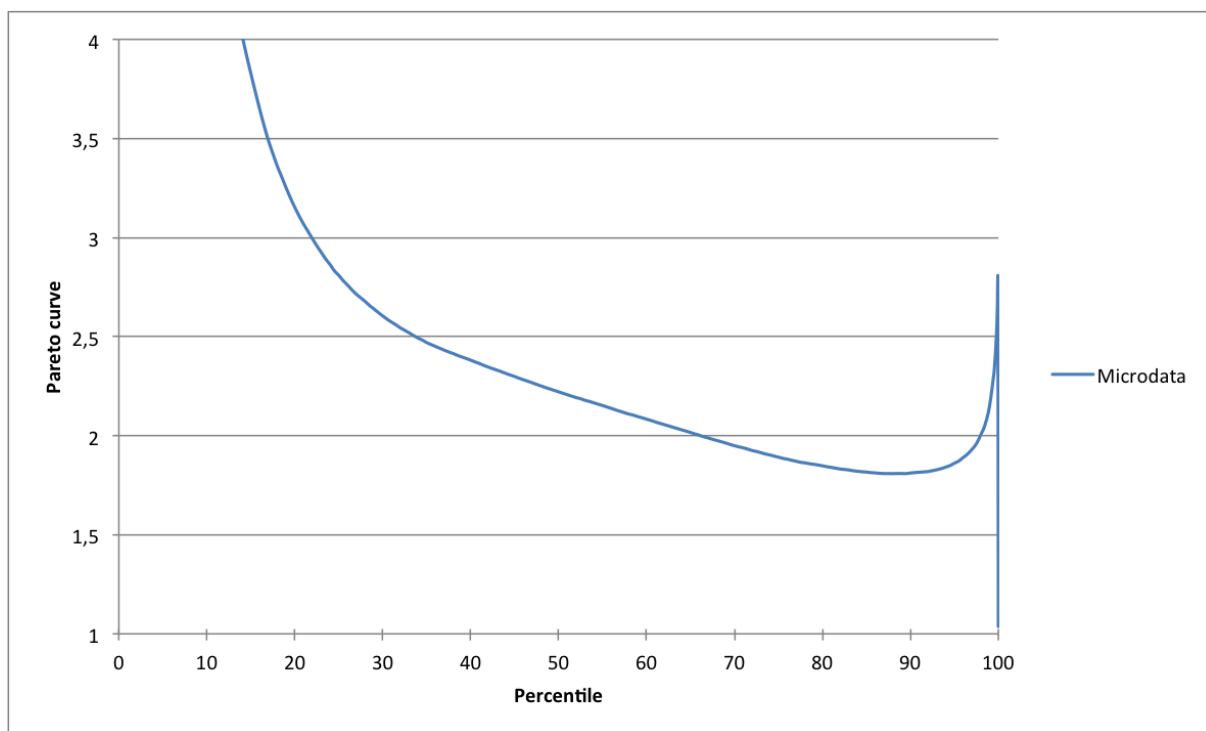
### 3.2.1 The Pareto curve

#### 3.2.1.1 Shape

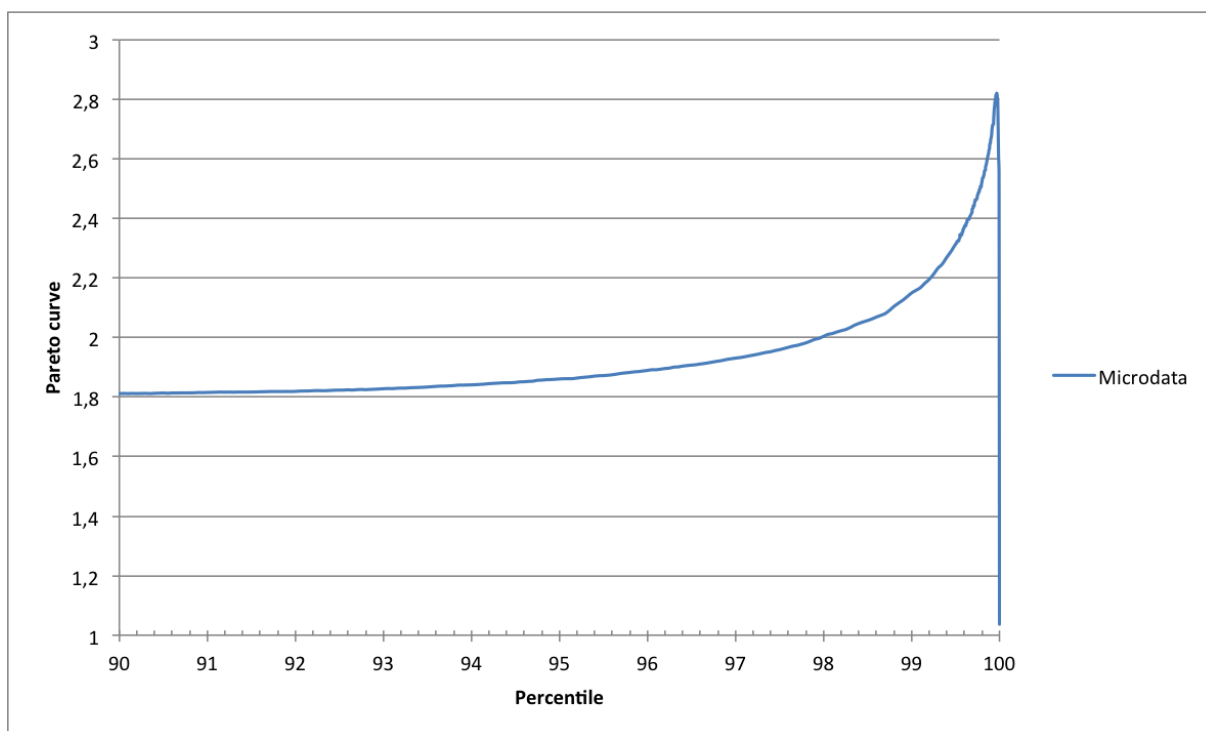
We generate the empirical Pareto curve with the data from the 2006 French micro-files (see figure 3.2a). At each point  $p$  corresponding to an income  $y$ , we compute the mean of the revenue of all individuals earning at least  $y$ , and we get the empirical value of the Pareto curve at  $p$ , namely  $b(p)$ , by dividing this mean by  $y$ .

---

<sup>1</sup>Available online at <http://www.revolution-fiscale.fr/annexes-simulateur/Fichiers/>.

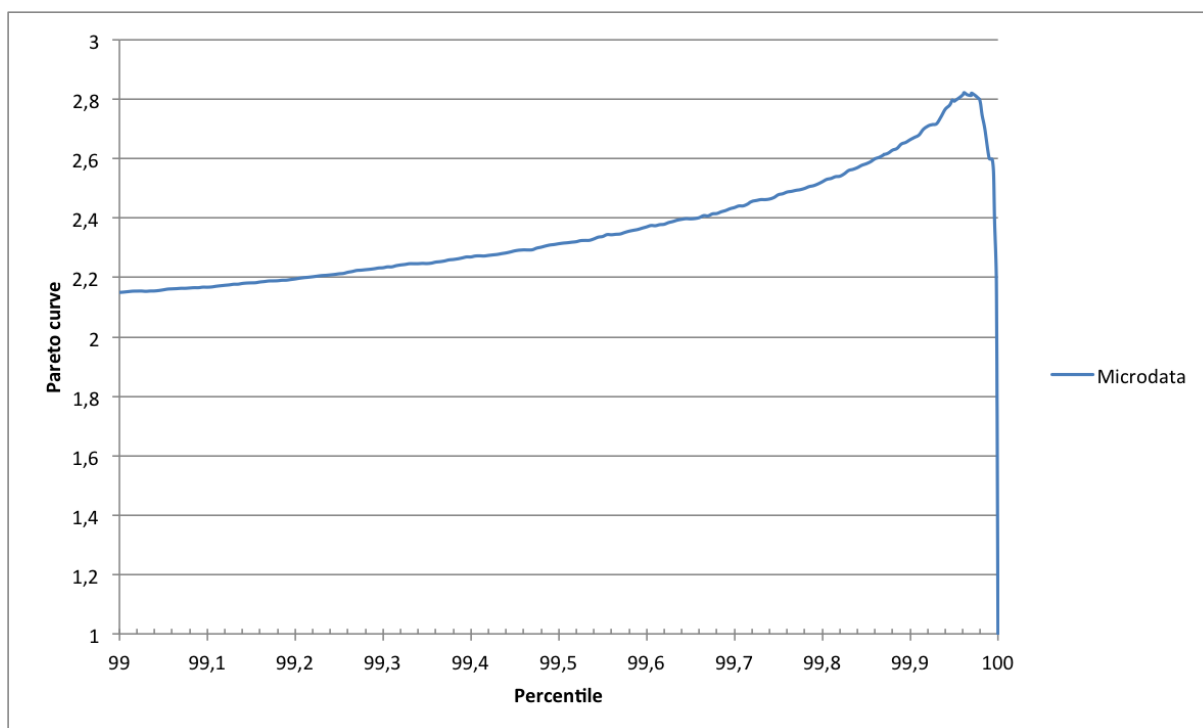


(a) Generalized Pareto curve  
Spacing: 0.5%.

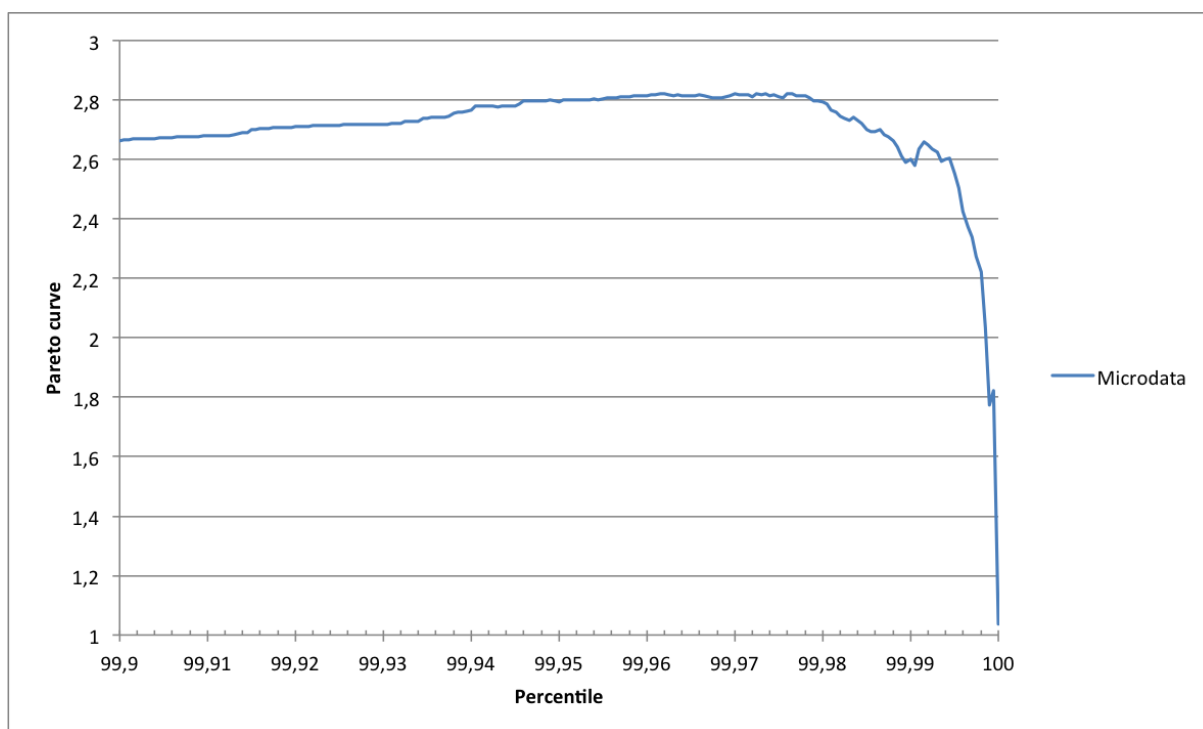


(b) Zoom on the top 10 percent  
Spacing: 0.05%.

Figure 3.2: Pareto curve, France 2006  
Source: Micro-files provided by tax authorities.



(a) Zoom on the top 1 percent  
Spacing: 0.005%.



(b) Zoom on the top 0.1 percent  
Spacing: 0.0005%.

Figure 3.3: Pareto curve, France 2006 - Zoom  
Source: Micro-files provided by tax authorities.

The graph has a vertical asymptote near the bottom 10 percents (the part before corresponds to people who have declared no income). Then, it is steadily decreasing until the top 10 percent of taxpayers. There, it is roughly stable up to percentile 95 and finally rises sharply around the top 1 percent.

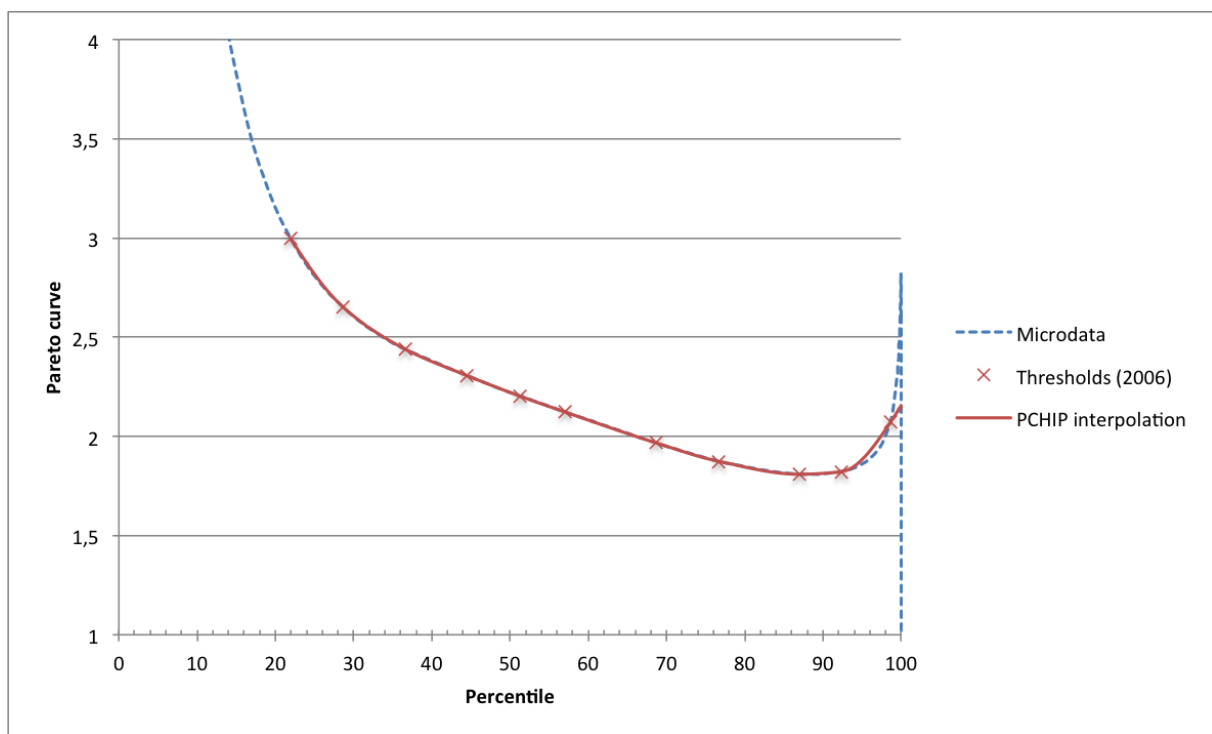
We also notice that  $b(p)$  abruptly falls to 1 within the last fractiles of the curve.

### 3.2.1.2 Approximation of the Pareto curve

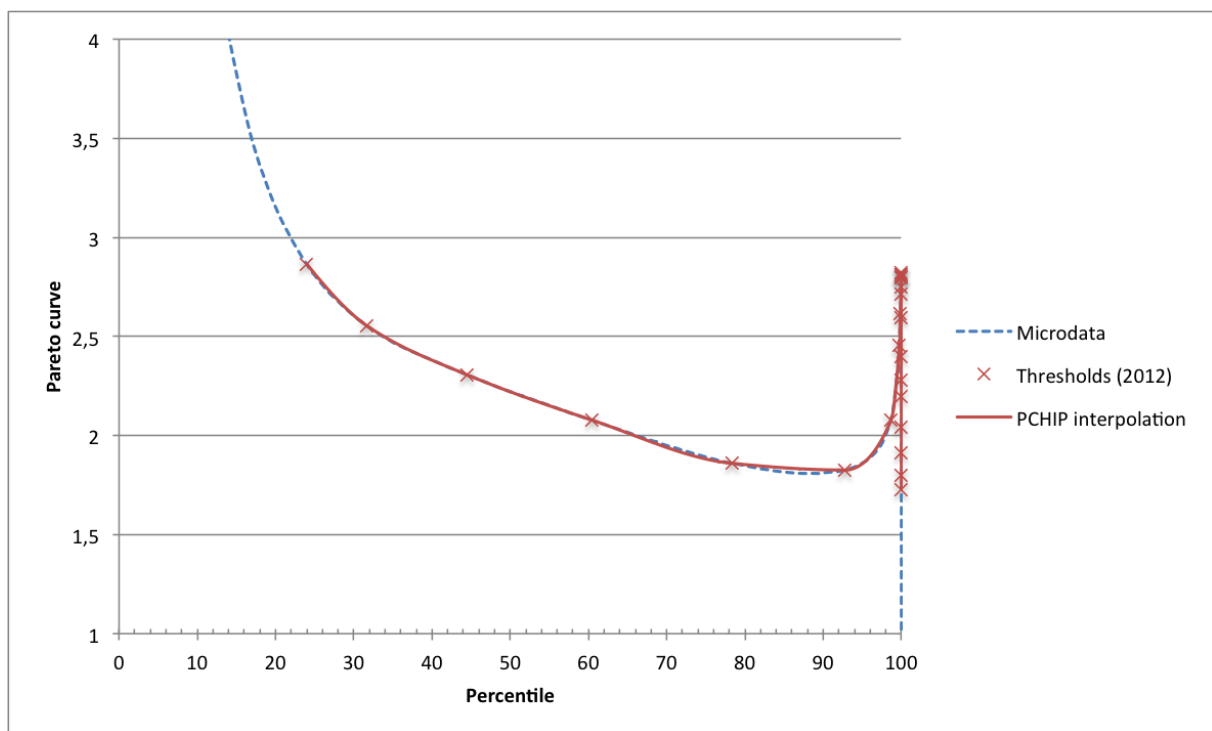
In order to apply the method described in section 3.1, we first have to approximate numerically the empirical Pareto curve from data in tax tabulations.

We have tried to approximate the Pareto curve by a suited functional form, but the resulting approximating curves appeared to be unstable. A small modification of a data point could lead to large changes in the approximating function. Another idea is to interpolate the Pareto curve. The simplest way to do so is to linearly interpolate the points of the tax scale. However, to be more realistic and to better fit to the actual form, we would like to get a smooth curve in the end. Polynomial interpolation techniques could be more satisfactory. Nevertheless, the curves resulting from such polynomial interpolations do not necessarily preserve the shape of the data and often oscillate or overshoot points on intervals where we would expect the Pareto curve to be monotonic (Runge's phenomenon is a well-known example of such a behavior). To address these shortcomings, methods of monotone cubic interpolation have been developed in numerical analysis following Fritsch and Carlson [1980]'s seminal paper. Thereafter, we will use a *Piecewise Cubic Hermite Interpolating Polynomial* (PCHIP) to interpolate our tax data. Our choice is justified extensively in appendix B, and a description of the PCHIP method is provided.

The two plots below correspond to the interpolation of the Pareto curve with the thresholds of the 2006 and 2012 tax tabulation respectively. Indeed, there are many thresholds for highest incomes in the tabulation of the year 2012, which allows to observe the behavior of our interpolant near rapid variations of the tabulated data.



(a) Thresholds of the 2006 tax scale



(b) Thresholds of the 2012 tax tabulation

Figure 3.4: Interpolation of the Pareto curve, France 2006

Method: Piecewise Cubic Hermite Interpolating Polynomial. Spacing: 0.5%. Source: Micro-files provided by tax authorities.

### 3.2.2 Asymptotic decline of the Pareto curve for finite populations

As noticed earlier, the curve declines abruptly when  $p$  is close to 1.

The question is whether this final decrease comes from the underlying distribution of incomes or from the finiteness of the population. In fact, it turns out that it is mechanically driven by the finiteness of the sample.

To see this, let  $F$  be a CDF (say, the CDF associated with a Pareto curve  $b$ ). The population of taxpayers may be seen as a large number  $N$  (roughly 35 millions) of draws  $Y_1, \dots, Y_N$  from this distribution. Assume that the  $Y_i$  are ordered. Then, the empirical Pareto curve  $\tilde{b}$  that we observe for this population is given at points  $i/N$ ,  $1 \leq i \leq N$  by the formula:

$$\tilde{b}\left(\frac{i}{N}\right) = \frac{1}{(N-i+1)Y_i} \sum_{j=i}^N Y_j. \quad (3.13)$$

If we set  $M = Y_N$  the larger income in the sample and  $\beta = \lim_{p \rightarrow 1} b(p)$  which is assumed to be strictly greater than 1, we see that:

$$\tilde{b}\left(\frac{i}{N}\right) \leq \frac{M}{Y_i}. \quad (3.14)$$

Thus, as soon as  $Y_i$  is larger than  $\frac{M}{\beta}$ ,  $\tilde{b}$  mechanically decreases to 1.

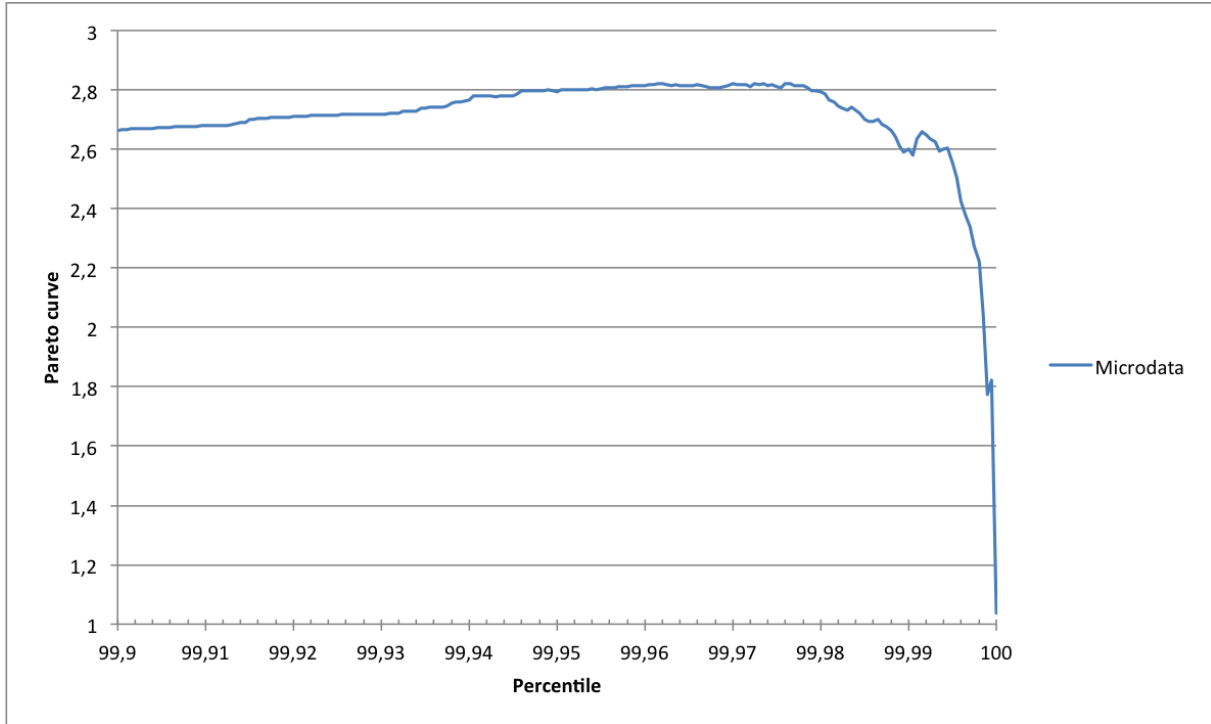


Figure 3.5: Zoom on the top 0.1 percent of the Pareto curve, France 2006

Source: Micro-files provided by tax authorities. Spacing: 0.0005%.

We simulate a population with the income distribution corresponding to the Pareto curve  $b(p)$  obtained by interpolating a mesh of the 2006 micro-file.

It turns out that the observed decrease of the empirical Pareto curve within the last percentiles

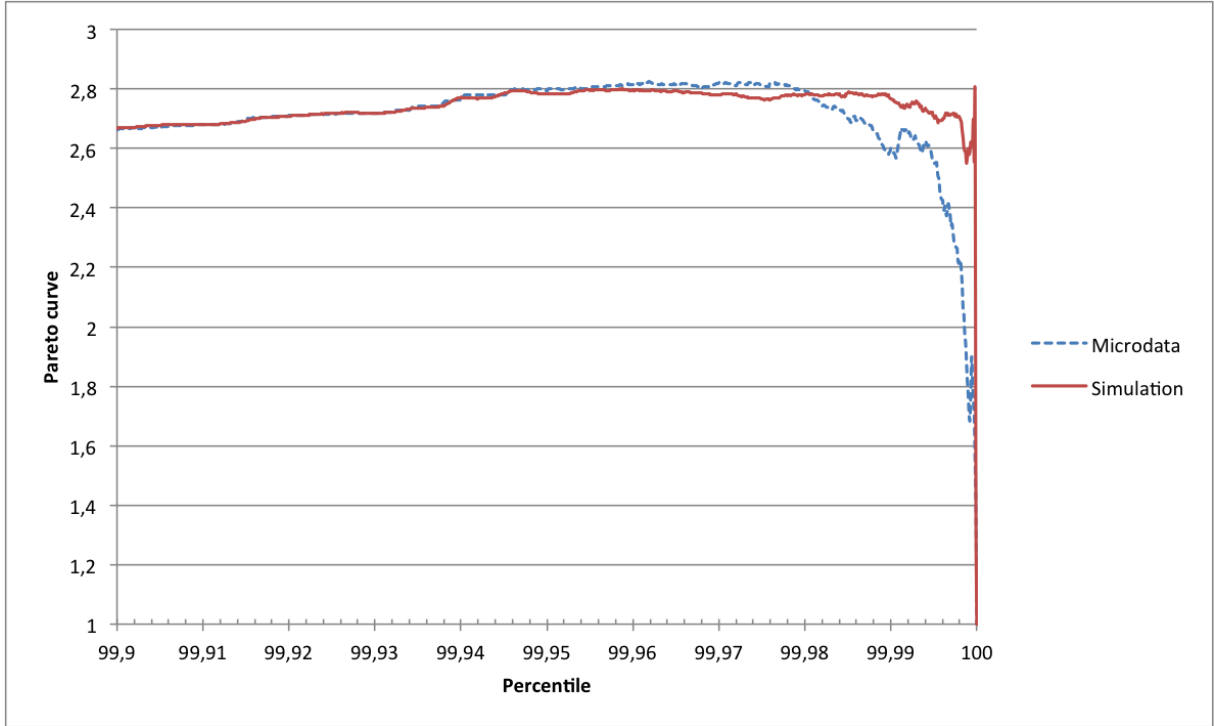


Figure 3.6: Final drop - Comparison with simulation

Source: Micro-files provided by tax authorities and simulated distribution. Spacing: 0.0001%.

of the distribution is not fully explained by a mechanical motive. Indeed, the Pareto curve of microdata starts to go down long before the Pareto curve of simulated data. As income considered here is the sum of labor income and capital income, it would be worth exploring if this empirical fact is related to the wage distribution or to the wealth distribution.

### 3.2.3 Estimations using tabulations of the income tax

We use the tabulation of 2006 in France to interpolate the Pareto curve  $b(p)$  and to predict the Lorenz curve  $L(p)$ .

We notice that the tabulation that would correspond to the micro-files is slightly different from the official table released by the tax administration. Those discrepancies would disturb the testing of the estimation method. So we will use the tabulation depicted below where the figures are computed from the micro-files to interpolate the Pareto curve.

Revenu fiscal de référence par tranche (en euros)	Nombre de foyers fiscaux	Revenu fiscal de référence des foyers fiscaux
0 à 9 400	9 576 833	42 788 853 184
9 401 à 11 250	2 392 121	24 751 590 663
11 251 à 13 150	2 543 364	31 206 544 304
13 151 à 15 000	2 693 578	37 863 149 160
15,001 à 16 900	2 337 336	37 223 221 346
16 901 à 18 750	1 920 632	34 180 761 353
18 751 à 23 750	3 576 320	75 528 414 688
23 751 à 28 750	2 776 438	72 547 037 751
28 751 à 38 750	3 458 939	114 819 782 315
38 751 à 48 750	1 778 550	76 836 621 599
48 751 à 97 500	2 086 577	133 240 253 524
Plus de 97 500	493 163	100 478 584 827
<b>Total</b>	<b>35 633 851</b>	<b>781 464 814 714</b>

Table 3.1: Income tax tabulation, France 2006  
Source: Income tax tabulation issued by the tax administration.

Revenu fiscal de référence par tranche (en euros)	Nombre de foyers fiscaux	Revenu fiscal de référence des foyers fiscaux (en millions d'euros)
0 à 9 400	7 592 031	34 490
9 401 à 11 250	2 305 258	23 880
11 251 à 13 150	2 769 638	33 850
13 151 à 15 000	2 701 325	37 980
15 001 à 16 900	2 357 246	37 500
16 901 à 18 750	1 976 702	35 200
18 751 à 23 750	4 008 824	84 500
23 751 à 28 750	2 788 768	72 800
28 751 à 38 750	3 538 988	117 500
38 751 à 48 750	1 846 420	79 700
48 751 à 97 500	2 179 706	139 100
Plus de 97 500	481 207	97 200
<b>Total</b>	<b>34 546 115</b>	<b>793 700</b>

Table 3.2: Tabulation corresponding to microdata - Tax scale: France 2006  
The thresholds are those of the tax scale in effect in 2006. Source: Micro-files provided by tax authorities.

Revenu fiscal de référence par tranche (en euros)	Nombre de foyers fiscaux	Revenu fiscal de référence des foyers fiscaux (en millions d'euros)
0 à 10 000	8 286 660	41 240
10,001 à 12 000	2 662 826	29 380
12,001 à 15 000	4 418 879	59 560
15,001 à 20 000	5 510 439	95 490
20,001 à 30 000	6 203 368	151 600
30,001 à 50 000	4 961 616	187 900
50,001 à 100 000	2 047 140	133 900
Plus de 100 000 dont :		
100 001 à 200 000	357 736	46 770
200 001 à 300 000	52 536	12 610
300 001 à 400 000	18 236	6 254
400 001 à 500 000	8 776	3 909
500 001 à 600 000	4 600	2 515
600 001 à 700 000	2 772	1 798
700 001 à 800 000	1 960	1 458
800 001 à 900 000	1 384	1 174
900 001 à 1 000 000	912	864
1 000 001 à 2 000 000	4,012	5 493
2 000 001 à 3 000 000	936	2 220
3 000 001 à 4 000 000	452	1 547
4 000 001 à 5 000 000	248	1 110
5 000 001 à 6 000 000	120	650
6 000 001 à 7 000 000	84	546
7 000 001 à 8 000 000	60	456
8 000 001 à 9 000 000	52	447
Plus de 9 000 000	310	4 769
<b>Total</b>	<b>34 546 115</b>	<b>793 700</b>

Table 3.3: Tabulation corresponding to microdata - Tax scale: France 2012  
The thresholds are those of the tax scale in effect in 2012. Source: Micro-files provided by tax authorities.

### 3.2.3.1 Method

The first step is to interpolate the empirical Pareto curve using the PCHIP method.

Then, using formula (3.5) with each of the thresholds  $(p_1, \theta_1), \dots, (p_\omega, \theta_\omega)$  found in the tax tabulation as a starting point gives us  $\omega$  quantile functions  $Q_1, \dots, Q_\omega$ :

$$Q_i(p) = \begin{cases} 0 & \text{if } 0 \leq p \leq p_{min}, \\ \theta_i \frac{(1-p_i)b(p_i)}{(1-p)b(p)} \exp\left(-\int_{p_i}^p \frac{1}{(1-q)b(q)} dq\right) & \text{if } p > p_{min}. \end{cases} \quad (3.15)$$

If we had interpolated exactly  $b(p)$  and found the true Pareto curve, we should obtain  $\omega$  identical functions. However, small discrepancies appear as the consequences of the approximative estimation of the Pareto curve. We note that  $Q_i(p_j)$  is not exactly equal to  $\theta_j$  when  $j \neq i$ , even if the two values are very close.

We could select arbitrarily one starting point  $p_i$  and draw all our estimations from the associated quantile function  $Q_i$ . The problem is that if some other observation  $p_j$  coincides with a point of interest such as  $p = 99.5\%$ , we waste valuable information provided by the administrative tabulations.

As for each point of the distribution, the two best estimations are obtained when starting from the consecutive thresholds that are bracketing it, we approximate the underlying quantile function  $Q$  with a weighted average of these two points. Formally, if  $p$  lies between  $p_i$  and  $p_{i+1}$ , we set:

$$Q(p) = \frac{p_{i+1} - p}{p_{i+1} - p_i} Q_i(p) + \frac{p - p_i}{p_{i+1} - p_i} Q_{i+1}(p). \quad (3.16)$$

To sum up, for each threshold  $\theta_i$  we use the approximation of the Pareto curve  $b(p)$  to estimate  $Q_i(p)$  which passes exactly through the point  $(p_i, \theta_i)$ . The  $Q_i$  are the same up to a scalar multiplication and they all have  $b(p)$  as a Pareto curve. But  $Q$ , which is a combination of the  $Q_i$  does not have exactly the same Pareto curve.

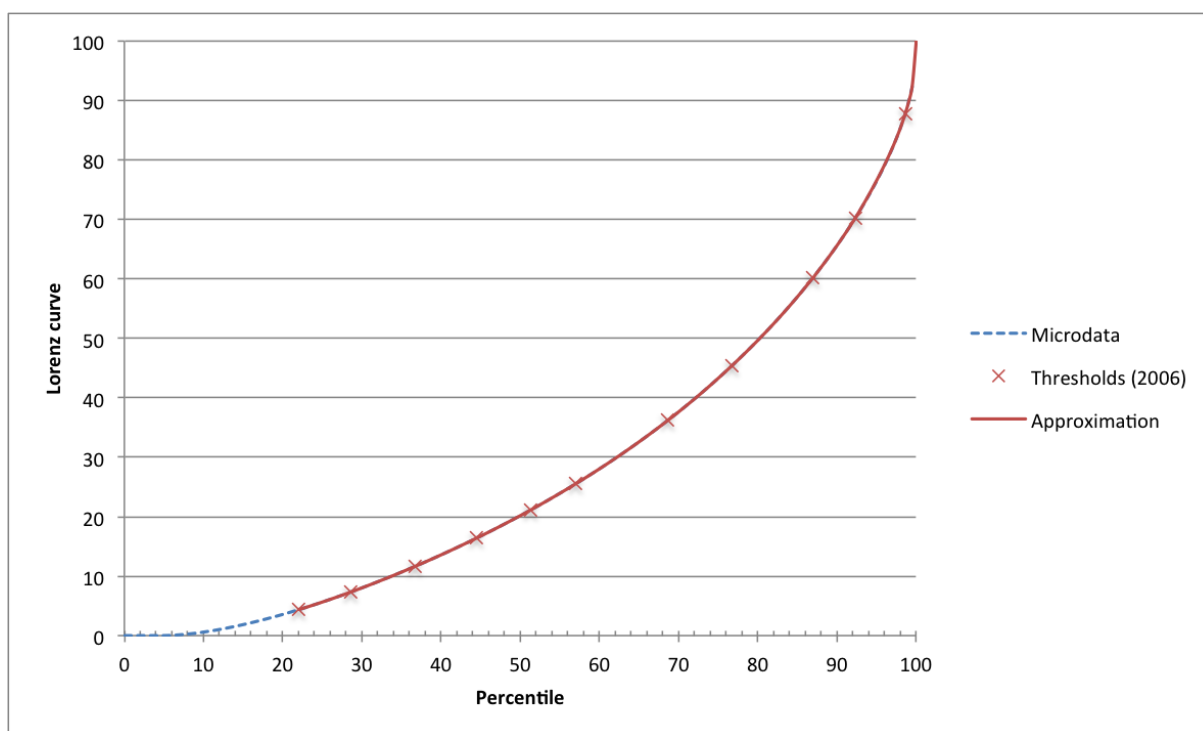
### 3.2.3.2 Estimation of declared taxable income

We display three tables giving thresholds, average incomes and shares accruing to different deciles and percentiles of the population. The column "Microdata" indicates their "true" value, namely their value in the micro-file distribution. The columns "2006 tax scale" and "2012 tax scale" provide the estimations obtained by applying our method with the thresholds of the tax scales in effect in 2006 and 2012 respectively.

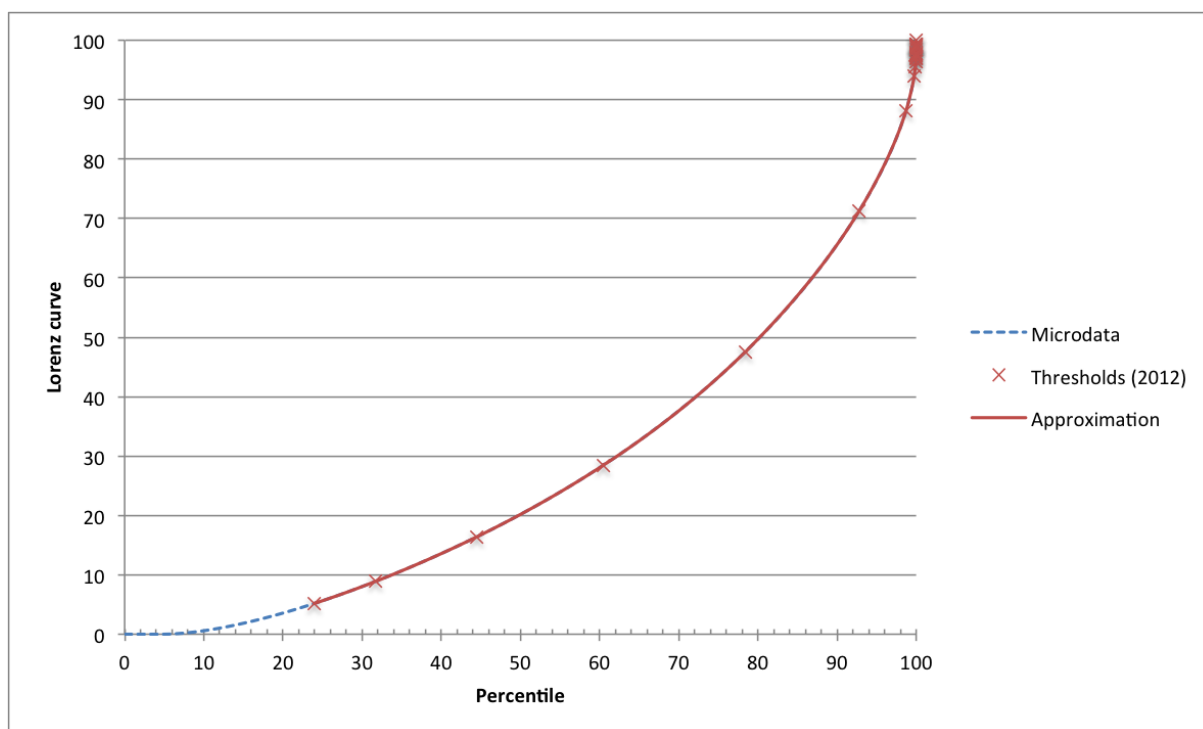
Approximations appear to be especially accurate. The only discrepancies emerge for the very top of the distribution (99.9% and 99.99%) when the tax scale of 2006 is used: thresholds and average income are underestimated. This is due to the fact that the extrapolation of the extrapolated Pareto curve is itself underestimated for top percentiles in this case.

Graphs 3.7 represent the Lorenz curves approximated using our method with the 2006 and the 2012 tax scales. In both cases, the approximations perfectly fit the microdata Lorenz curve.

We notice that the approximations of thresholds are typically less precise than approximations of average income or shares.



(a) Thresholds of the 2006 tax tabulation



(b) Thresholds of the 2012 tax tabulation

Figure 3.7: Approximation of the Lorenz curve, France 2006  
Spacing: 0.5%. Source: Micro-files provided by tax authorities.

Percentile	Microdata	2006 tax scale	2012 tax scale
30%	11,589	11,582	11,584
40%	13,895	13,920	13,903
50%	16,511	16,512	16,515
60%	19,834	19,848	19,831
70%	24,498	24,487	24,628
80%	31,280	31,324	31,173
90%	43,584	43,544	43,214
95%	58,110	57,166	58,117
99%	111,784	114,531	112,421
99.5%	152,161	162,611	150,883
99.9%	346,354	378,884	346,465
99.99%	1,535,357	1,296,334	1,501,065

Table 3.4: Thresholds corresponding to different deciles and percentiles of the population  
The table compares the values corresponding to microdata, and the values predicted by applying our method with the 2006 and 2012 tax scales. Source: Micro-files provided by tax authorities.

Percentile	Microdata	2006 tax scale	2012 tax scale
30%	30,186	30,186	30,186
40%	33,090	33,090	33,090
50%	36,676	36,676	36,676
60%	41,322	41,323	41,322
70%	47,756	47,757	47,770
80%	57,804	57,810	57,777
90%	78,962	78,963	78,968
95%	108,141	107,681	107,952
99%	240,287	239,853	240,017
99.5%	352,046	345,437	351,644
99.9%	922,673	813,969	922,603
99.99%	3,993,587	2,791,930	3,994,573

Table 3.5: Average income above different deciles and percentiles of the population  
The table compares the values corresponding to microdata, and the values predicted by applying our method with the 2006 and 2012 tax scales. Source: Micro-files provided by tax authorities.

Percentile	Microdata	2006 tax scale	2012 tax scale
30%	8.03	8.03	8.03
40%	13.58	13.58	13.58
50%	20.18	20.18	20.18
60%	28.05	28.05	28.05
70%	37.64	37.64	37.62
80%	49.68	49.67	49.70
90%	65.63	65.63	65.63
95%	76.46	76.56	76.51
99%	89.54	89.56	89.55
99.5%	92.34	92.48	92.35
99.9%	95.98	96.46	95.98
99.99%	98.26	98.78	98.26

Table 3.6: Values taken by the Lorenz curve at different deciles and percentiles of the population  
The table compares the values corresponding to microdata, and the values predicted by applying our method with the 2006 and 2012 tax scales. Source: Micro-files provided by tax authorities.

### 3.2.3.3 Comparison with the old method

Figures 3.8 depict the ratios of estimates found with our method and the piecewise Pareto method developed by Piketty [2001] over true microdata values for various deciles and percentiles of the distribution. As the mesh of the 2006 and 2012 tax scales is narrow, the two methods provide good results: predicted values are close to true values. The relative good performance of the old method stems from the fact that the tax scale mesh is narrow. However, the estimations obtained with the new method are always better than the others. The difference is more striking in the case of thresholds. Indeed, thresholds estimations are generally more sensitive to errors in approximating  $b(p)$  than average income estimations.

Threshold corresponding to percentile  $p$  is given by the quantile function:

$$Q(p) = \theta_i \frac{(1-p_i)b_i}{(1-p)b(p)} \exp \left( - \int_{p_i}^p \frac{1}{(1-q)b(q)} dq \right) \quad (3.17)$$

for a threshold  $i$  of the tax scale. This is the estimate obtained with our nonparametric method, where  $b(p)$  is the interpolation of tabulation points.

Piecewise Pareto method (PP) assumes that the Pareto coefficient is locally constant. Analytically, the estimator of the threshold writes:

$$Q^{\text{PP}}(p) = \theta_i \frac{(1-p_i)}{(1-p)} \exp \left( - \int_{p_i}^p \frac{1}{(1-q)b_i} dq \right) \quad (3.18)$$

$$= \theta_i \left( \frac{1-p_i}{1-p} \right)^{(b_i-1)/b_i}. \quad (3.19)$$

The ratio of these two estimators can be expressed as:

$$\frac{Q^{\text{PP}}(p)}{Q(p)} = \frac{b(p)}{b_i} \exp \left( \left( \int_{p_i}^p \left( \frac{1}{b(q)} - \frac{1}{b_i} \right) \frac{dq}{1-q} \right) \right). \quad (3.20)$$

Similarly, the estimator for the new method of average income above percentile  $p$  is:

$$A(p) = b(p)Q(p) \quad (3.21)$$

$$= b_i \theta_i \frac{(1-p_i)}{(1-p)} \exp \left( - \int_{p_i}^p \frac{1}{(1-q)b(q)} dq \right), \quad (3.22)$$

and the estimator of the Piecewise Pareto method is:

$$A^{\text{PP}}(p) = b_i Q^{\text{PP}}(p) \quad (3.23)$$

$$= b_i \theta_i \frac{(1-p_i)}{(1-p)} \exp \left( - \int_{p_i}^p \frac{1}{(1-q)b_i} dq \right) \quad (3.24)$$

$$= b_i \theta_i \left( \frac{1-p_i}{1-p} \right)^{(b_i-1)/b_i}. \quad (3.25)$$

Their ratio is:

$$\frac{A^{\text{PP}}(p)}{A(p)} = \exp \left( \int_{p_i}^p \left( \frac{1}{b(q)} - \frac{1}{b_i} \right) \frac{dq}{1-q} \right). \quad (3.26)$$

Consequently, the estimation errors in the case of the Piecewise Pareto method come from two factors:

1. the ratio  $\frac{b(p)}{b_i}$ ;
2. the exponential term  $\exp \left( \int_{p_i}^p \left( \frac{1}{b(p)} - \frac{1}{b_i} \right) \frac{dq}{1-q} \right)$ .

If we denote  $\delta = \frac{b_i - b(p)}{b(p)}$  the relative error in approximating  $b$ , we have:

$$\frac{b(p)}{b_i} = \frac{1}{1 + \delta} \simeq 1 - \delta, \quad (3.27)$$

and

$$\exp \left( \int_{p_i}^p \left( \frac{1}{b(p)} - \frac{1}{b_i} \right) \frac{dq}{1-q} \right) \simeq \left( \frac{1-p}{1-p_i} \right)^{\delta/b_i} \simeq 1 + \frac{p_i - p}{b_i(1-p_i)} \delta \quad (3.28)$$

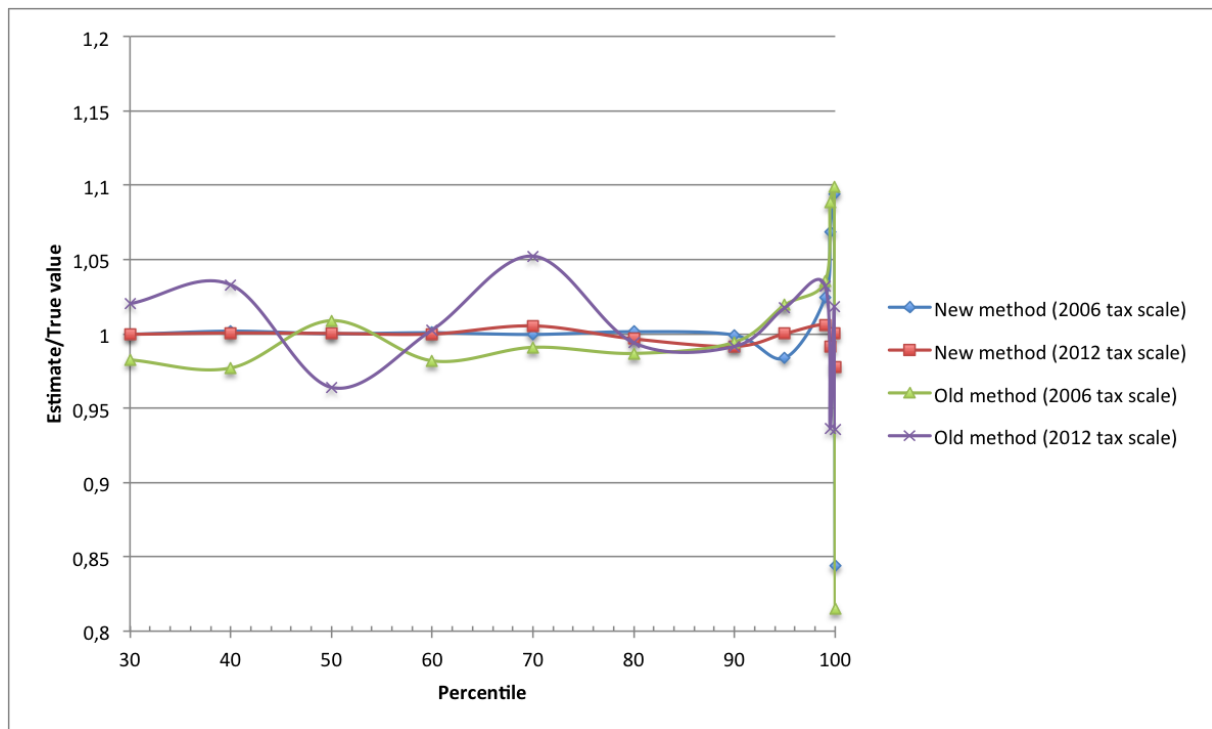
at the first order.

Therefore, when  $p \simeq p_i$ , the first term dominates.

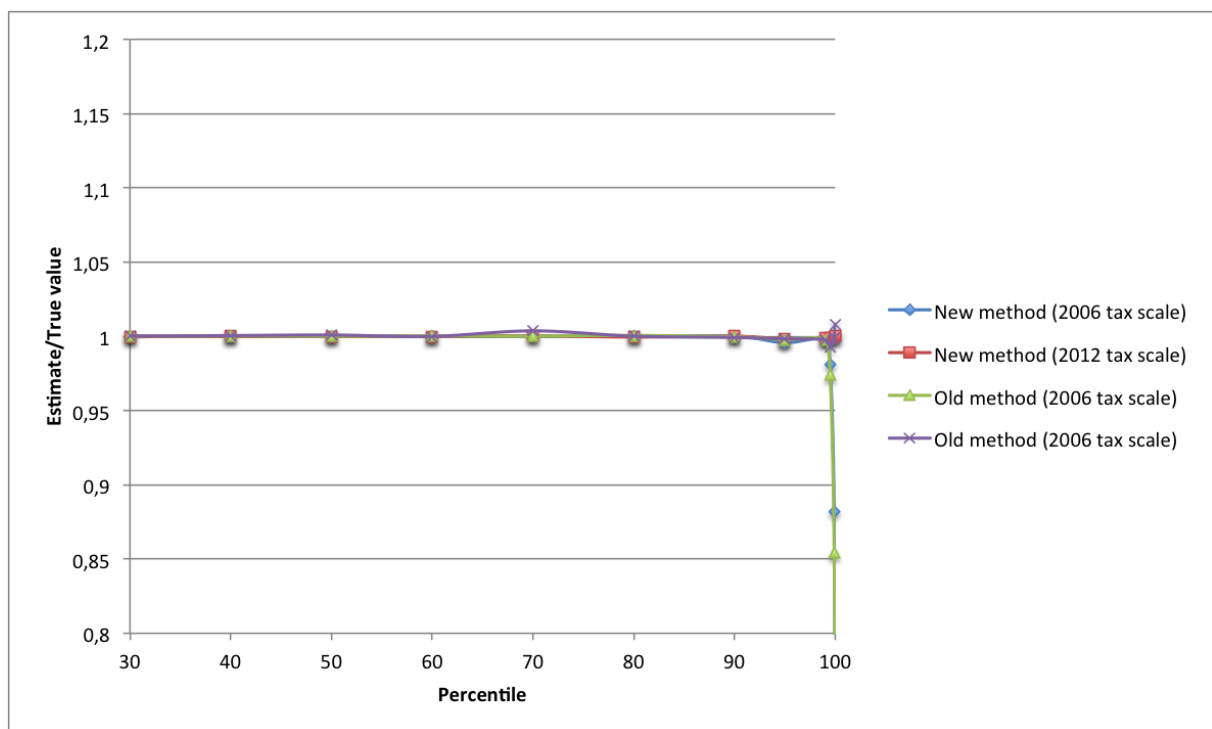
In the case of average income estimators, the errors only spring from the second factor.

That is why the observed discrepancies are generally larger for threshold estimates than for average income estimates.

For very top incomes, that is when  $p \simeq 1$ , the multiplicative factor  $\frac{p_i - p}{b_i(1-p_i)}$  can still be very high, due to the term  $\frac{1}{1-p_i}$ . Then, the second factor dominates. Gaps can be very large for top incomes both for threshold and average income estimations. Therefore, we must take great care to extrapolate accurately the Pareto curve for top incomes.



(a) Thresholds



(b) Average income

Figure 3.8: Ratios of estimated values of true values for different deciles and percentiles of the population

Source: Micro-files provided by tax authorities.

## Section 4

# Application to French income and inheritance tax tabulations 1901-2012

### 4.1 Application to income tabulations for France 1915-2012

In this section, we first recall the evolution of the income tax legislation in France, as described by Piketty [2001]. We will focus on the measures that affect our methodology to estimate shares and the precision of our results: exemptions, changes in the number of tax brackets, modifications of the general allowance, measures to take into account the family situation... Second, we will detail the corrections that we have performed to offset these various deductions. Ultimately, we will compare our estimations to the estimations found in [Piketty, 2001].

#### 4.1.1 The income tax in France

##### 4.1.1.1 The system of the "quatre vieilles" (1792-1914)

The income tax was introduced in France by the law of 15 July 1914. It broke with the taxation system dating back to the French Revolution. This tax scheme, which applied from 1792 to 1914, was composed of four direct taxes known as the *contributions directes* or "*quatre vieilles*". Crucially, they were never depending directly on the revenues of the taxpayer. This index-based taxation system relied on indications of the contributory capacity and not on the income itself which was never reported to the tax administration. The *contribution des portes et fenêtres* was a property tax based on the number of doors and windows in a house. The *contribution foncière* on all developed and not developed properties (houses, buildings, lands, forests...) and the *contribution personnelle-mobilière* on the principal residence were based on the rental value of the estate. The *contribution des patentes* was paid by all the merchants, craftsmen and manufacturers and was calculated on a scale established for each profession as a function of the size of their business (and not on profits). Unlike modern fiscal systems, there was no tax rate applicable to a fixed base. Instead, the government set each year the amount to be levied which was somehow apportioned between taxpayers.

Resulting both from the imperfect correspondance between individual incomes and the in-

dication used to compute the amount of the tax to be paid and from the random variations of effective tax rates from region to region, we can not use the data released at that time by tax authorities to infer any information about the distribution of incomes. Furthermore, effective tax rates remained very low (about 1 or 2%), so that these taxes did not affect the wealth accumulation process (all the more so that there was virtually no inflation during this period).

This system was enforced until the law of 31 July 1917 setting up the schedular tax system (*impôt cédulaires*), when the portion of revenues due to the government was repealed. The share due to the department and municipalities was preserved. The ruling of 7 January 1959 modified these taxes official names.

#### **4.1.1.2 The IRVM (law of 29 June 1872)**

A first institutional break with the legacy system occurred with the implementation of the *impôt sur le revenu des valeurs mobilières* (IRVM) by the law of 29 June 1872. This tax on income from securities had an extensive (and fixed) tax base as it affected all revenues from securities with a fixed rate (initially equal to 3%, and then to 4%). It was withheld at source.

#### **4.1.1.3 The progressive inheritance tax (law of 25 February 1901)**

Until 1901, the inheritance tax in France was fully proportional. It was created in 1799. The rate depended on the degree of relationship to the deceased. While instituting the progressive inheritance tax, the law of 25 February 1901 implemented the first nationwide progressive tax in France. From then onwards, the tax authorities started publishing statistical tables corresponding to the tax brackets. This allows to study the evolution of large inheritances throughout the XX<sup>th</sup> century. Here again, the tax rates remained very low.

#### **4.1.1.4 The income tax (law of 15 July 1914)**

The law adopted on 15 July 1914 to institute the income tax was based on the draft initially submitted in 1907 by the French Minister of Finance from the radical party Joseph Caillaux. The Chamber of Deputies passed the bill on 9 March 1909 but it was thereafter obstructed by the Senate. The law was at last approved by the Senate in 1914, while international tensions and the pressing needs for national defense imposed an additional financial burden on the country. The income tax (*impôt général sur le revenu* or IGR) became effective on 1 January 1916, affecting the revenues of 1915.

Taxpayers had to declare their revenues to the fiscal administration. More specifically, each "family head" had to report his income and the income of all his dependants. This notion of taxpayer is close to the definition of taxable household (*foyer fiscal*) which is still used today in France.

The IGR was designed as a progressive tax targeting a small minority of affluent households. Its scale was defined in terms of marginal tax rates.

Just as for the progressive inheritance tax, statistical tables were released by the fiscal administration. However, in the early years, these tables gave only information on a very tiny

number of top-earning households. When the law was first applied, the revenues of only 260 000 households (1.7% of the population at the time) were taxable.

To take into account the family situation of the taxpayers, a dual system was set up. Flat-rate deductions from the taxable income and proportional tax cuts were specified, both depending on the family size. Flat-rate deductions amounted to 2 000 francs for married couples, 1 000 francs per dependent child up to the 5<sup>th</sup>, and then 1 500 francs for each additional child. These deductions resulted de facto in a raise in the taxation threshold. After taking into account these abatements, the amount due calculated with the tax scale was reduced by 5% if the taxpayer had one dependant, 10% if he had two, and then 10% more for each additional dependant until the 6<sup>th</sup>.

Another measure of the law of 15 July 1914 provided that taxpayers could deduce from their taxable income subject to the IGR the total amount of all direct taxes paid on revenues of the previous year. These direct taxes included the IGR itself, the *quatre vieilles* until 1917 and then the schedular taxes. At first, these exemptions were negligible. But as soon as the income tax reached higher levels, these deductions became substantial and this measure induced artificial fluctuations in the amount due by each taxpayer. Although the Popular Front which ruled France from 1936 to 1938 intended to remove this measure, it remained until the Liberation.

Besides, the system of the *quatre vieilles* was maintain until 1917. It was only the law of 31 July 1917 that substituted it with the system of schedular taxes that applied from revenues of 1917. The *contribution foncière* was replaced with a schedular tax on land incomes. The IRVM served as a schedular tax on securities. The law created four other schedular taxes: a wage tax (*impôt sur les traitements, salaires, pensions et rentes viagères* or *impôt sur les salaires*) on all revenues of wage earners, a tax on all manufacturing and commercial earnings (*impôt sur les bénéfices industriels et commerciaux* or BIC), a tax on agricultural profits (*impôt sur les bénéfices agricoles* or BA) and a tax on non-commercial profits (*impôt sur les bénéfices non commerciaux* or BNC) that applied to mixed earnings of self-employed workers. Contrary to the IGR, the tax unit was the individual. While the IGR was intended as a progressive tax targeting a minority of wealthy taxpayers, the schedular system was designed to tax a large number of taxpayers at almost proportional rates. All revenues were to be taxed by this exhaustive system. A notable exception was the interests on public debt, which were spared by the schedular taxes (but were taxed by the IGR).

#### **4.1.1.5 The French income tax between 1915 and 1944**

Let us now overview the chronology of the income tax from 1915 to 1944. During this first half of the XX<sup>th</sup> century, the income tax experienced dramatic changes, both institutional and quantitative.

	Number of thresholds	Share of taxable households
1915	9	1.7
1916	12	3.1
1917	7	3.9
1918	7	4.6
1919	10	3.6
1920	10	6.5
1921	10	7.3
1922	10	6.6
1923	10	7.7
1924	10	9.4
1925	10	12.1
1926	10	16.0
1927	10	17.9
1928	9	12.1
1929	9	11.7
1930	9	13.0
1931	9	12.4
1932	9	11.5
1933	9	11.4
1934	9	10.4
1935	9	9.7
1936	11	9.7
1937	11	13.5
1938	11	16.5
1939	11	13.0
1940	11	11.6
1941	13	17.8
1942	25	25.0
1943	24	13.4
1944	24	18.4

Table 4.1: Share of taxable households and number of tax brackets by year, France 1915-1944  
Source: [Piketty, 2001].

**Revenues of 1915-1916** After its first application for revenues of 1915, a new scale with additional thresholds was instituted for incomes of 1916 (law of 30 December 1916). The general allowance was lowered to 3 000 francs and marginal rates up to 10% were applied to incomes falling into the new brackets.

**Revenues of 1917-1918** Then, the law of 29 June 1918 established a new scale with fewer thresholds. This time the scale was defined in terms of average rates (and not of marginal rates), which actually raised effective tax rates. The maximal rate reached 20%. This new system applied to incomes earned in 1917 and 1918.

**Revenues of 1919-1935** A major quantitative break happened with the law of 25 June 1920. A new scale, formulated in terms of marginal rates, allowed for rates up to 50%. This reform was voted by the right-wing majority of the *Bloc national* to respond to the difficult financial situation following WWI. As the number of affected taxpayers was extremely low, this measure was primarily symbolic. The rates of schedular taxes was also increased, but remained moderate. As part of a pro-birth policy, a system to penalize all single or married for more than two years taxpayers without children was introduced. The surcharge was respectively equal to 25% or 10% of the amount due. This disposition, which was toughened in 1934 and then tempered in 1936, applied for revenues of years 1919 to 1938. It turned into a family compensation tax (*Taxe de compensation familiale* or TCF) during WWII.

The period from 1920 to 1936 (corresponding to the taxation of revenues from 1919 to 1935) was characterized by considerable and continual fluctuations in tax rates applicable to highest incomes. But the frame remained the law of 25 June 1920. Indeed, this law had defined separately the overall structure of tax brackets and the level of rates that applied. The amount due was calculated as follows:  $0/25^{\text{th}}$  of the fraction of income between 0 and 6 000 francs was kept,  $1/25^{\text{th}}$  between 6 000 and 20 000 francs,  $2/25^{\text{th}}$  between 20 000 and 40 000 francs... until  $25/25^{\text{th}}$  (that is, all) the fraction of income above 550 000 francs. Then the global rate (initially 50%) was applied to the total. Throughout this period, the only changes were the variations of the global rate, and the reassessments of the general allowance (which was equal to 6 000 for the taxation of incomes of years 1919-1921, 7 000 for incomes of years 1922-1927, 10 000 for incomes of years 1928-1935). The unique tax rate was maintained at 50% from the taxation of 1919 to the taxation of 1922. After various exceptional increases, the upper marginal rate exceeded 90% for the taxation of incomes of 1924. The President of the Council Raymond Poincaré brought down the global rate of the IGR to 30% for incomes of 1926 (decrease that was offset by other tax hikes).

André Tardieu, President of Council in 1929-1930, decided to raise the general allowance and to institute new deductions for dependants (finance laws of 31 December 1928 and of 29 December 1929). He also slightly reassessed the global rate of the IGR. This "prosperity policy" ended when the global economic crisis hit France in 1930. The government of Édouard Daladier set up an exceptional increase of 10% (law of 28 February 1933) that applied for revenues of 1932 and 1933.

After coming into office, Gaston Doumergue abrogated the exceptional surcharge of 10% and decreased the rate of the IGR from 33.33% to 24% (law of 6 July 1934 and decree of 20 July 1934, effective for the revenues of 1933). These measures applied from the taxation of revenues of 1934. In return, the government took measures to extend the mass of taxable incomes. An abatement of 10% for occupational expenses was instituted to control excessive deductions. This abatement still applies. The law also repealed the tax cuts for dependants while it raised significantly the level of flat-rate deductions for dependants and reassessed substantially the existing surcharges for taxpayers without children.

In 1935, Pierre Laval decided of exceptional increments for top-earning taxpayers by the decree-laws of 16 July and 26 July.

**Revenues of 1936-1941** The Popular Front, which ruled France from 1936 to 1938, undertook a major reform of the income tax with the finance law of 31 December 1936. It applied from revenues of 1936. A new tax scale was established, formulated in average rates. A maximal effective tax rate of 30% was created to limit the effects on top-earning taxpayers. The deductions for dependants, which until 1936 were flat-rate and did not depend on the income of the taxpaying households, were reduced for the more affluent families. Surcharges for taxpayers without children were softened.

This legislation was not undermined by the Vichy regime and applied from 1936 to 1941. The only changes were the establishment of successive exceptional surcharges.

**Revenues of 1942-1944** The progressiveness of the tax scale was even strengthened by the law of 24 October 1942. The new scale of the IGR applied for the taxation of revenues from 1942 to 1944. It was defined in marginal rates, which induced a slight decrease of the tax burden, but the upper marginal rate reached 70%. As in the law of 25 June 1920, the overall structure of tax brackets and the level of rates that applied were defined separately. The law of 30 January 1944 rose the general allowance, initially equal to 10 000 francs, up to 20 000 francs for the revenues of 1943. As part of a pro-birth policy, the reduction of deductions for dependants decided by the Popular Front was repealed (law of 13 January 1941). The family situation was also taken into account with the family compensation tax (*Taxe de compensation familiale* or TCF) which applied for revenues from 1939 to 1944. Indeed, Daladier government had removed the IGR surcharges for taxpayers without children, and replaced them by an equivalent tax, the TCF (decree-law of 29 July 1939).

#### **4.1.1.6 The income tax since 1945**

Since 1945, the evolution of the income tax in France has been much smoother. The decisions taken at the Liberation essentially determined the frame of the progressive income tax up to the present day.

	Number of thresholds	Share of taxable households
1945	10	10.2
1946	10	25.1
1947	9	8.9
1948	10	16.0
1949	9	20.1
1950	8	17.5
1951	8	14.8
1952	8	19.5
1953	8	17.8
1954	8	18.0
1955	8	21.3
1956	8	24.7
1957	11	24.6
1958	11	27.4
1959	12	27.4
1960	11	29.3
1961	13	32.5
1962	13	35.5
1963	13	39.5
1964	13	42.2
1965	12	42.8
1966	13	44.4
1967	13	47.2
1968	13	51.2
1969	13	50.7
1970	9	50.0
1971	10	51.6
1972	10	53.1
1973	10	55.2
1974	10	57.6
1975	10	60.3
1976	10	63.3
1977	10	61.7
1978	10	63.5
1979	5	64.7
1980	5	65.2
1981	12	63.4
1982	13	63.7
1983	13	62.8
1984	12	61.9
1985	12	60.7
1986	12	52.1
1987	12	50.8
1988	12	50.3
1989	12	50.7
1990	12	51.0
1991	12	51.2
1992	12	50.8
1993	12	50.4
1994	12	49.9
1995	12	50.6
1996	12	48.8
1997	12	49.7
1998	12	52.7

Table 4.2: Share of taxable households and number of tax brackets by year, France 1945-1998  
Source: [Piketty, 2001].

**Finance law of 31 December 1945** The finance law of 31 December 1945 set the ingredients of the modern income tax. The structure of the tax scale and the family quotient survived throughout the second half of the XX<sup>th</sup> century. It was also decided to remove the possibility of deducting from taxable income the tax paid the previous year.

First, the structure of the tax scale stabilized after 1945. It is defined in marginal rates, and the number of tax brackets (5 for revenues between 1945 and 1948, 8 between 1949 and 1972, 12 between 1974 and 7 between 1993 and 1998) as well as the corresponding levels of taxation has not changed much. The upper marginal rate steadied around 55-65%. The only fluctuations were due to a few provisional exceptional increases of the tax rates.

**Family quotient** The family quotient system was set up in 1945 to replace the system of flat-rate deductions for dependants that had applied for revenues of years 1915-1944. Each household is assigned a number according to its family situation. The tax scale is applied to the income of the household divided by the number of tax shares, and finally the amount to be paid is multiplied by the number of shares.

Singles had one tax share, married couples two, and each additional child gave an additional half-share. This system reduces the average tax rate which actually applies owing to the progressivity of the income tax. Raymond Barre granted an extra half-share to large families for the fifth dependent child for the taxation of the revenues of the year 1979 (law of 18 January 1980), and another extra half-share for the third dependent child as from the taxation of 1980 revenues. The government led by Jacques Chirac granted an extra half-share for each child from the third (law of 20 December 1986). But, as a legacy of the TCF, married couples without dependent child after three years of marriage had only one share and a half (and not two). This disposition applied to incomes of the years 1945-1949 and was repealed by the law of 24 May 1951. Another point is that single taxpayers were given one share and a half (and not one) if at least one of their children had reached 16. This family quotient system was opened to civil partners when the PACS was adopted in October 1999.

The family quotient system is still applied in France. In 1981, the socialist government led by Pierre Mauroy reformed the family quotient system (law of 30 December 1981). As from the revenues of the year 1981, a mechanism ensured that deductions provided by the family quotient were capped. Lionel Jospin decided to lower drastically the level of the capping (law of 30 December 1998, which affected the revenues of 1998).

**Deductions of the amount paid the previous year** The law of 31 December 1945 also decided that taxpayers would be able to deduce only half of the amount they had paid the previous year from their taxable income. According to the law of 23 December 1946, this possibility was totally repealed for the incomes of 1946. But for the incomes of 1947, one quarter of the amount paid in 1946 could be deduced. The reform of 1948 (ruling of 9 December 1948) removed the possibility of deducing any fraction of the amount paid for the "progressive surtax" (formerly IGR). The possibility to deduce schedular taxes from the taxable income ended in 1970, but the amounts at stake were small.

**The reforms of 1948 and 1959** The ruling of 9 December 1948, which affected incomes from the year 1948, substituted the unique *impôt sur le revenu des personnes physiques* (IRPP) for the former IGR and schedular taxes. Actually, this new tax was divided in a proportional tax (*taxe proportionnelle*) that replaced the schedular taxes, and a progressive surtax (*surtaxe progressive*) that replaced the IGR.

The law of 28 December 1959, which applied to earnings from the year 1959, annulled the proportional tax. But it created instead a complementary tax and a 5% surtax that affected the same revenues. These new taxes only disappeared from the taxation of the year 1972.

The issue at stake with the successive reforms of the schedular taxes was the unequal treatment of wage-earners and self-employed earners.

**The taxation of capital incomes** A growing proportion of capital incomes has been removed from the income tax base. These waivers concern mostly top-earning households, and corrections have to be made to take them into account. Initially, all revenues from securities were included in the income tax base. The only exceptions were the capital gains, which have always been shielded from the income tax.

The law of 13 March 1924 introduced some exemptions for interests on Treasury bills and National Defence bills. From the 1920s and the 1930s, the list of exemptions lengthened, so that in the late 1950s, a large majority of short-term public bonds, as well as a sizeable portion of long-term government borrowings, were fully exempted from the income tax. Since the law of 29 November 1965, all the bonds and debt securities, and more generally all securities yielding fixed revenues, whoever the issuer, are exempted from the income tax if the holder agrees to pay a proportional tax (*prélèvement libératoire*, in general amounting to 15 or 25%) withheld at source. The received incomes do not appear on tax returns.

Since the exemption of the *livret A* in 1952, the number of savings accounts and plans that have been fully exempted from the income tax boomed.

At the end of the 1990s, the only securities still subject to the progressive income tax were the dividends received by shareholders. However, these revenues make up the bulk of the capital incomes received by wealthy households. Since the law of 12 July 1965, shareholders benefit by the *avoir fiscal* paid by the State, that corresponds to the profit tax settled by the firm before apportioning dividends. They also profit since the 1990s from a lump-sum allowance on all revenues from dividends.

Land incomes also have benefited from abatements and derogations. Since the law of 23 December 1964, imputed rents are exempted from the income tax.

#### 4.1.2 Corrections

We describe now the corrections that we have performed to estimate the shares of income accruing to different deciles and percentiles. The correcting rates that we apply have been calculated by Piketty [2001]. We follow his methodology.

#### 4.1.2.1 Truncated distributions

Only the taxable households fall into the scope of the tabulations issued by the tax administration. Households who benefit from dependency deductions may be nontaxable due to their family situation. These truncations underrate the number of taxpayers in each tax bracket. They also distort the distribution as reported in the tax tabulations. Indeed, The income thresholds above which households are taxable depend on the family characteristics of each household.

The methodology adopted to correct for this bias depends on the information available in tax tabulations about the distribution of family structures in the tax brackets. Three distinct periods have been distinguished in [Piketty, 2001]: 1915-1918, 1919-1944, and 1945-1998.

**Revenues of 1945-1998** Since the taxation of revenues of the year 1945, tax tabulations contain detailed information about the distribution of family structures among tax brackets. For every bracket and every number of tax shares, the number and the revenues of the corresponding taxable households are disclosed.

Taxpayers with at least 6 tax shares (that is, households with at least 8 dependent children) account for a minute portion of the overall taxpaying population. They are neglected in the corrections.

Then, let  $[\theta_i, \theta_{i+1}]$  be the lower bracket such that all households with less than 6 tax shares and whose income lies between  $\theta_i$  and  $\theta_{i+1}$  are taxable. There is no need to correct this bracket. For the previous bracket  $[\theta_{i-1}, \theta_i]$ , the total number of (taxable and nontaxable) households with 5.5 tax shares can be estimated under the hypothesis that the ratio of the total number of households with 5.5 tax shares over the total number of households with 5 tax shares is the same as in the bracket  $[\theta_i, \theta_{i+1}]$ . We carry on step by step, by correcting successively for households with 5 tax shares, 4.5 tax shares, etc, and moving from  $[\theta_{i-1}, \theta_i]$  to  $[\theta_{i-2}, \theta_{i-1}]$ ,  $[\theta_{i-3}, \theta_{i-2}]$ , etc.

We proceed in this way to construct the estimations since the taxation of revenues of 1945 from distribution tabulations corrected.

**Revenues of 1919-1944** For this period, tax tabulations give the number and the total amount of dependency deductions granted in each tax bracket. A table providing the number of taxpayers with 1, 2, ..., 13 dependent children in each bracket for the taxation of incomes of the year 1937 has also been issued by tax authorities. Tax tabulations assure that the distribution of family structures within tax brackets evolved slowly between WWI and WWII. Therefore, corrections for taxations of incomes of the years 1919-1944 are made with the same method as for the post-war period using the 1937 distribution table.

**Revenues of 1915-1918** Our estimations for the years 1915-1918 are restricted to the fractiles above 99%. Thus, there is no need to correct for truncations due to dependency deductions. However, another problem appears for the early years of the income tax. The tax tabulations have been released before the tax was levied for all taxable households. Consequently, some taxable households do not appear in the tax tabulations of the years 1915-1918. Piketty [2001] compares the total number of taxpayers and the total amount they paid that are given in tax

tabulations with definitive figures issued afterwards. He obtains correction rates for the number of taxable households: 1.57 for the year 1915, 1.29 for 1916, 1.35 for 1917 and 1.28 for 1918. Yet, if all thresholds and shares are corrected with these figures, they are underestimated for the year 1915 and overestimated for the years 1917 and 1918. In fact, top-earning taxpayers are overrepresented among latecomers in 1915, and underrepresented in 1917 and 1918. We use the corrections described in [Piketty, 2001] to amend these distortions.

#### 4.1.2.2 Shift from taxable income to fiscal income

The estimates we have obtained so far are estimates of taxable income. To construct homogenous series, we have to assess fiscal income, i.e. income before any abatement or deduction.

First, the deductibility of the taxes paid the previous year has to be taken into account. During the period from 1916 to 1947, IGR was deductible from taxable income. Scharlar taxes have been deductible from 1918 to 1970.

Moreover, income estimations have to be corrected for category reductions and abatements.

**Deductibility of IGR of the previous year (revenues 1916-1947)** Piketty [2001] assesses the rates of the tax cuts based on the deductibility of the IGR paid the previous year under the assumption that taxpayers lied in the same fractile the previous year. He deduces the corrective rates that have to be applied to estimates of taxable income to take into account this specific rule.

For instance, the corrective rate for the fraction of taxpayers above percentile 99.9% in 1930 (32.1%) is calculated as follows. The average tax rate that affected the fraction of the population above percentile 99.9% in 1929 (29.2%) is multiplied by the ratio between the average taxable income of the fraction of population above 99.9% in 1929 and the average taxable income of the share of taxpayers above percentile 99.9% in 1930:

$$29.2\% \times \frac{1,472,839}{1,336,715} = 32.1\%. \quad (4.1)$$

Of course, corrective rates in Piketty [2001] are computed with his estimations of taxable income. The corrective rates that we would obtain with our new estimates of taxable income are almost equal to the old ones, as our estimates of taxable income are very close to old estimates. So we use the corrective rates given in [Piketty, 2001].

	90%-95%	95%-99%	99%-99.5%	99.5%-99.9%	99.9%-99.99%	99.99%-100%
1916			0.2	0.7	1.3	1.4
1917			1.1	2.3	4.8	8.4
1918			1.9	3.6	12.1	19.2
1919	0.0	0.0	1.6	3.2	8.1	13.7
1920	0.0	0.4	1.3	3.6	14.4	34.0
1921	0.1	0.7	2.3	6.3	18.6	50.5
1922	0.1	0.7	2.0	4.7	14.3	35.2
1923	0.1	0.7	2.1	5.0	14.8	34.5
1924	0.2	1.0	3.1	7.9	22.4	53.9
1925	0.3	1.2	4.2	10.7	25.8	51.9
1926	0.4	1.3	3.7	8.6	19.3	37.2
1927	0.3	0.8	2.7	6.3	13.3	24.8
1928	0.3	0.8	2.4	5.6	12.9	25.0
1929	0.2	0.9	3.1	6.9	16.3	30.9
1930	0.2	0.8	2.7	6.9	15.3	32.1
1931	0.3	1.0	3.1	7.2	17.2	33.7
1932	0.2	0.8	2.6	6.1	14.6	32.3
1933	0.2	0.8	2.6	5.8	14.1	29.3
1934	0.2	0.9	2.7	5.7	13.8	31.7
1935	0.1	0.5	1.7	3.8	11.2	21.6
1936	0.1	0.5	1.6	3.6	11.3	25.8
1937	0.1	0.6	2.0	4.7	16.8	33.6
1938	0.3	0.8	2.8	7.0	22.3	49.0
1939	0.5	1.2	3.7	9.0	23.0	42.9
1940	0.4	1.1	3.1	8.4	28.4	69.5
1941	0.2	0.6	1.9	4.4	15.9	35.7
1942	0.5	1.2	3.6	9.9	27.5	59.8
1943	0.7	1.9	5.9	13.9	33.5	56.2
1944	0.5	1.8	6.6	14.6	34.7	68.6
1945	0.3	0.8	2.3	4.3	8.2	11.7
1946	0.0	0.0	0.0	0.0	0.0	0.0
1947	0.8	1.3	2.0	3.0	5.6	11.3

Table 4.3: Deductibility of the IGR: corrective rates for incomes of the years 1916-1947  
Source: [Piketty, 2001].

**Deductibility of the schedular taxes of the previous year (revenues 1918-1970)** In the case of schedular taxes, average rates affecting the different fractiles of the taxpaying population are difficult to estimate. Indeed, composition tabulations only appear after 1948. Furthermore, only indirect information about the level of individual category revenues is found in these tabulations.

However, the average rates of schedular taxes are quite low, so that corrections are much less significant than in the case of the deductibility of the IGR.

Piketty [2001] obtains rough estimates of the corrective rates to be applied by hypothesizing average rates of schedular taxes affecting the different parts of the population. The resulting rates are consistent with totals reported in composition tabulations.

We give below these rates combined with the rates correcting for the deductibility of the IGR.

	90%-95%	95%-99%	99%-99.5%	99.5%-99.9%	99.9%-99.99%	99.99%-100%
1916			0.2	0.7	1.3	1.4
1917			2.7	4.8	8.2	13.2
1918			3.7	6.3	16.0	24.5
1919	0.0	0.0	5.1	8.4	14.1	21.5
1920	0.8	2.8	5.3	9.4	22.6	42.6
1921	1.0	3.6	7.3	13.7	28.3	62.6
1922	1.0	3.4	6.6	11.1	22.4	44.3
1923	1.0	3.4	6.4	10.9	22.6	43.1
1924	1.0	3.8	7.7	14.4	31.4	64.9
1925	1.2	3.7	8.8	17.3	34.3	61.0
1926	1.3	4.0	8.0	14.6	26.8	44.8
1927	1.2	3.7	7.8	13.3	22.0	34.2
1928	1.2	3.7	7.1	12.2	21.5	34.5
1929	1.1	3.7	8.1	13.8	25.7	41.4
1930	1.1	3.7	7.6	14.1	24.5	43.1
1931	1.3	4.2	8.6	15.1	27.8	45.7
1932	1.3	3.9	7.9	13.6	24.5	44.0
1933	1.2	3.8	7.7	13.0	23.5	39.3
1934	1.3	4.1	8.0	13.1	23.2	42.7
1935	1.2	3.6	6.8	11.0	20.4	31.0
1936	1.1	3.3	6.3	10.2	19.6	35.0
1937	1.0	3.3	6.4	10.7	24.6	42.0
1938	1.2	3.6	7.3	13.5	31.1	59.6
1939	1.6	4.4	9.2	16.4	31.6	52.0
1940	1.5	4.2	8.3	16.1	38.9	83.4
1941	1.0	3.0	5.6	9.5	22.8	43.5
1942	1.3	3.7	7.7	16.0	36.1	70.6
1943	1.6	4.4	10.4	20.4	42.5	66.2
1944	1.4	4.4	11.3	21.5	44.1	80.7
1945	0.8	2.4	5.0	8.2	12.8	16.0
1946	0.6	1.8	2.7	3.6	4.1	4.1
1947	1.6	3.5	5.9	8.6	13.0	20.6
1948	0.5	1.9	3.1	4.2	5.3	5.7
1949	0.8	2.4	4.0	5.5	6.8	7.2
1950	0.9	2.6	4.3	6.1	7.9	8.6
1951	0.8	2.3	3.9	5.6	7.3	8.4
1952	0.9	2.5	4.2	5.8	7.7	9.0
1953	1.0	3.1	5.1	7.1	9.1	9.8
1954	1.0	2.9	4.8	6.7	8.7	9.5
1955	0.9	2.7	4.5	6.4	8.2	9.1
1956	0.9	2.8	4.5	6.4	8.3	9.2
1957	0.9	2.6	4.5	6.3	8.1	9.0
1958	0.9	2.7	4.5	6.4	8.2	9.5
1959	1.0	2.8	4.7	6.5	8.5	9.3
1960	0.8	2.5	4.1	5.7	7.2	8.0
1961	0.8	2.2	3.7	5.3	6.8	7.5
1962	0.7	2.0	3.4	4.9	6.3	7.3
1963	0.6	1.8	3.0	4.3	5.6	6.2
1964	0.5	1.6	2.6	3.7	4.8	5.4
1965	0.5	1.4	2.3	3.3	4.2	4.6
1966	0.4	1.2	2.0	2.8	3.6	3.8
1967	0.3	0.9	1.5	2.1	2.7	2.9
1968	0.2	0.7	1.2	1.7	2.2	2.4
1969	0.2	0.5	0.8	1.1	1.4	1.5
1970	0.1	0.2	0.4	0.5	0.7	0.8

Table 4.4: Deductibility of the IGR and schedular taxes: global corrective rates for incomes of the years 1916-1970

Source: [Piketty, 2001].

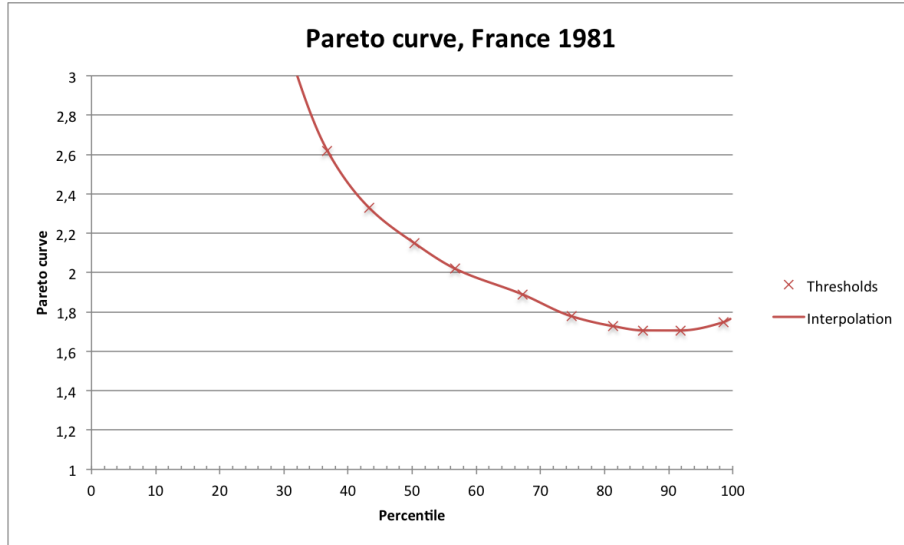


Figure 4.1: Pareto curve of the income distribution, France 1981

**Category deductions and abatements** Last, various category deductions and abatements have to be taken into account to obtain fiscal income. The main evolutions throughout the XX<sup>th</sup> century concern wage income. The 10% compensation for business expenses has been instituted in 1934. The abatement rate was 0% for wage revenues of the years from 1915 to 1952, 10% in 1953, 15% between 1954 and 1958, 19% in 1959, and 20% after 1960.

Piketty [2001] computes the following corrective rates, based on legislative developments and on ratios of fiscal income over taxable income observed since the 1970s.

	0%-100%	90%-95%	95%-99%	99%-99.5%	99.5%-99.9%	99.9%-99.99%	99.99%-100%
1915-1952	18	18	16	16	16	14	11
1953	25	25	22	19	16	14	11
1954-1958	33	33	30	25	22	16	11
1959	41	41	37	32	27	18	11
1960-1998	43	43	39	33	28	19	11

Table 4.5: Category abatements: corrective rates for incomes of the years 1915-1998

Source: [Piketty, 2001].

### 4.1.3 Estimations

#### 4.1.3.1 Sources

The French Finance Ministry periodically issued statistic tabulations describing the distribution of taxpayers and incomes among the income tax brackets. Tax tabulations are found in the *Bulletin de Statistique et de Législation Comparée* (taxation of incomes of the years 1915-1937), in the *Renseignements Statistiques Relatifs aux Impôts Directs* (incomes of the years 1923-1929) in the *Bulletin de Statistique du ministère des Finances* (incomes of the years 1938-1945), in the *Statistiques et Études Financières* (incomes of the years 1946-1981), and in the *États 1921* (incomes from 1982).

Tax tabulations are now published on statistics section of the official website <http://www.impots.gouv.fr/>.

#### 4.1.3.2 Results

Figure 4.2 gives the ratios of our estimations of the distribution of taxable income over estimations obtained with the piecewise polynomial method found in [Piketty, 2001]. They are very close, especially estimations of average incomes above different percentiles of the distribution. Indeed, the multiplication by the Pareto coefficient of the threshold estimation offsets approximation errors due to local variations of  $b$ . Predicted values of the average income are less sensitive to errors in  $b$  than predicted values of thresholds.

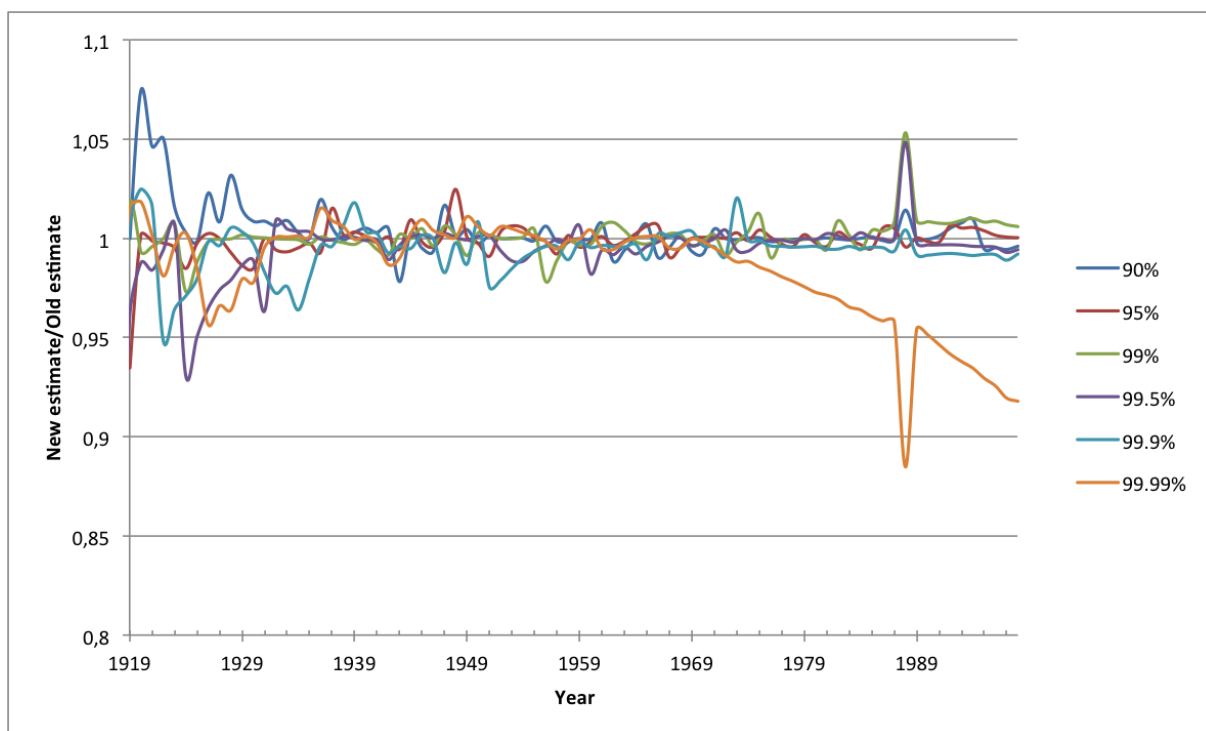
The downward trend followed by our estimates for percentiles 99.9% and 99.99% since the 1970s is explained by the fact that Piketty [2001] applied corrective rates to his values in order to get results closer to samples provided by the tax administration. This proves that our values extrapolated for the top of the distribution are in fact too low, and encourage us to deepen new extrapolation methods such as the use of splines with tension (see appendix B).

Another irregularity occurs for the year 1988. Actually, the data in the income tax tabulation of this specific year include the proportional rates capital gains (*plus-values à taux proportionnels*), which were not included elsewhere. Therefore Piketty did not use tax data for his estimations of incomes of the year 1988.

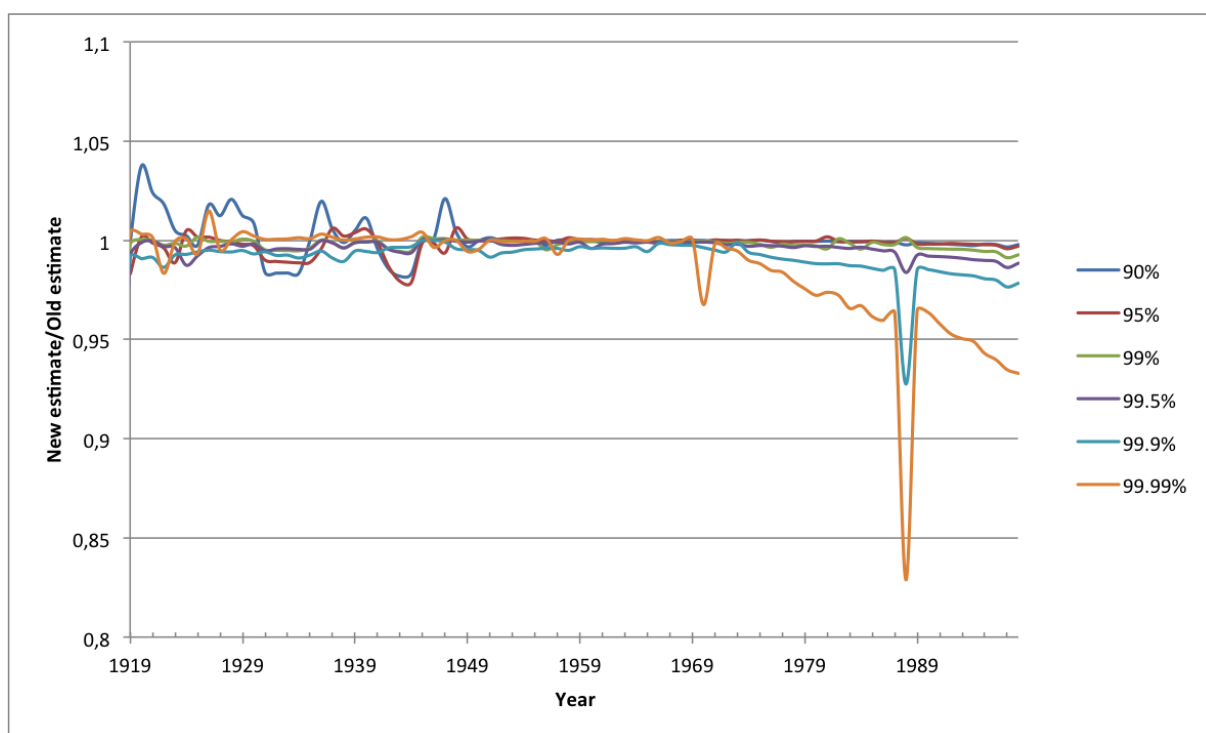
We do not provided similar ratios for fiscal income: we have applied exactly the same corrective rate as Piketty [2001], so the ratios would be identical.

We also compare in figure 4.3 the evolution of shares of total income accruing to the the top 10%, top 1%, top 0.1% of taxpayers for the old and new time series.

Note that we did not construct new estimates for the years 1915-1918: indeed, the corrections that had to be applied after Paretian approximation were too rough to expect improving the precision with a change of method.

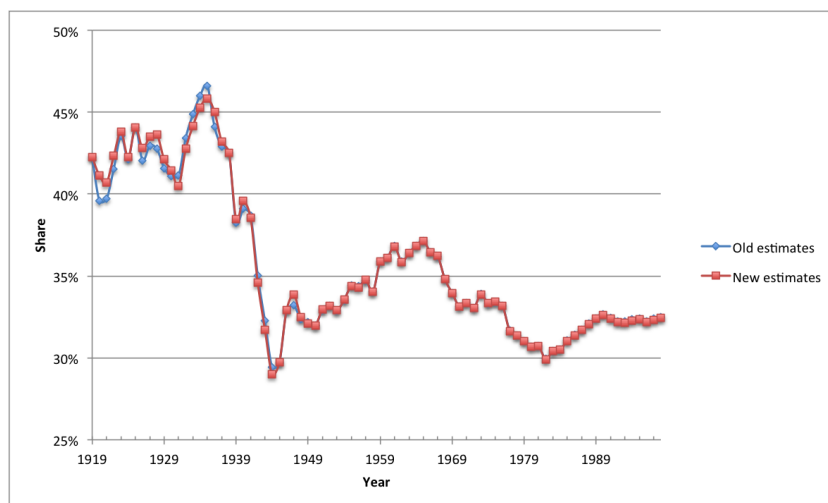


(a) Thresholds

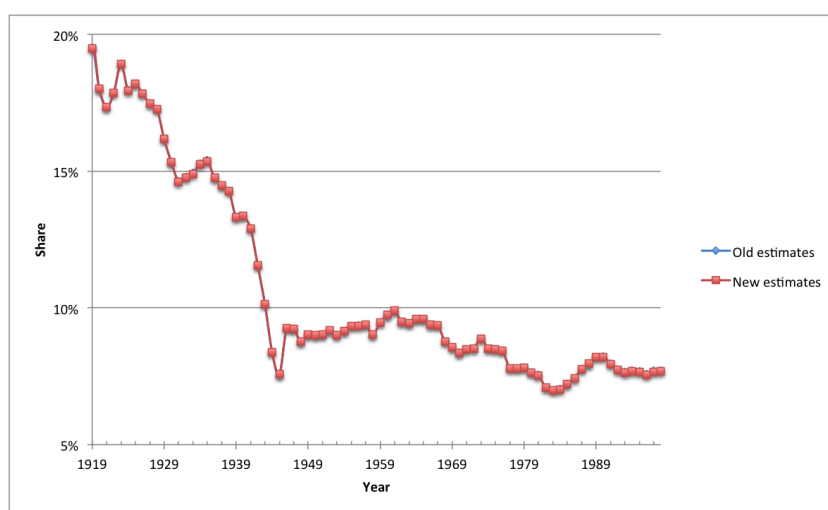


(b) Average income above

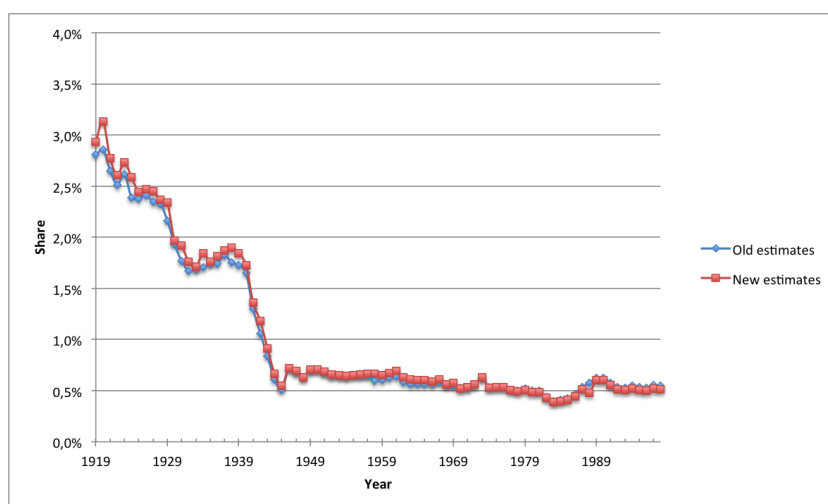
Figure 4.2: Comparison of new estimations of taxable income with estimations of Piketty [2001]



(a) Top 10%



(b) Top 1%



(c) Top 0.01%

Figure 4.3: Evolution of the top shares of income accruing to different percentiles of the population

## 4.2 Application to inheritance tax tabulations 1902-1994

We exploit inheritance tax tabulations with our new method. By doing so, we are able to estimate accurately the distribution of inheritance left by different deciles and percentiles of the population. Associated with demographical considerations, these estimations give a hint about the distribution of wealth within the population.

We first overview the legislation of the inheritance tax since its establishment in France in 1901, using the historical review given in [Piketty, 2001]. We then provide estimations of the distribution of inheritance using raw tabulations provided by tax administration.

### 4.2.1 The inheritance tax in France

The inheritance tax has been the first countrywide progressive tax introduced in France. It was established by the law of 25 February 1901.

The primary characteristic of the French inheritance tax is that tax rates that apply are based on the actual estate value which is paid to the heirs (the *part successorale*), and not on the total legacy left by the deceased.

In addition, there have always existed several tax scales, depending on the degree of kinship between the deceased and the legatee: transfers to lineal descendants, to spouses, to relatives in the collateral line or to unrelated persons are not taxed at the same rates. Here, we will estimate the level of the different fractiles of the total legacy left, irrespective of heirs.

A second feature of the law of 25 February 1901 is that bequests and donations were initially treated unequally. Legacies were taxed according to progressive scales, while donations were taxed according to proportional rates (that also varied depending of the degree of kinship). The law of 14 March 1942 rectified this flaw. Since then, donations are subject to the same progressive tax scales as inheritance. Prior donations are "recalled" and added to the inheritance bequeathed at the moment of death to calculate the tax owed. However, some tax benefits are still granted to donations. The donor can pay himself the tax due for the donation, which is not "recalled" at time of the heritage. The value of a donation, is not in general discounted at the time of the heritage, which is profitable if inflation is high. Finally, some categories of donations benefited from preferential tax regimes.

This preferential treatment of donations can bias estimations of wealth from inheritance tabulations if the most affluent use more than others the opportunity to donate a part of their fortune to avoid heavier taxation.

	Positive legacies	Thresholds
1902	65.0%	8
1903	69.5%	12
1904	68.1%	12
1905	66.3%	12
1907	65.6%	12
1909	63.9%	12
1910	65.5%	12
1911	61.6%	12
1912	65.6%	12
1913	65.4%	12
1925	65.9%	12
1926	69.3%	12
1927	67.8%	12
1929	63.0%	12
1930	65.4%	12
1931-32	64.1%	12
1932	65.9%	12
1933	61.9%	12
1935	63.7%	12
1936	63.9%	12
1937	64.9%	12
1938	65.8%	12
1939	59.6%	12
1940	45.0%	12
1941	58.3%	12
1942	60.4%	12
1943	60.8%	14
1944	53.6%	13
1945	58.0%	13
1946	61.7%	8
1947	67.4%	8
1948	63.4%	8
1949	57.1%	8
1950	58.0%	8
1951	55.6%	8
1952	60.3%	9
1953	50.7%	8
1954	60.2%	9
1955	59.6%	9
1956	58.2%	9
1957	60.5%	9
1958	67.4%	8
1959	67.1%	8
1960	59.1%	8
1962	59.7%	8
1964	66.1%	8
1984	56.2%	7
1994	63.8%	8

Table 4.6: Share of positive legacies and number of thresholds in the inheritance tax tabulations by year

Source: [Piketty, 2001].

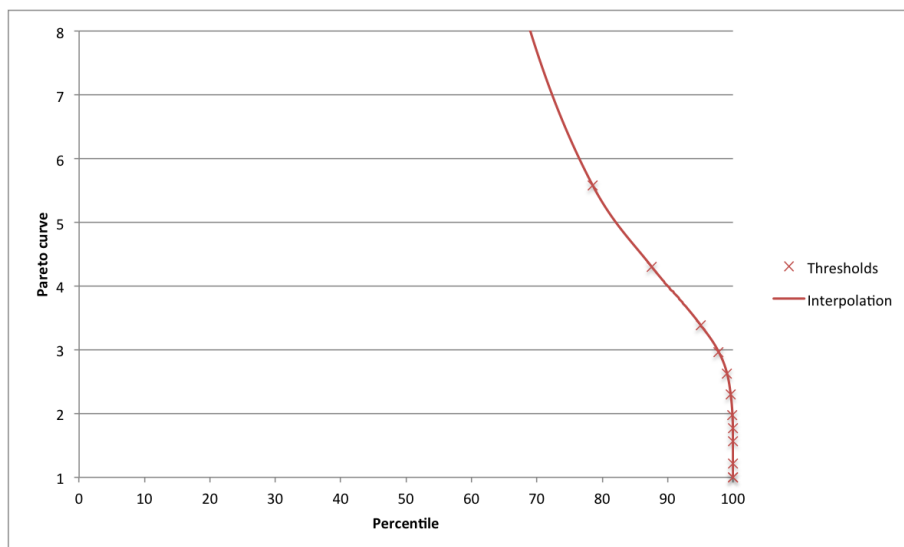


Figure 4.4: Pareto curve of the inheritance distribution, France 1943

## 4.2.2 Estimations

### 4.2.2.1 Sources

Since the institution of the inheritance tax in France in 1901, tax authorities have periodically processed inheritance declarations. They issued several series of statistic tabulations giving the number and the distribution of bequests within the different tax brackets. They also release composition tabulations indicating the different types of goods inherited.

We have exploited distribution tabulations. The statistics are missing for a significant number of years. They have been successively published in the *Bulletin de Statistique et de Législation Comparée* (ministère des Finances, successions of the years 1902-1905, 1907, 1909-1913, 1925-1927, 1929-1933, 1935-1938), in the *Bulletin de Statistique du ministère des Finances* (ministère des Finances, successions of the years 1939-1946), in the *Statistiques et Études Financières* (ministère des Finances, successions of the years 1947-1960, 1962, 1964), in *"L'imposition du capital", 8ème Rapport au Président de la République* (Conseil des Impôts, successions of the year 1984), and in *"L'imposition du patrimoine", 16ème Rapport au Président de la République* (Conseil des Impôts, successions of the year 1994).

### 4.2.2.2 Results

We have applied our estimation method to raw data found in the tabulations.

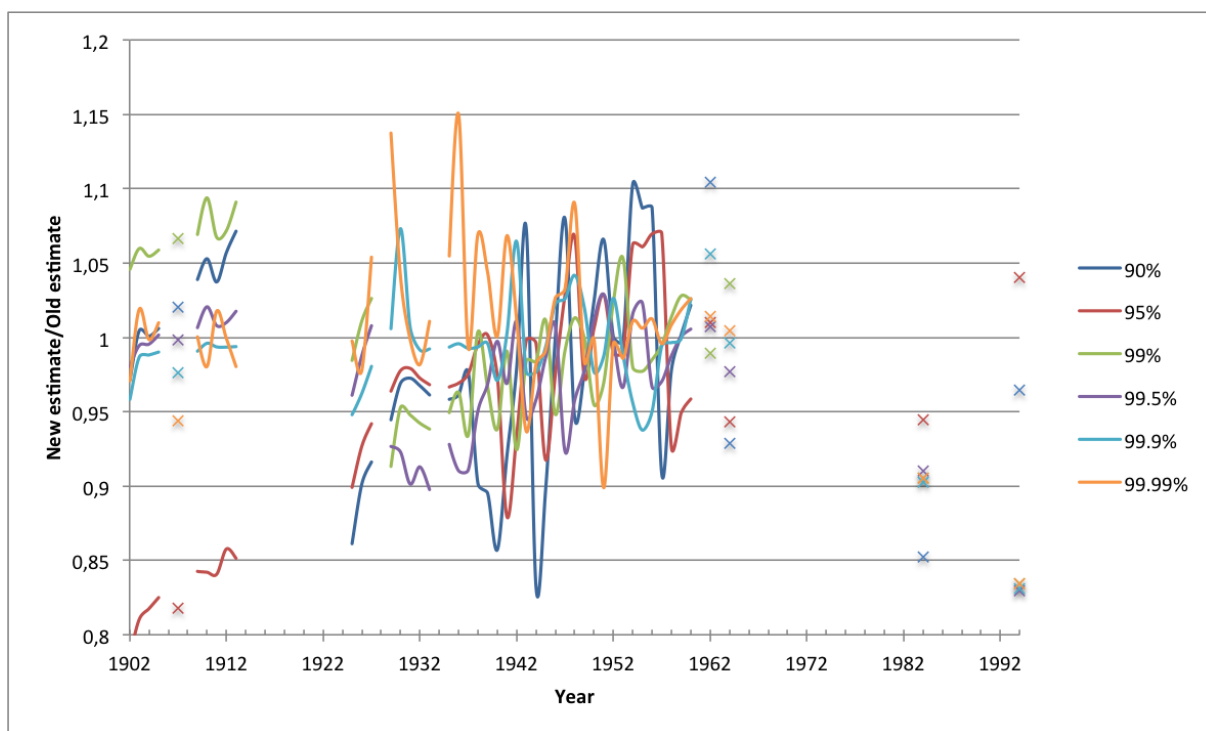
The main difference with Pareto curves of the income distribution is that a large part of the population leaves little or no inheritance to its descendants. As a smaller sample of the population is observed each year (there have been roughly 500 000 adults deaths a year throughout the XX<sup>th</sup> century), the Pareto coefficient starts to decrease for lower percentiles than in the case of incomes.

We have constructed the shares series using the average inheritance series assessed with the number of adults deaths (provided in [Piketty, 2001]).

The ratios of new estimates over old estimates, both for thresholds and average income of top inheritance groups, are plotted in figure 4.5. The series of average income above different percentiles of the inheritance distribution appear to be very close for old and new estimates. For the years 1984 and 1994, the raw estimates are actually very close, but Piketty [2001] applies

correcting rates.

Figure 4.6 compares the evolutions of top inheritance shares throughout the XX<sup>th</sup> century between old and new estimates.



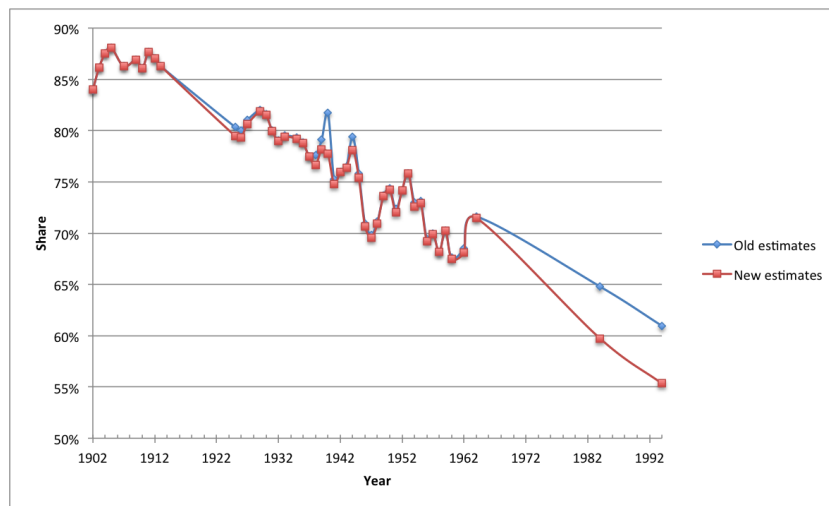
(a) Thresholds corresponding to different percentiles of the distribution



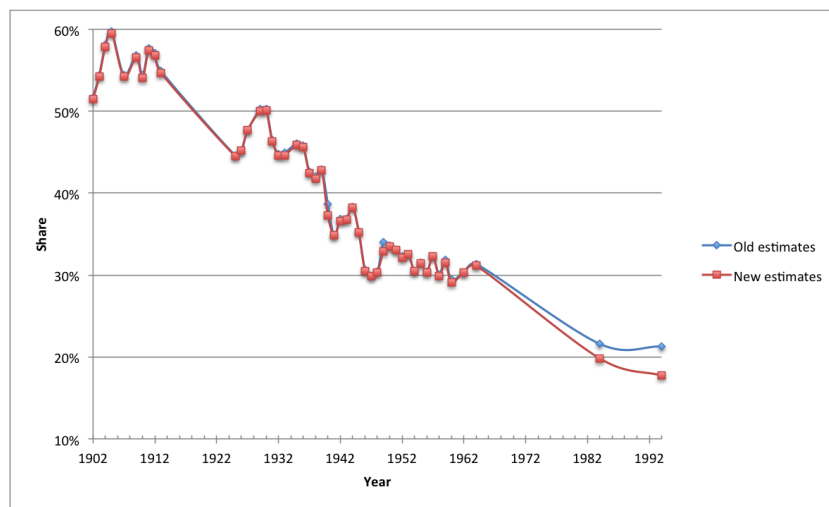
(b) Average income above different percentiles of the distribution

Figure 4.5: Comparison of new estimations of the inheritance distribution with estimations of Piketty [2001]

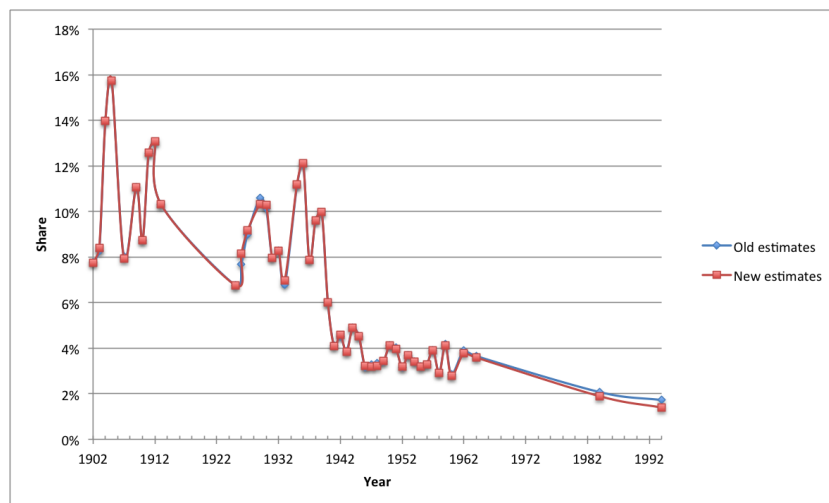
Source: Interpolation from raw inheritance tax tabulation.



(a) Top 10%



(b) Top 1%



(c) Top 0.01%

Figure 4.6: Evolution of the top shares of inheritance accruing to different percentiles of the population

## Section 5

# Conclusion

We worked out a new nonparametric method to assess the income and wealth distributions from tax tabulations data. While the literature until now had concentrated on parametric estimations, we have been able to relax any prior assumption about the functional form of the distribution. To do so, we have first defined the *generalized Pareto curve* that we can approximate numerically using tabulations data. This generalized Pareto curve characterizes the overall distribution.

We have suggested several applications. Our method provides precise estimations of shares of income and wealth accruing to different deciles and percentiles of the distribution. Their accuracy has been demonstrated with microdata provided by tax authorities in France for the year 2006. Contrary to usual parametric methods, our technique is not restricted to the Pareto-like top of the distribution, and allows to construct estimates for whole taxpaying population. It also permits the generation of synthetic micro-files representing the entire population with incomes distributed just as the actual population. A last application is the homogenization of series obtained for individual-based tax systems and for household-based tax systems.

This research into tax-based estimation methods can be extended into two directions.

First, the analytic methods used to estimate the empirical generalized Pareto curves could be refined. This would permit both to take full advantage of the information provided in tax tabulations and to incorporate empirical findings to complete gaps in the tax-based data, in particular for the bottom and the very top of the distribution. Indeed, tax tabulations do not tell us anything about the shape of the Pareto curve for highest incomes or wealth above the last threshold, or about low incomes and inheritance that are exempted. Hence, we have to extrapolate the Pareto curve outside the range of the tax scale. For higher percentiles of the distribution, a rise of the Pareto coefficient near 1 appears to be well-investigated. This empirical fact has to be taken into account when there are not enough thresholds in the tax scale for the top of the distribution to be assessed. Besides, depending on the level of the general allowance, a large part of the population may be exempted from paying the income or inheritance tax. To fill in this gap, the shape of the Pareto curve below the lower tax threshold has to be extrapolated using both additional data (for example the average and the median incomes) and observations for years when reliable micro-files have been made available by tax authorities. Interpolation technique to approximate the Pareto curve between tax thresholds may also be improved. The curve  $b(p)$  appears to be tighter for the middle part of the distribution than for the top. The varying curvature may be integrated if we employ more general interpolants such as tension splines.

Second, the two-dimensional nature of the income distribution has to be investigated. Indeed, income decompose into two components, labor income, which relates to the wage distribution, and capital income, which relates to the wealth distribution. The appropriate mathematical tools to

deal with this two-dimensional distribution are the copulas<sup>1</sup>. These functions join multivariate distribution functions to their one-dimensional margins. Aaberge et al. [2015] prescribe their use to analyze finely the dramatic transition from a class system, in which top incomes were predominantly made of capital income, to system were top wage earners and capital owners appear to co-habitate the top of the income distribution. The question is to determine whether the two groups are merging, or if the rentiers are being elbowed out of the top income groups. As stated by the authors, four ingredients affect the distribution of personal incomes: the respective shares of labor and capital income in the national product, the marginal distribution of wages, the marginal distribution of capital incomes, and the correlation between earned and capital incomes. Copulas allow to disentangle these effects. They isolate the last element so that its evolution can be studied separately. The exploitation of composition tax tabulations may provide information about the composition of earnings of top income groups. But they tell us little about how the two dimensions of the income distributions relate at the individual level. Further examination of these tabulations would help to realize to which extent we can ascertain the copula corresponding to the empirical income distribution using only tax data. Estimating copulas can also help finding the appropriate corrections to pass from taxable income to fiscal income. Last, copulas can be used to generate two-dimensional micro-files that correspond to the true distributions of labor and capital incomes.

---

<sup>1</sup>See [Nelsen, 2006] for a reference.

## Appendix A

# Pareto curves of usual parametric distributions

We display in this appendix the generalized Pareto curves of the main parametric distributions that have been used in the literature to mirror the income and wealth distributions for various choices of parameters.

In each case, the Pareto curve  $b(p)$ ,  $0 \leq p \leq 1$ , has been computed numerically from the formula:

$$b(p) = \frac{1}{(1-p)Q(p)} \int_p^1 Q(r)dr \quad (\text{A.1})$$

where  $Q = F^{-1}$  is the quantile function associated to the distribution.

We can discern the Paretian distributions, for which:

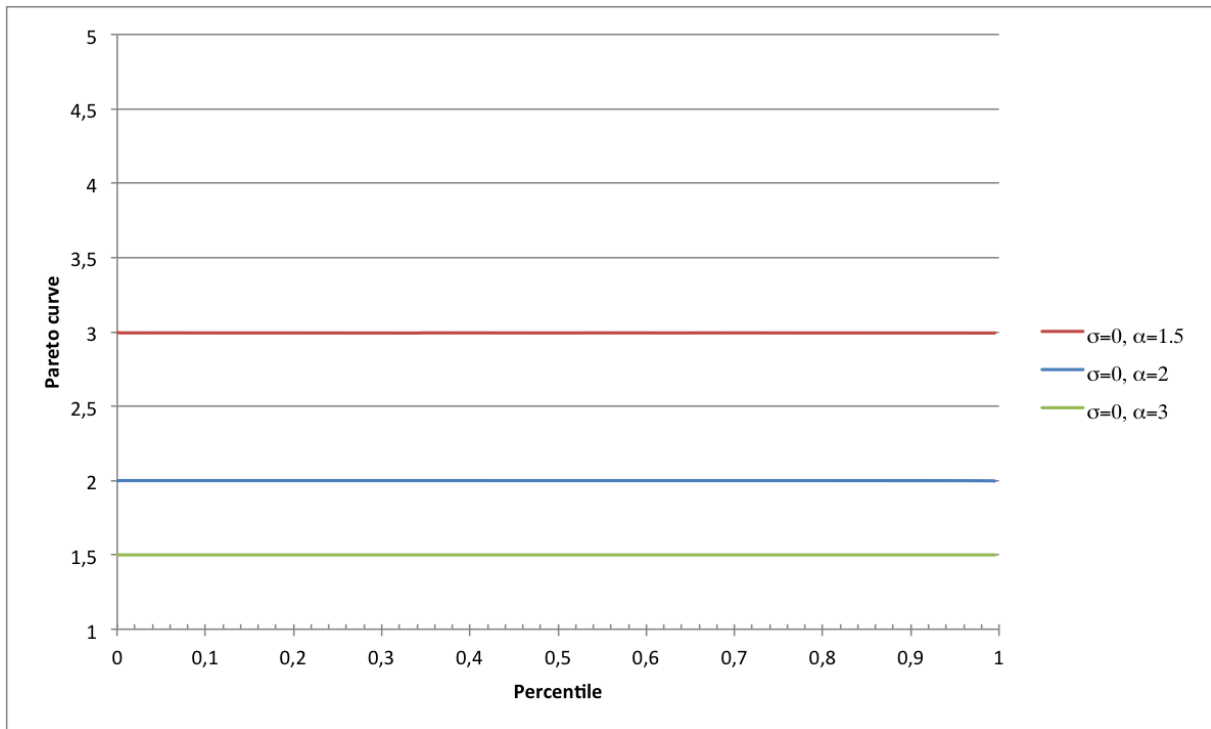
$$\lim_{p \rightarrow 1} b(p) > 1. \quad (\text{A.2})$$

All types of Pareto distributions, the Champernowne distribution, the Sech<sup>2</sup> distribution and the Singh Maddala distribution satisfy this criterion.

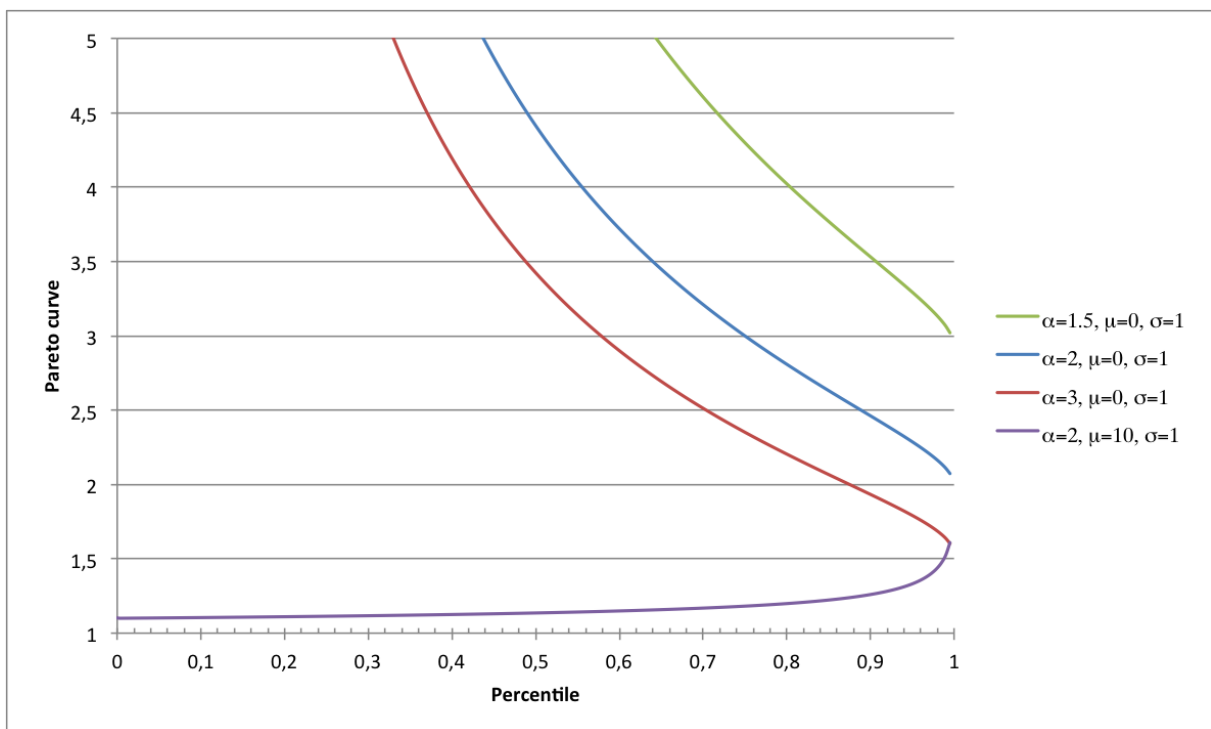
Unlike them, the lognormal and the three-parameter lognormal distributions, the Gamma and the generalized Gamma distributions and the Weibull distribution are characterized by the behavior near 1 of their Pareto curves:

$$\lim_{p \rightarrow 1} b(p) = 1. \quad (\text{A.3})$$

We notice that the Pareto curve of the Pareto type III and IV, Champernowne, Sech<sup>2</sup> and Singh-Maddala distributions do resemble the empirical Pareto curve observed in microdata provided by the French tax administration for the year 2006.

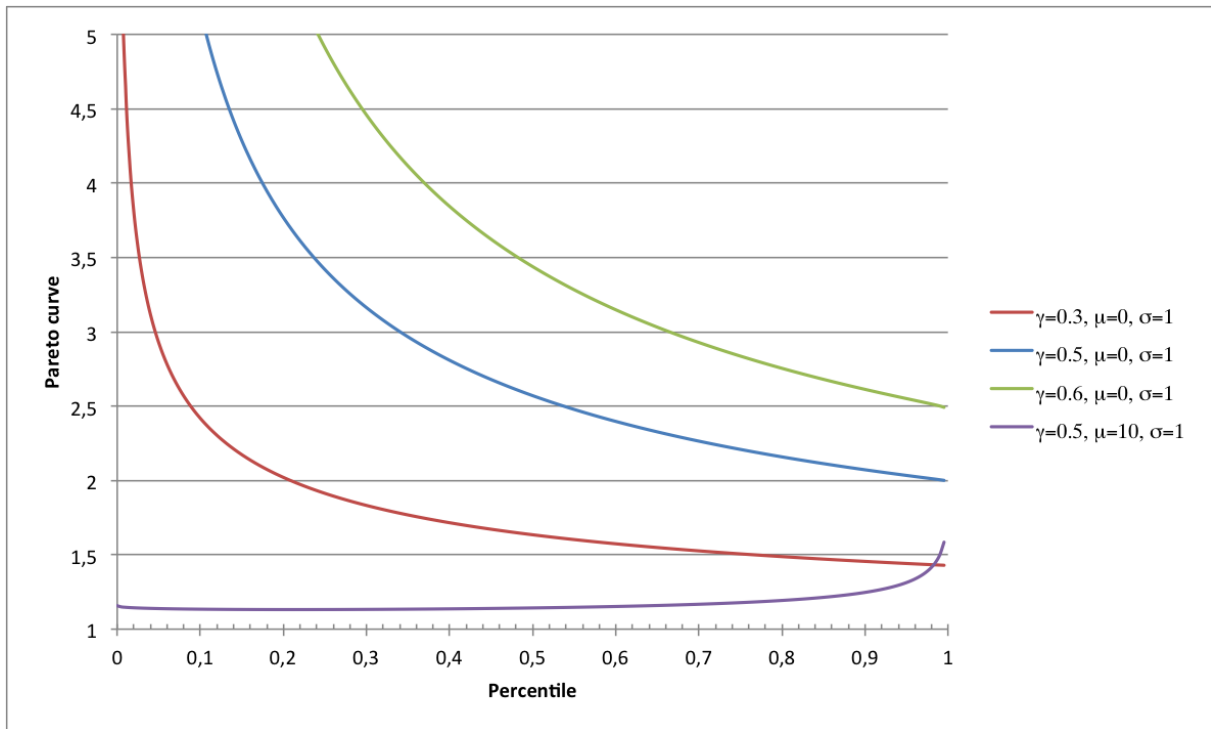


(a) Pareto type I distribution

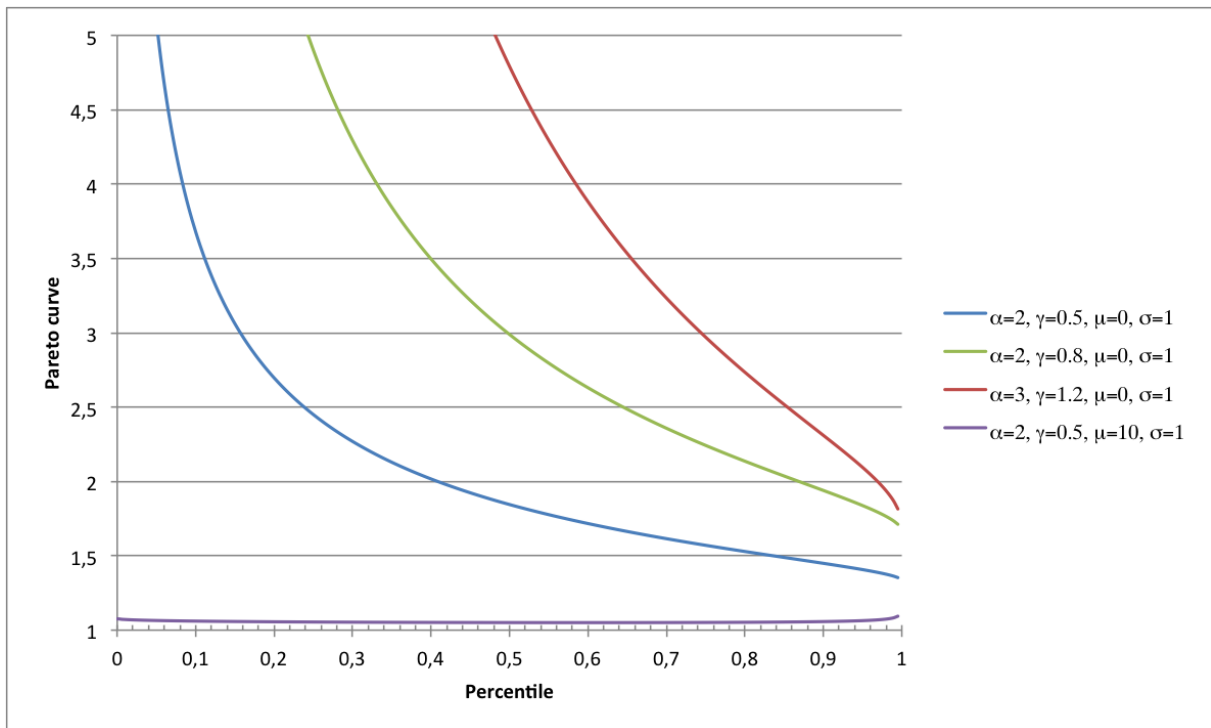


(b) Pareto type II distribution

Figure A.1: Pareto curves of Pareto type I and type II distributions

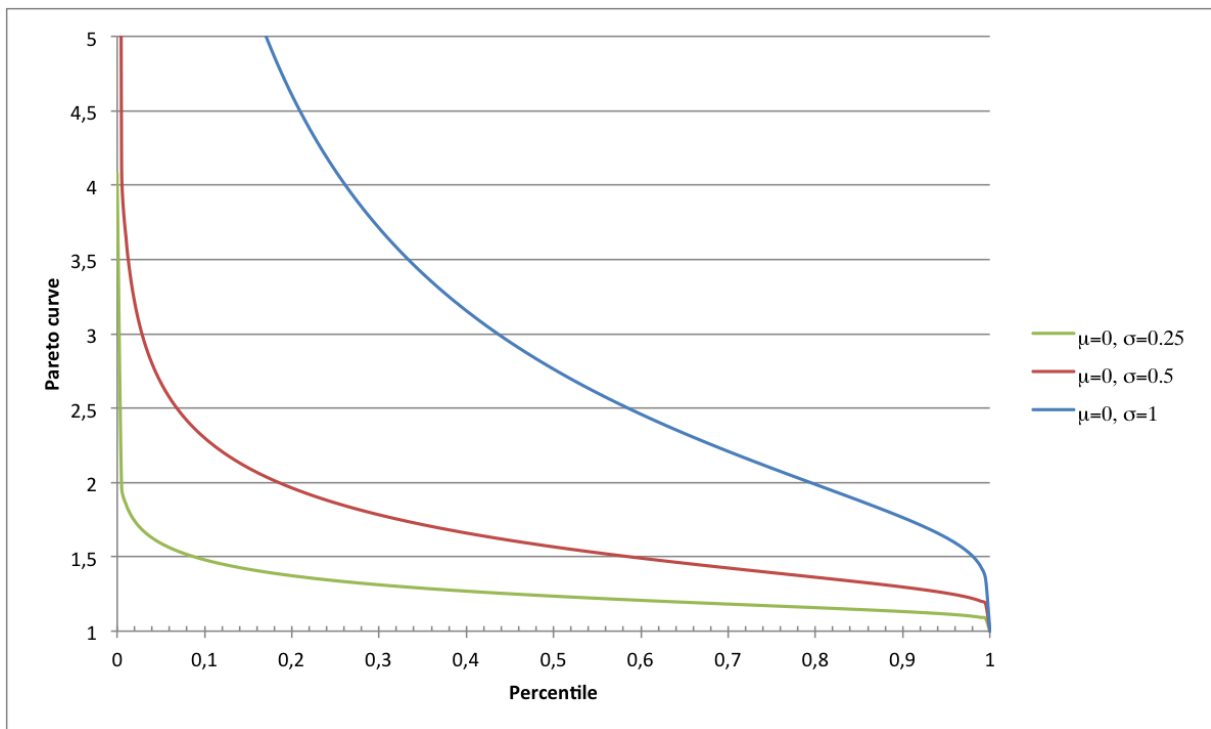


(a) Pareto type III distribution

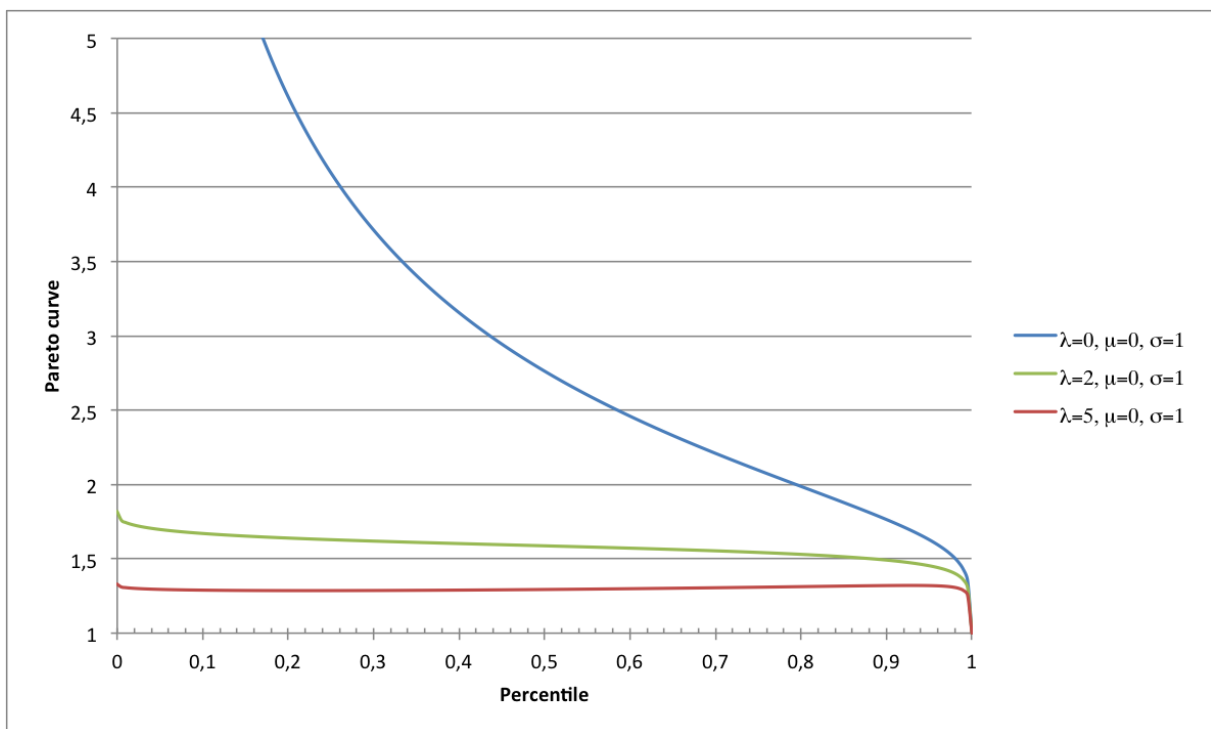


(b) Pareto type IV distribution

Figure A.2: Pareto curves of type III and type IV Pareto distributions

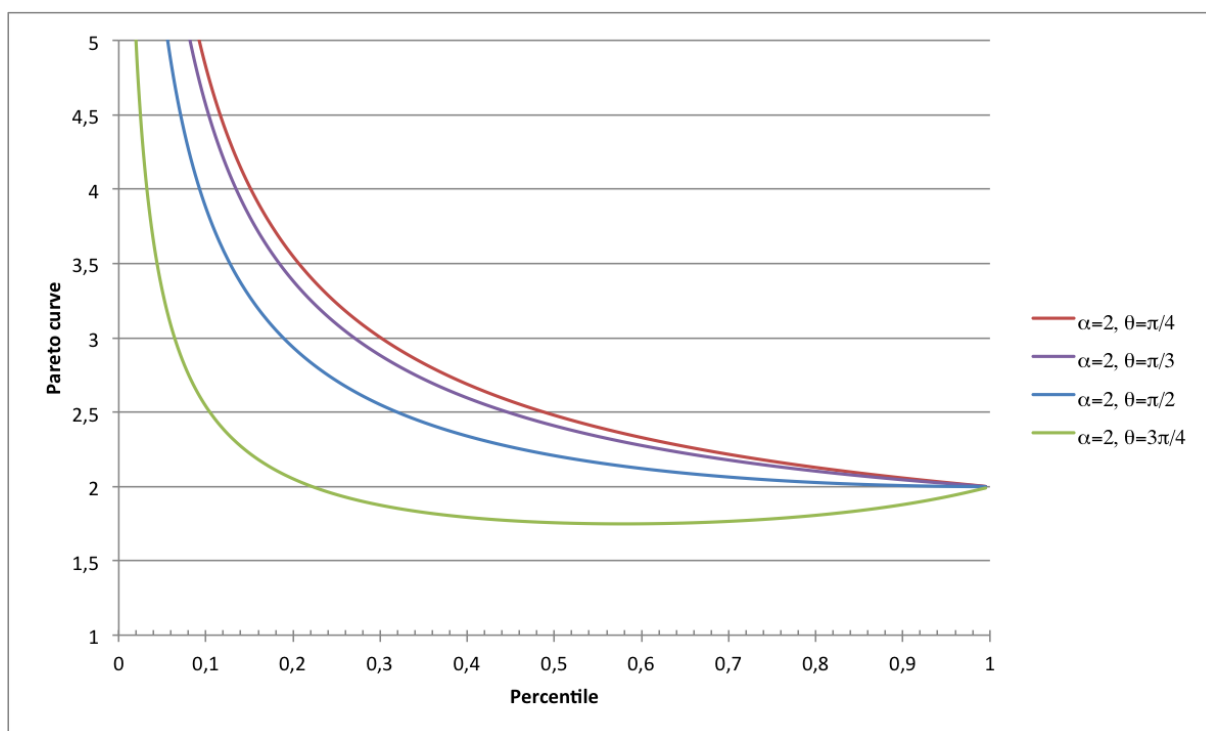


(a) Lognormal distribution

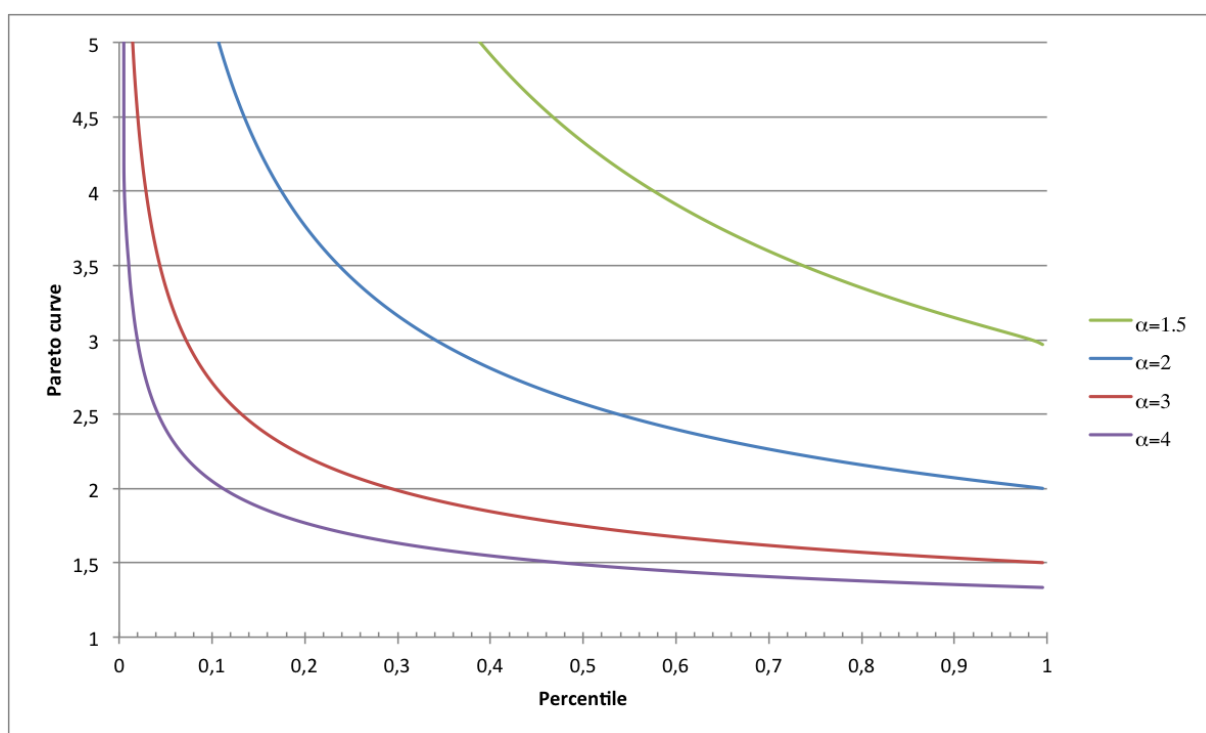


(b) Three-parameter lognormal distribution

Figure A.3: Pareto curves of lognormal distributions

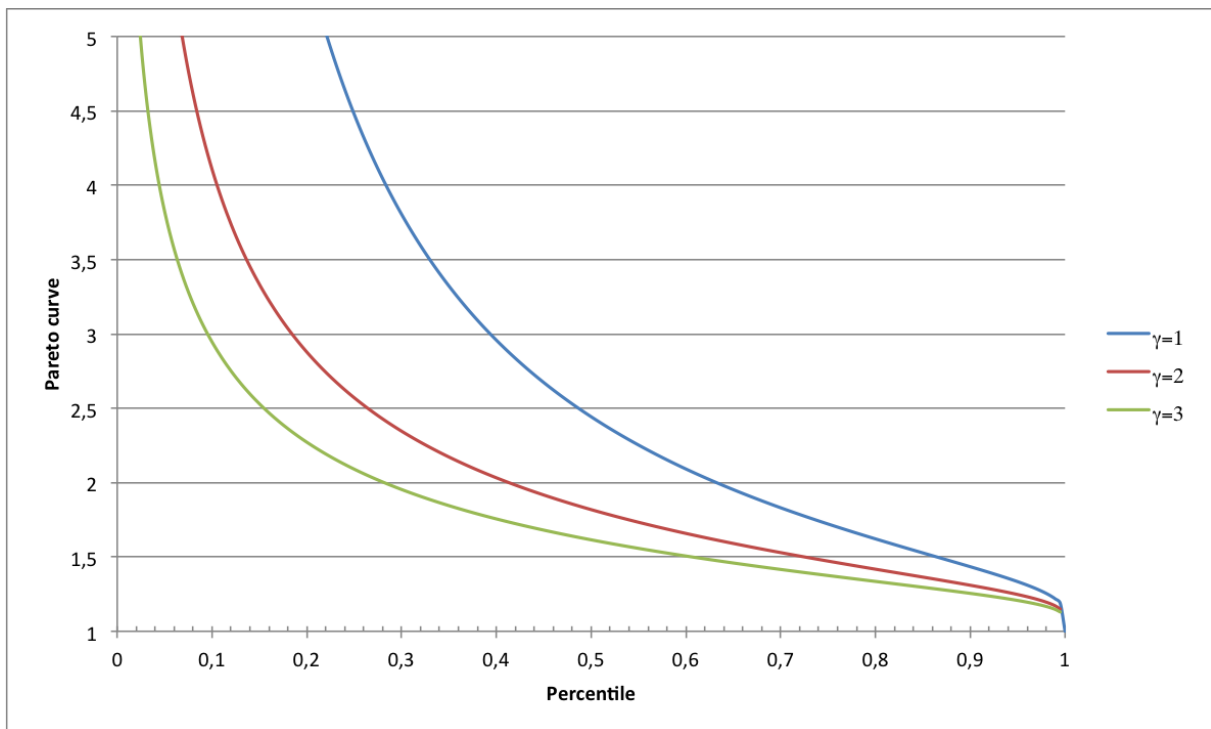


(a) Champernowne distribution

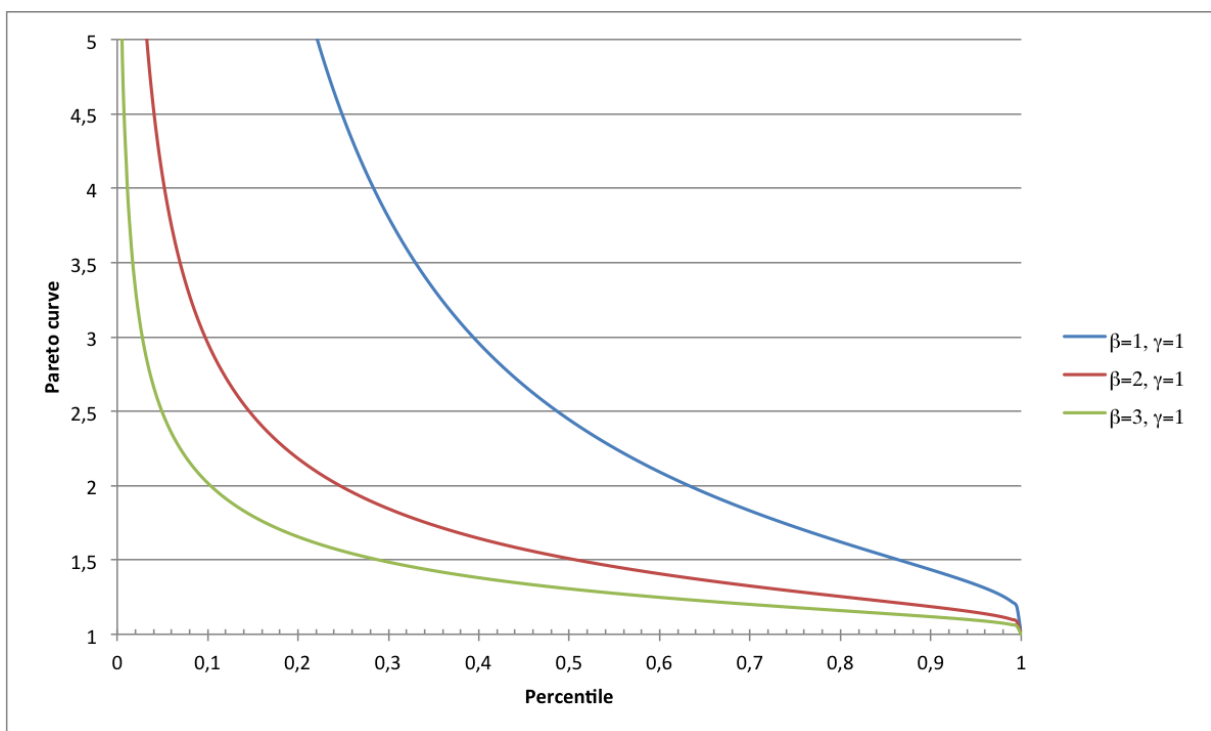


(b)  $\text{Sech}^2$  distribution

Figure A.4: Pareto curves of Champernowne and  $\text{Sech}^2$  distributions

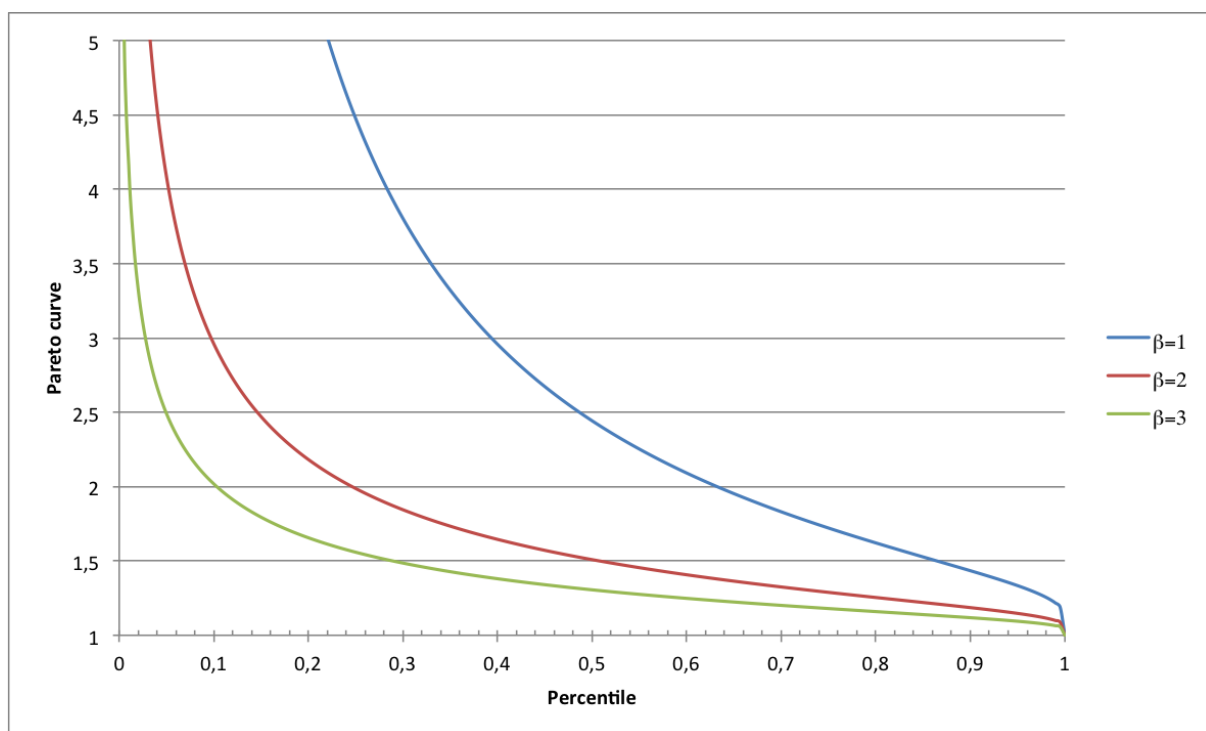


(a) Gamma distribution

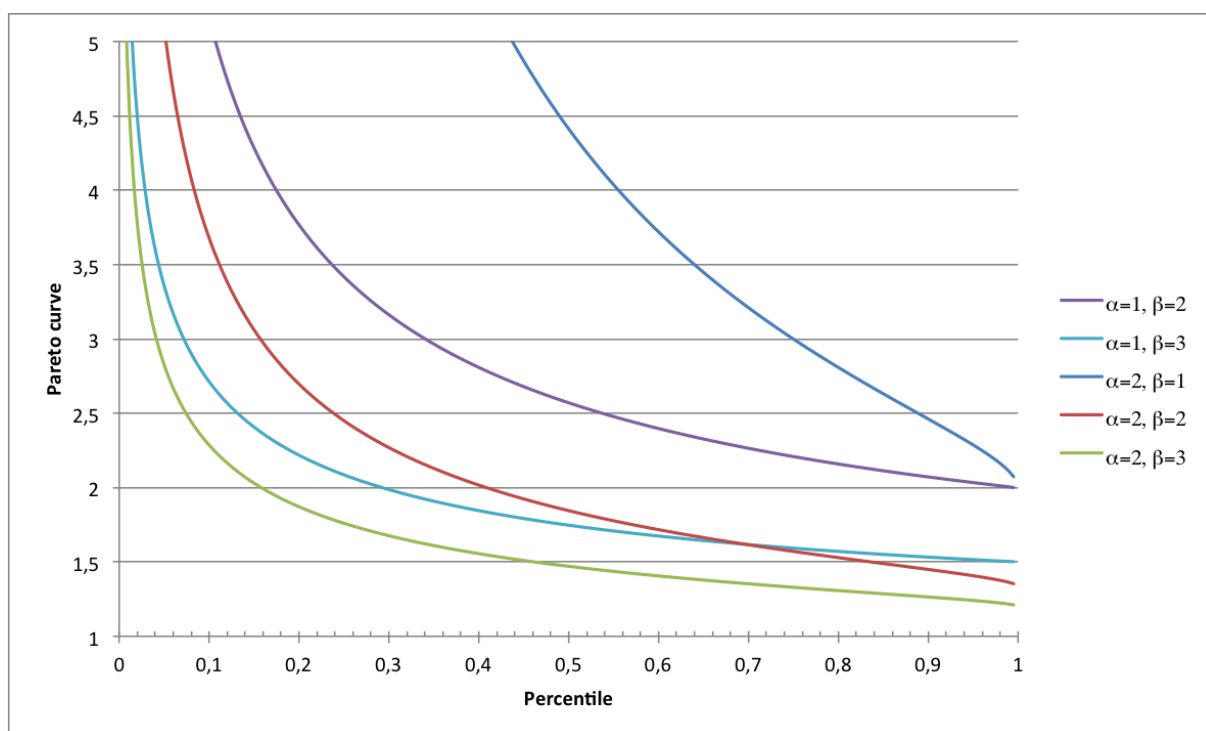


(b) Generalized Gamma distribution

Figure A.5: Pareto curves of Gamma distributions



(a) Weibull distribution



(b) Singh-Maddala distribution

Figure A.6: Pareto curves of Weibull and Singh-Maddala distributions

## Appendix B

# Estimating the generalized Pareto curve

It is often required to estimate the value of a function for intermediate values of the independent variable given a discrete set of data points where this function is known. Curve fitting refers to numerical analysis methods designed to construct new points within the range of a number of known values of the independent variable. It involves either *smoothing*, where a smooth function is built which approximately fits the data, or *interpolation*, which necessitates an exact fit to the data.

We explain in this appendix how to approximate numerically the empirical Pareto curve given the data found in tax tabulations. Specifically, we know the value of the empirical Pareto curve  $b(p)$  at a number of thresholds of the tax scale which correspond to percentiles  $p_1, \dots, p_\omega$  and where  $b(p)$  is equal to  $b_1, \dots, b_\omega$  respectively. In practice, the application of the method described in section 3.1 crucially rests on the quality of the approximation.

### B.1 A first try: approximation by a suited functional form

A first strategy is to seek a manageable functional form to approximate the Pareto curve. A natural candidate is:

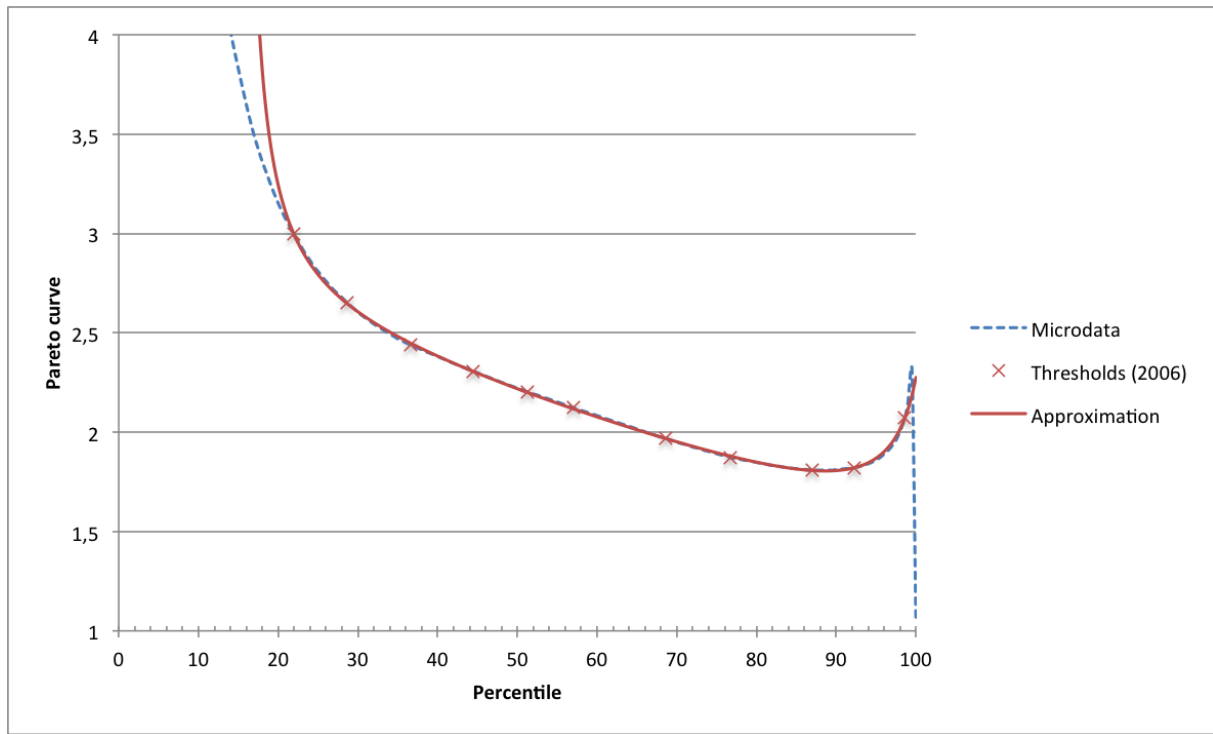
$$b(p) = a + \frac{b}{p - p_{min}} + c \cdot p + \frac{d}{1 + \varepsilon - p}, \quad 0 \leq p \leq 1 \quad (\text{B.1})$$

where  $a, b, c, d, p_{min}$  and  $\varepsilon$  are the parameters to estimate.

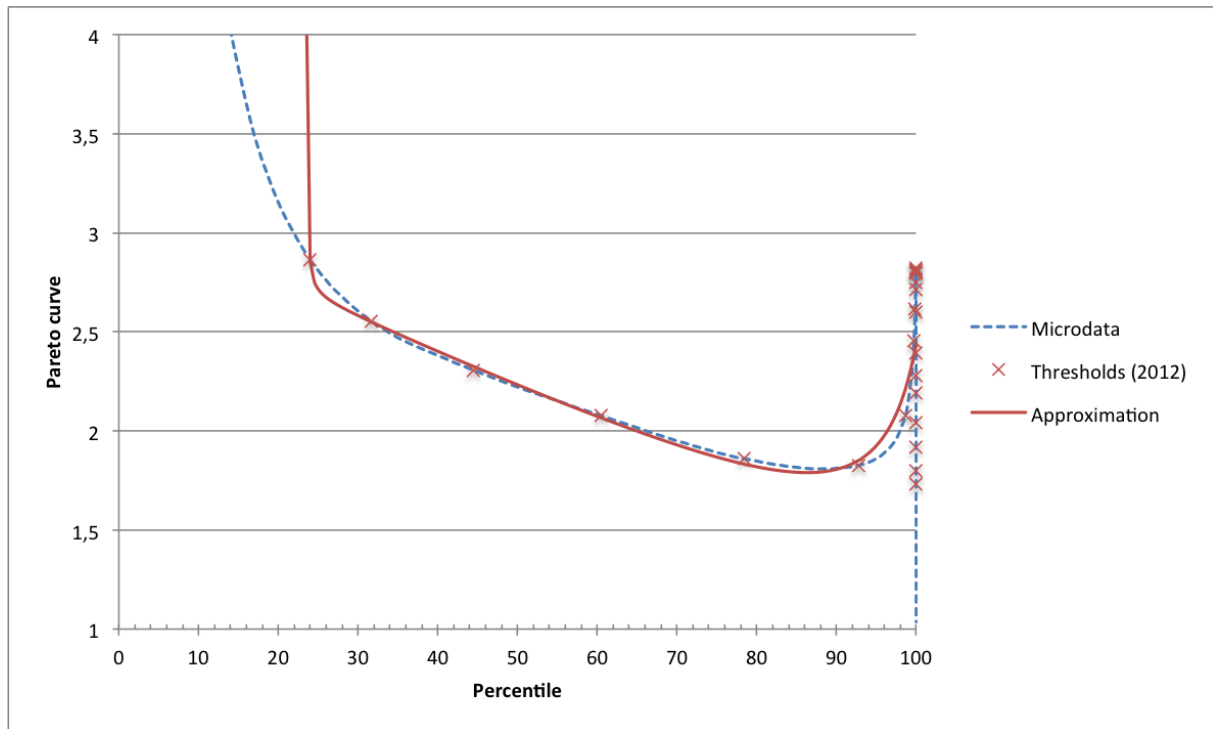
The term  $\frac{b}{p - p_{min}}$  tries to capture the asymptote at  $p_{min} \simeq 6\%$  while the term  $\frac{d}{1 + \varepsilon - p}$  is designed to fit the final increase around the top 1 percent. The linear term  $c \cdot p$  matches the global downward trend of the curve.

We provide the resulting approximation. They are obtained with non-linear regression of  $b_1, \dots, b_\omega$  on the chosen functional form. The microdata used were provided by tax authorities. They correspond to the French taxpaying population in the year 2006. The first figure stems from the thresholds in the fiscal scale of the year 2006, the second from the thresholds in the scale of the year 2012 that were more numerous at the top of the distribution.

In both cases, the approximating curves fit quite well the central part of the Pareto curve. Below the first threshold, neither of them manages to capture the shape around the vertical asymptote. The approximation in that part appears to randomly depend on the rest of the curve. In the top percentiles, the two approximations underestimate the rise of the empirical Pareto curve. That failing was expected with the 2006 tax data: how could we guess, given only the thresholds in the 2006 tax tabulation, that the empirical Pareto coefficient increases so



(a) Thresholds of the 2006 tax scale.



(b) Thresholds of the 2012 tax scale.

Figure B.1: Approximation of the generalized Pareto curve, France 2006

Functional form:  $b(p) = a + \frac{b}{p - p_{min}} + c \cdot p + \frac{d}{1 + \varepsilon - p}$ ,  $0 \leq p \leq 1$ . Spacing: 0.5%.

sharply in the very top? But with the 2012 scale, we have information on the variations of the curve until the last fractiles. The downward bias at the top of the curve is due to fact that the last points correspond to the final decrease of  $b(p)$ , which flaws the regression.

Here is the approximation obtained if we exclude top thresholds for which the Pareto curve goes down from the tabulation. The top of the curve is now quite well approximated.

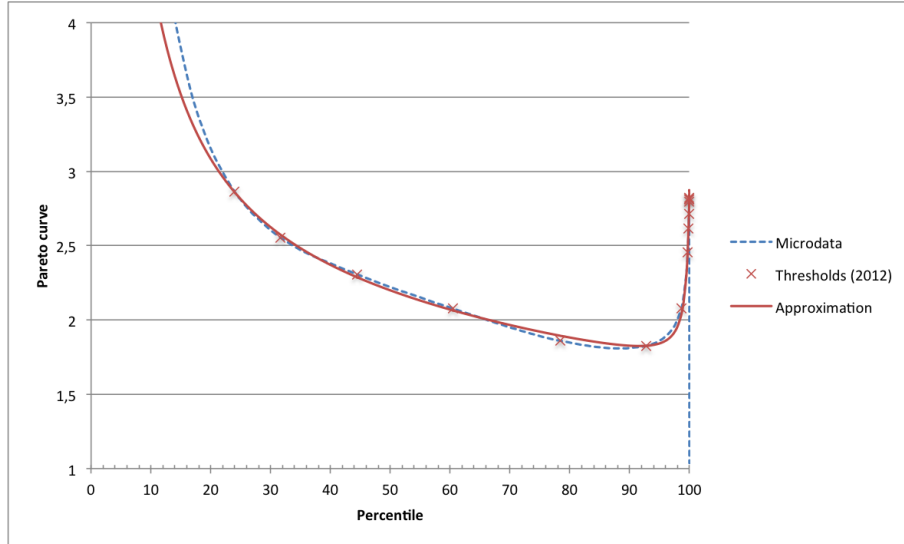


Figure B.2: Approximation of the generalized Pareto curve, France 2006

Thresholds of the 2012 tax scale, top thresholds excluded.

Functional form:  $b(p) = a + \frac{b}{p-p_{min}} + c \cdot p + \frac{d}{1+\varepsilon-p}$ ,  $0 \leq p \leq 1$ . Spacing: 0.5%.

Note that these approximations are not interpolations: the curves do not pass exactly through the thresholds points.

## B.2 Shape-preserving interpolation

The smoothing approach remained somehow unsatisfactory. The resulting approximations diverged from the empirical curve in places. As the functional form was not flexible enough, the curves in the bottom part of the distribution varied randomly depending on the thresholds elsewhere. Small changes in a given threshold could lead to large changes in the approximating curve. This method also necessitated at least 6 thresholds in the tax scale, so that the regression could be run. Moreover, the  $b_i$  found in tax tabulations are exact up to misreporting of incomes. Ideally, we would like our approximative function to pass through these points.

The interpolation approach gives much robust and stable results.

Formally, the interpolation problem is as follows. An unknown real-valued function  $f$  is defined on an interval  $[a, b]$ . Data is stored in tabular form  $(x_i, f_i)$  for  $i = 1, \dots, N$ , where  $f_i = f(x_i)$  and the points  $x_i$ , which are called the interpolation nodes, form an ordered sequence  $a = x_1 < x_2 < \dots < x_N = b$ . Typically, we are looking for a function  $P_N$  selected from a predetermined class of functions that passes through the data points so that for each  $i$ ,  $P_N(x_i) = f_i$ .

At first glance, it may seem easy to find such a function, as one could simply draw by hand a curve passing through a given set of data points. But finding a closed-form formula turns out to be more complex, especially if one wishes to obtain a visually pleasing curve (that is, if it is required for the interpolating function to keep to the overall shape of the data).

We review below the main interpolation methods. This will justify our choice of using shape-preserving cubic splines to interpolate generalized Pareto curves from the income tax tabulations. These techniques are thoroughly described by Kvasov [2000].

## B.2.1 Review of basic interpolation methods

### B.2.1.1 Polynomial interpolation

The traditional method to obtain a smooth interpolant is the construction of a polynomial  $P_N$  of order  $N + 1$  which goes through the nodes. The Lagrange formula gives the explicit solution:

$$P_N(X) = \sum_{j=0}^N f_j L_j(X) \quad (\text{B.2})$$

where the Lagrange coefficient polynomial  $L_j$  are, for  $j = 0, \dots, N$ ,

$$L_j(X) = \frac{(X - x_1) \dots (X - x_{j-1})(X - x_{j+1}) \dots (X - x_N)}{(x_j - x_1) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_N)} \quad (\text{B.3})$$

and satisfy

$$L_j(x_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{B.4})$$

However, for arbitrary nodes such that  $a \leq x_1 < \dots < x_N \leq b$ , the Lagrange polynomial interpolant does not in general converge uniformly to the unknown function  $f$  on the interval  $[a, b]$  as the number of nodes  $N$  grows to infinity. A famous example is given by Runge [1901]. The Lagrange polynomial interpolants do not converge uniformly to the function defined on  $[-1, 1]$  by  $f(x) = 1/(1 + 25x^2)$  if the nodes are equally spaced. In particular, polynomial interpolants appear to be bad approximation functions as oscillations can occur between nodes with high degree polynomials.

### B.2.1.2 Splines

As global interpolation by a unique polynomial does not guarantee convergence, interpolation methods which are used in practice involve piecewise polynomials. Spline interpolation avoids Runge's phenomenon, as different polynomials are defined within each interval.

**Linear interpolation** The simplest method is linear interpolation. If the coordinates of two consecutive known points are  $(x_i, f_i)$  and  $(x_{i+1}, f_{i+1})$ , the linear interpolant  $P$  is the straight line between these two points. For  $x$  in the interval  $[x_i, x_{i+1}]$ ,  $P(x)$  is given by:

$$P(x) = f_i \frac{x_{i+1} - x}{h_i} + f_{i+1} \frac{x - x_i}{h_i}, \quad (\text{B.5})$$

with  $h_i = x_{i+1} - x_i$ .

Linear interpolation on the set of points  $(x_1, f_1), \dots, (x_N, f_N)$  is obviously defined by the concatenation of linear interpolants between each pair of consecutive data points. The resulting curve is continuous, but not differentiable in general.

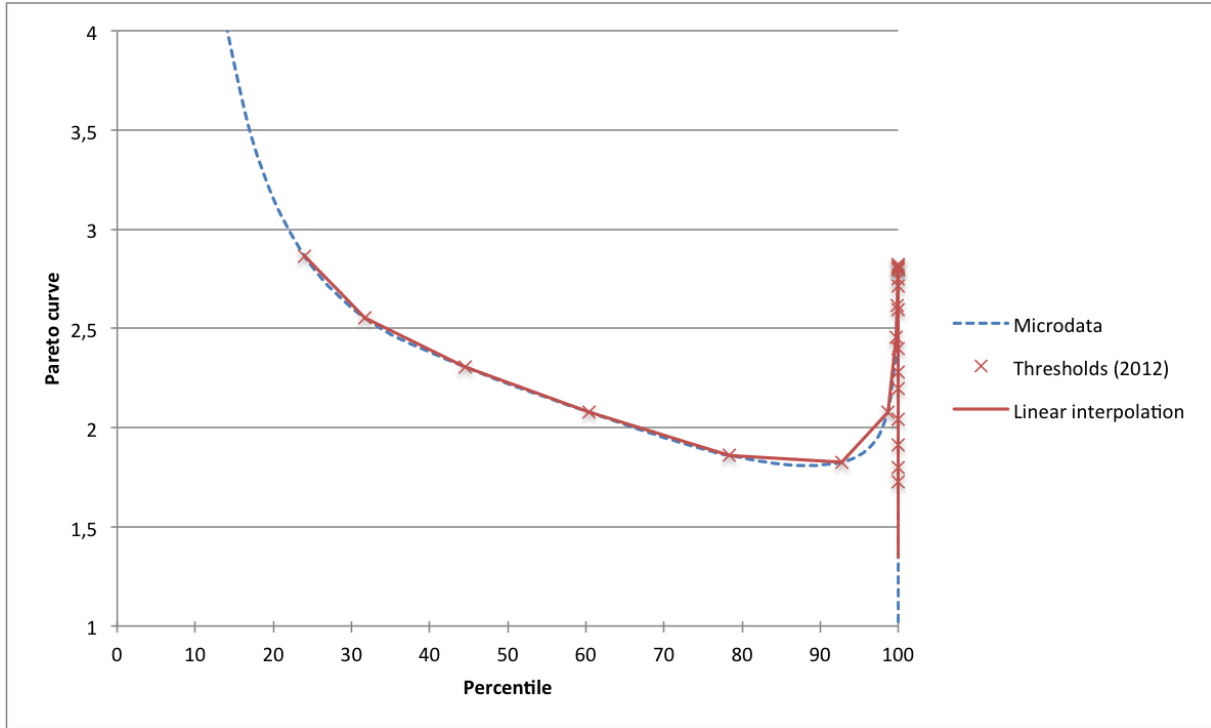


Figure B.3: Linear interpolation of the generalized Pareto curve  
Spacing: 0.5%.

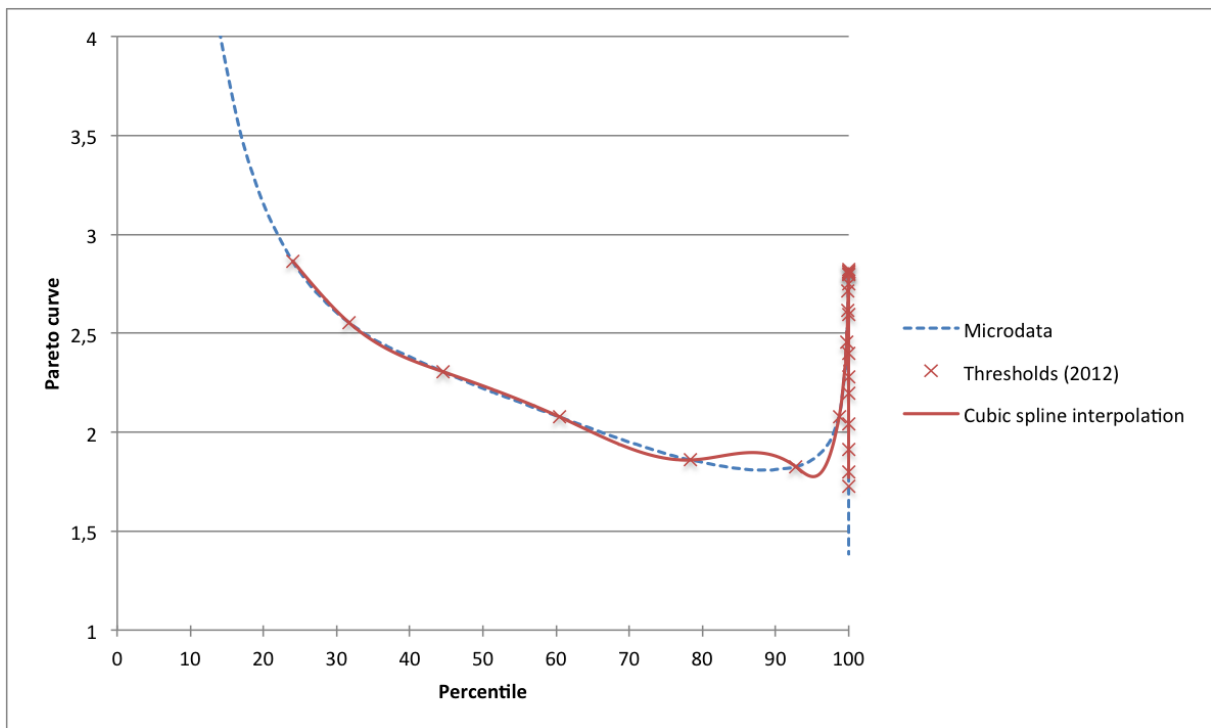


Figure B.4: Cubic spline interpolation of the generalized Pareto curve  
Spacing: 0.5%.

**Cubic splines** Even if linear interpolation provides quite satisfactory results, we see in figure B.3 that the interpolant should be more curved to follow accurately the Pareto curve. The

approximation is improved by replacing piecewise linear interpolants by cubic splines. On each interval, the function  $b$  is approximated by a polynomial of order 3. Additional conditions are given at the nodes to force both continuity and differentiability. The Stata routine `csipolate` based on the method introduced by Herriot and Reinsch [1973] provides a cubic spline interpolant.

Formally, the restriction of the interpolant  $P_N$  on each interval  $[x_{i-1}, x_i]$ ,  $i = 2, \dots, N$  is a polynomial of order 3  $S_i$  so that:

$$S_1(x_1) = f_1, \quad S_N(x_N) = f_N, \quad S_i(x_i) = S_{i+1}(x_i), \quad \forall 1 \leq i \leq N-1 \quad (\text{B.6})$$

and

$$S'_i(x_i) = S'_{i+1}(x_i), \quad \forall 1 \leq i \leq N-1. \quad (\text{B.7})$$

The second statement guarantees the differentiability of the interpolant  $F_N$ . This method is based on Hermite polynomials, i.e. polynomials that are determined by their values and the values of their derivatives at both ends of the intervals.

The curve obtained by interpolating the 2006 Pareto curve from the thresholds used for the 2012 tax scale is displayed below. Unfortunately, the interpolant exhibits a spurious behavior and does not respect the shape of the data above percentile 80.

**Shape-preserving methods** As we have seen, piecewise polynomial interpolants do not necessarily preserve the shape of the given data. To approximate accurately the Pareto curve, we would like the interpolant to be monotonic on intervals where the data is monotonic and to be convex where the data is convex.

Those geometric considerations translate into constraints that have to be satisfied by the derivatives of the Hermite polynomials splines at the data points<sup>1</sup>. Fritsch and Carlson [1980] give an algorithm to constrain the interpolating polynomials to meet the conditions that imply the desired properties (see also [Dougherty et al., 1989]). The interpolation method that we used to interpolate the Pareto curves is based on their work. The so-called *Piecewise Cubic Hermite Interpolating Polynomial* (PCHIP) provides visually pleasing interpolants, and has the advantage of being already implemented in Stata, with `pchipolate`, and in Matlab, with `pchip`.

The interpolant for the French Pareto curve (microdata 2006) with thresholds of the 2012 tax scale is displayed below.

---

<sup>1</sup>Two other approaches have been suggested in the literature. The first one consists in adding new mesh points so as to increase the number of polynomial splines [McAllister et al., 1977]. One can also construct shape-preserving interpolants by increasing the degree of the polynomials [Hyman and Larrouturou, 1982].

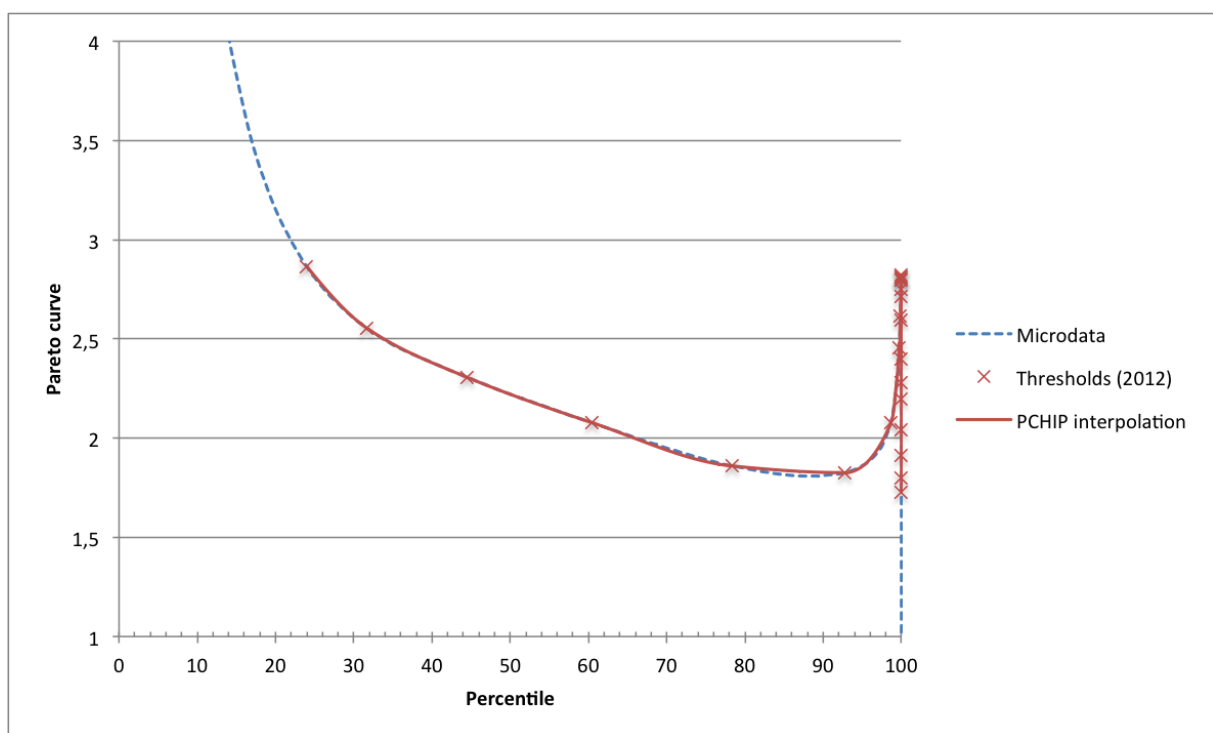
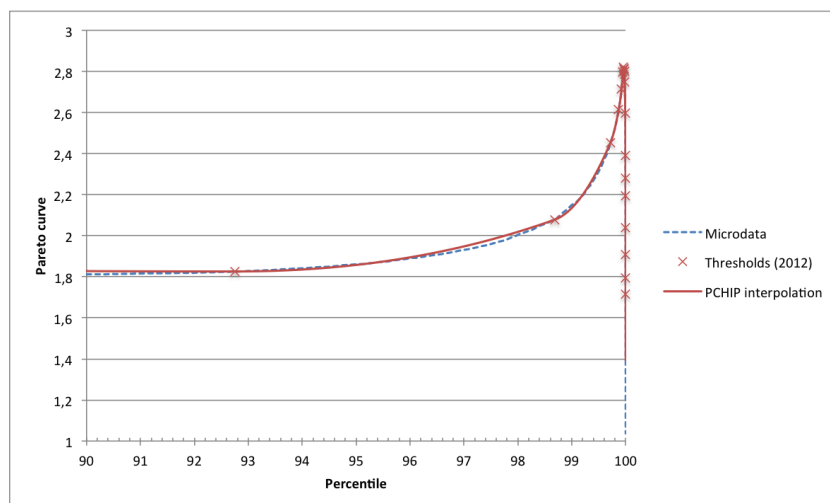
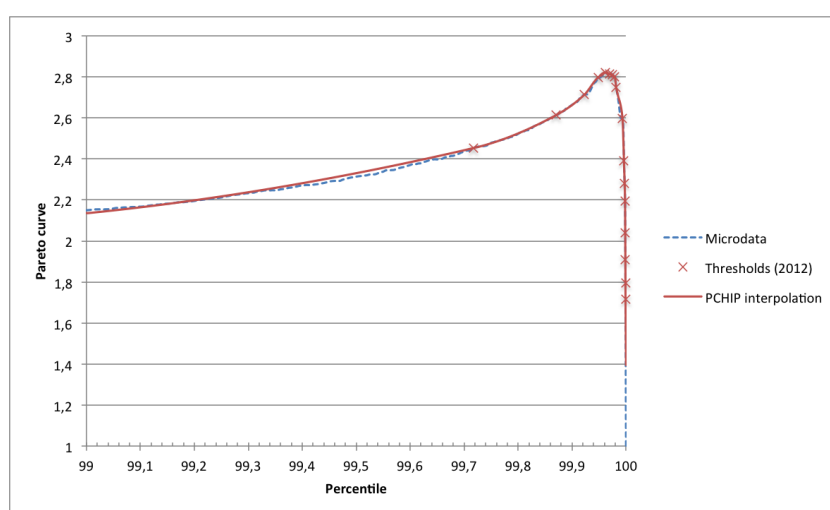


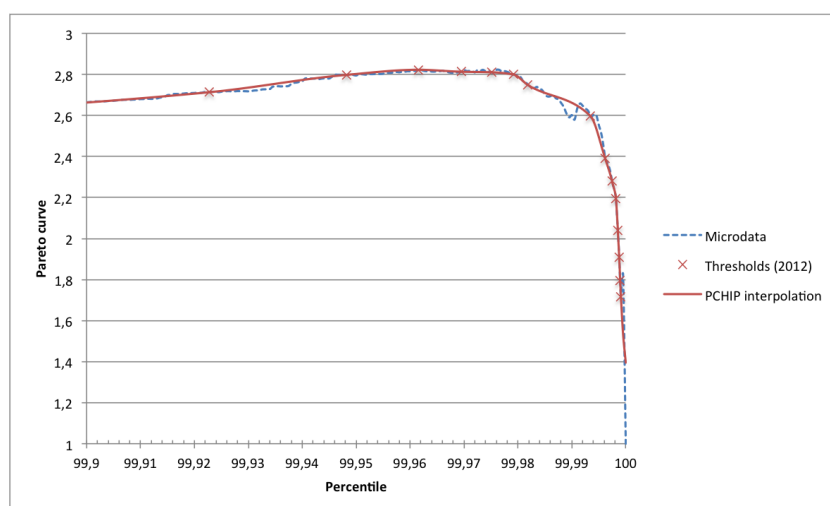
Figure B.5: PCHIP interpolation of the Pareto curve  
Spacing: 0.5%.



(a) Top 10%  
Spacing: 0.05%.



(b) Top 1%  
Spacing: 0.005%.



(c) Top 0.1%  
Spacing: 0.0005%.

Figure B.6: PCHIP interpolation of the Pareto curve - Zoom on top percentiles

## B.2.2 Piecewise cubic Hermite polynomial interpolation

The routines `pchipolate` in Stata and `pchip` in Matlab are both based on the algorithm developed by Fritsch and Carlson [1980]. These authors claim that their algorithm produces a "visually pleasing"  $\mathcal{C}^1$  monotone piecewise cubic interpolant.

We first examine the properties of the interpolants drawn under this procedure, and then detail the steps of the algorithm.

### B.2.2.1 Properties

Each spline of the PCHIP interpolant is a cubic Hermite polynomial. It is a polynomial of order 3 defined by its values and the values of its derivative at both ends of the interval.

Formally, to keep previous notations, the interpolant  $P$  satisfies:

$$P(x_i) = f_i, \quad i = 1, \dots, N. \quad (\text{B.8})$$

The spline  $S_i$  on each subinterval  $[x_i, x_{i+1}]$  can be represented as follows:

$$S_i(X) = f_i H_1(X) + f_{i+1} H_2(X) + d_i H_3(X) + d_{i+1} H_4(X) \quad (\text{B.9})$$

where  $d_j = S'_i(x_j)$ ,  $j = 1, 2$ , and the  $H_k(X)$  are the standard cubic Hermite basis functions for the interval  $[x_i, x_{i+1}]$ :

$$H_1(X) = \phi\left(\frac{x_{i+1} - X}{h_i}\right), \quad (\text{B.10})$$

$$H_2(X) = \phi\left(\frac{X - x_i}{h_i}\right), \quad (\text{B.11})$$

$$H_3(X) = -h_i \psi\left(\frac{x_{i+1} - X}{h_i}\right), \quad (\text{B.12})$$

and

$$H_4(X) = h_i \psi\left(\frac{X - x_i}{h_i}\right) \quad (\text{B.13})$$

with  $h_i = x_{i+1} - x_i$ ,  $\phi(X) = 3X^2 - 2X^3$ , and  $\psi(X) = X^3 - X^2$ .

The procedure constructs a visually pleasing interpolant in the sense that if  $f_i \leq f_{i+1}$  (respectively  $f_i \geq f_{i+1}$ ), the spline  $S_i$  is monotone increasing on  $[x_i, x_{i+1}]$  (respectively monotone decreasing). There are no extraneous overshoots, bumps or wiggles.

The problem of finding such an interpolant boils down for cubic Hermite splines to the construction of derivative values  $d_1, \dots, d_N$  such that each spline is monotone. In the paper [Fritsch and Carlson, 1980], necessary and sufficient conditions are derived that constrain interpolating splines to be monotone.

The PCHIP algorithm produces the values  $d_i$  using information on the neighboring points only. Consequently, this procedure is local: a single change in the data will affect the interpolant only in neighboring intervals. This property ensures stability of the interpolants.

### B.2.2.2 Description of the PCHIP algorithm

**Necessary and sufficient conditions for monotonicity** Fritsch and Carlson [1980] give the following characterization for monotonicity of the spline  $S_i$  on the interval  $[x_i, x_{i+1}]$ .

First, they notice that an obvious necessary condition for monotonicity is:

$$\text{sgn}(d_i) = \text{sgn}(d_{i+1}) = \text{sgn}(\Delta_i). \quad (\text{B.14})$$

**Proposition B.2.1**

Let  $\Delta_i = \frac{f_{i+1}-f_i}{h_i}$ ,  $\alpha_i = \frac{d_i}{\Delta_i}$  and  $\beta_i = \frac{d_{i+1}}{\Delta_i}$ .

- If  $\alpha_i + \beta_i - 2 \leq 0$ , then  $S_i$  is monotone on  $[x_i, x_{i+1}]$  if and only if (B.14) is satisfied.
- If  $\alpha_i + \beta_i - 2 > 0$  and (B.14) is satisfied, then  $S_i$  is monotone on  $[x_i, x_{i+1}]$  if and only if one of the following conditions is satisfied:
  - (i)  $2\alpha_i + \beta_i - 3 \leq 0$ ; or
  - (ii)  $\alpha_i + 2\beta_i - 3 \leq 0$ ; or
  - (iii)  $\alpha_i - \frac{1}{3} \frac{(2\alpha_i + \beta_i - 3)^2}{\alpha_i + \beta_i - 2} \geq 0$ .

Let  $\mathcal{M}$  be the region of all points  $(\alpha_i, \beta_i)$  that produce a monotone interpolant.

**Algorithm** Fritsch and Carlson [1980] advise the following two-step procedure.

1. Initialization of the derivatives  $d_i$ ,  $i = 1, \dots, N$  so that  $\text{sgn}(d_i) = \text{sgn}(d_{i+1}) = \text{sgn}(\Delta_i)$ . If  $\Delta_i = 0$ , set  $d_i = d_{i+1} = 0$ . Otherwise, authors suggest to use the three-point difference formula:

$$d_i = \frac{d_{i+1} - d_i}{2(x_{i+1} - x_i)} + \frac{d_i - d_{i-1}}{2(x_i - x_{i-1})}. \quad (\text{B.15})$$

$$(\text{B.16})$$

2. For each interval  $[x_i, x_{i+1}]$  in which the  $(\alpha_i, \beta_i) \notin \mathcal{M}$ , modify  $d_i$  and  $d_{i+1}$  to  $d_i^*$  and  $d_{i+1}^*$  defined by:

$$\alpha_i^* = \tau_i \alpha_i, \quad \beta_i^* = \tau_i \beta_i \quad (\text{B.17})$$

where  $\tau_i = 3(\alpha_i^2 + \beta_i^2)^{-1/2}$ .

## B.3 Extrapolation

### B.3.1 Lower incomes

In this section, we try to approximate the Pareto curve below the lowest threshold in the tax scale. The resulting extrapolation is no more than a crude approximation: it is kind of a challenge to assess the Pareto curve of the nontaxable population from tax data.

The bottom part of the Pareto curve cannot be directly assessed using splines. Indeed, the curve  $b(p)$  diverges toward infinity for lower incomes. We present below an attempt to extrapolate this fraction of the Pareto curve.

**First step: Estimating  $L(p)$ ,  $p \leq p_1$**  The idea is to approximate the Lorenz curve  $L$  instead of the Pareto curve. Indeed, we know that  $L(p_0) = 0$ , and we can estimate the value of  $L(p_1)$

using tax tabulation and the average income  $\bar{y}$ :

$$L_1 = L(p_1) = 1 - \frac{(1 - p_1)b_1\theta_1}{\bar{y}}, \quad (\text{B.18})$$

where  $b_1 = b(p_1)$  is given.

In fact, we can also estimate  $L'(p_1)$ : one can check that the expression (3.10) is equivalent to:

$$L'_1 = L'(p_1) = \frac{1 - L_1}{(1 - p_1)b_1}. \quad (\text{B.19})$$

We will interpolate  $L$  on the interval  $[p_0, p_1]$  with a Hermite cubic spline. We want this spline to satisfy:

$$\begin{cases} L(p_0) &= 0 \\ L'(p_0) &= 0 \\ L(p_1) &= L_1 \\ L'(p_1) &= L'_1 \end{cases} \quad (\text{B.20})$$

The cubic polynomial satisfying this set of conditions can be expressed in the standard cubic Hermite basis defined before as:

$$L(X) = L_1 H_2(X) + L'_1 H_4(X), \quad (\text{B.21})$$

where

$$H_2(X) = \phi\left(\frac{X - p_0}{h_0}\right), \quad (\text{B.22})$$

$$H_4(X) = h_0 \psi\left(\frac{X - p_0}{h_0}\right), \quad (\text{B.23})$$

with  $h_0 = p_1 - p_0$ ,  $\phi(X) = 3X^2 - 2X^3$ , and  $\psi(X) = X^3 - X^2$ .

We now have to determine  $p_0$ . We know that the Lorenz curve is increasing convex. Dougherty et al. [1989] give the following requirement to ensure convexity of the spline:

$$-2(L'_1 - s_0) \leq L'_0 - s_0 \leq -\frac{1}{2}(L'_1 - s_0), \quad (\text{B.24})$$

with  $s_0 = \frac{L_1 - L_0}{p_1 - p_0}$ .

This condition becomes:

$$p_1 - \frac{5}{2} \frac{L_1}{L'_1} \leq p_0 \leq p_1 - \frac{3}{2} \frac{L_1}{L'_1}. \quad (\text{B.25})$$

We can thus choose  $p_0$  in this range of values and obtain an increasing convex spline on the interval  $[p_0, p_1]$ .

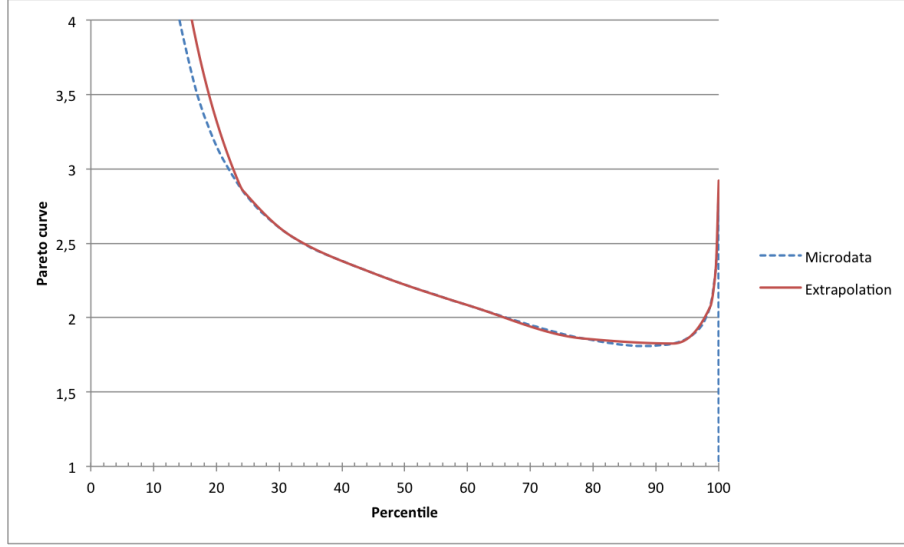


Figure B.7: Extrapolation of the lower part of the Pareto curve, France 2006  
Spacing: 0.5%. Source: Micro-files provided by tax authorities.

**Second step: Estimation of  $b(p)$  for  $p \leq p_1$**  Recall that we can express  $b$  using  $L$  with the formula derived in section 3.1:

$$\forall p \in [0, 1], \quad b(p) = \frac{1 - L(p)}{(1 - p)L'(p)}. \quad (\text{B.26})$$

This formula gives an estimation of the Pareto curve. The condition  $L'(p_0) = 0$  guarantees that  $b(p)$  grows to infinity as  $p$  approaches  $p_0$ .

A shortcoming of this method is that it does not in theory ensures that the shape of the resulting Pareto curve is satisfactory.

Another possibility is to add a point at  $p = 50\%$ : indeed, the share of income accruing to the bottom half of the population, as well as the ratio of median income over mean income, have remained quite stable throughout the XX<sup>th</sup> century.

### B.3.2 Top of the distribution

Another part of the distribution where PCHIP interpolation sometimes fails to provide good approximation of the Pareto curve lies in the last percentiles. Finding an accurate extrapolation above the higher threshold in the tax scale is all the more important as small errors in the approximation of  $b(p)$  for  $p$  near 1 lead to large discrepancies in predicted estimates. Tax tabulations give little guidance about the shape of the Pareto curve about this point, and specifically do not indicate the maximum value reached by the Pareto curve.

Future work will consist in incorporating empirical knowledge about the form and the values of the Pareto curve for top incomes in the extrapolation technique used. For instance, one could put an additional fictive point (say, at  $p = 99.99\%$ ) into the data interpolated to force the curve to rise sufficiently high.

Another avenue to explore is the interpolation by tension splines. Tension splines, initiated by Späth [1969], are a generalization of cubic splines that give the possibility to choose on each subinterval defined by interpolated data a tension factor, that determines how much the spline between two points is tight.

# Appendix C

## Simulation of synthetic micro-files

In section 3, we have described a method which permits to assess the quantile function from the tax tabulations data. Here, we show how to make use of these results to generate reliable synthetic micro-files representing the whole taxpaying population.

### C.1 Simulation of a population using tax tabulations

Taking as an input tabulated income data, we want to simulate a synthetic micro-file of a given number  $N$  of incomes that are distributed just as the incomes of the population of taxpayers. The code described below has been written to run on the software Matlab.

#### C.1.1 The inversion method

The generation of non-uniform continuous random variables is based on the following statement, given in [Devroye, 1986].

##### **Proposition C.1.1** (*The inversion principle*)

Let  $F$  be a continuous distribution function on  $\mathbb{R}$  with inverse  $Q = F^{-1}$ , called the quantile function, defined by

$$Q(u) = \inf\{y \in \mathbb{R} : F(y) = u, 0 < u < 1\}. \quad (\text{C.1})$$

If  $U$  is a uniform  $[0, 1]$  random variable, then  $Q(U)$  has cumulative distribution function  $F$ . Also, if  $Y$  has distribution function  $F$ , then  $F(Y)$  is uniformly distributed on  $[0, 1]$ .

PROOF:

- For all  $y \in \mathbb{R}$ ,

$$\begin{aligned} \mathbb{P}(Q(U) \leq y) &= \mathbb{P}(\inf\{z \in \mathbb{R} : F(z) = U\} \leq y) \\ &= \mathbb{P}(U \leq F(y)) \\ &= F(y). \end{aligned}$$

Therefore,  $Q(U)$  has CDF  $F$ .

- For all  $u \in (0, 1)$ ,

$$\begin{aligned}\mathbb{P}(F(Y) \leq u) &= \mathbb{P}(Y \leq Q(u)) \\ &= F(F^{-1}(u)) \\ &= u\end{aligned}$$

because  $F$  is continuous.

□

Numerical analysis softwares such as Matlab have built-in pseudorandom number generators that provide random numbers uniformly distributed on the interval  $[0, 1]$ . Thus, provided that the quantile function is explicitly known, we are able to generate random variables that follow a given distribution.

### C.1.2 Matlab code

Our code to generate micro-files of incomes distributed as the taxpaying population breaks down into three programs:

- the subroutine `pc.m` that interpolates the Pareto curve using tax tabulation data;
- the routine `quantile.m` that approximates the quantile function using the Pareto curve estimated with `pc.m`;
- the main part of the code `gen_pop.m` that uses the approximation of the quantile function provided by `quantile.m` to generate the population.

To generate a file of say 35 millions of incomes, the program will run for about 1 hour.

#### C.1.2.1 Program `gen_pop.m`

`gen_pop.m` is the main program. It imports tax tabulation data from an Excel spreadsheet, calls the Matlab function `quantile.m` to approximate the quantile function, and generates the synthetic micro-file by applying the inversion principle C.1.1.

#### Description of the code

**Input** An Excel spreadsheet workbook named `taxdata.xls` has to be placed in a subfolder, entitled "data", of the folder containing the Matlab program. The tabulated income data needs to have the following form.

In the worksheet named "Tabulation", a 3-column table has to give:

- the thresholds in a first column "thr";
- the corresponding percentiles in a second column "perc";
- and the corresponding coefficient  $b$  in a third column "b".

In another worksheet "Average income", the average income of the population has to be written.

thr	perc	b
9400	0,268757	3,01579
11250	0,335887	2,68161
13150	0,407262	2,45805
15000	0,482853	2,33289
16900	0,548446	2,2345
18750	0,602345	2,15837
23750	0,702708	1,97903
28750	0,780623	1,8927
38750	0,877692	1,83887
48750	0,927604	1,85842
97500	0,98616	2,08967

Table C.1: Input of the Matlab program - Worksheet "Tabulation"

$y_{av}$
21930,41

Table C.2: Input of the Matlab program - Worksheet "Average income"

**Output** The synthetic micro-file is a file named `sim_pop.txt` that is created in the folder where the Matlab program is. Each line corresponds to a taxpayer. To import it in Stata, click File, Import, Text data (delimited, \*.csv, ...), and select the Text file.

### Steps of the program

1. First, the path to access tabulated data is indicated to the program. The user has also to specify `Npop`, the size of the simulated population.
2. Then, the program loads the tabulated data from the file `taxdata.xls`. The shortcut `loadpath` gives the emplacement of this spreadsheet. The thresholds are stored in the table `tab.thr`, the percentiles in `tab.p` and the Pareto coefficients in `tab.b`. `T` is the total number of thresholds in the tax scale considered. `y_av` is the mean income in the population.
3. A vector `u` of `Npop` numbers uniformly distributed on the interval  $[0, 1]$  is generated. We sort the incomes stored in `u` in an increasing order.
4. The main program calls the routine `quantile.m` to recover the quantile function. The function `quantile.m` interpolates the quantile function corresponding to tax data at the points of vector `u`. It returns the vector `sim_pop` of `Npop` random numbers which are distributed as the incomes of the population of taxpayers.
5. The last step is the writing of the output file `sim_pop.txt`.

**Code** Here is the code of the program `gen_pop.m`.

```
% Simulation of a population of taxpayers using income tax data
%
% Npop=34546115 => about 1 hour
%
```

```

% Steps of the program:
% * 1. Preliminaries
% * 2. Loads data
% * 3. Generates a vector of random numbers uniformly distributed on [0,1]
% * 4. Simulates the population
% * 5. Exports the population
%
%% 1. Preliminaries
close all; clear all;
currentfolder=pwd;
% Path to load data
loadpath=[currentfolder, '/data'];

% Number of taxpayers to be generated
Npop = 34546115;

%% 2. Loads data
disp('Loads data');

cd(loadpath);
[tabulation] = xlsread('taxdata_pchip', 'Tabulation'); %thresholds
tab.thr = tabulation(:,1); %thresholds
tab.p = tabulation(:,2); %percentiles
tab.b = tabulation(:,3); %inverted Pareto coefficients
T = length(tab.thr); %number of thresholds in the tax tabulation

[y_av] = xlsread('taxdata_pchip', 'Average income'); %average income

clear tabulation;
cd(currentfolder);

%% 3. Generates a vector of random numbers uniformly distributed on [0,1]
disp('Generates a vector of random numbers uniformly distributed on [0,1]');

u = rand(Npop,1); %vector of Npop random numbers uniformly drawn on [0,1]
u = sort(u); %sorts the coefficients of u

%% 4. Simulates the population
disp('Simulates the population');

[sim_pop, p0] = quantile(tab.thr, tab.p, tab.b, y_av, u);

%% 5. Exports the simulated population
disp('Exports the simulated population');

dlmwrite('sim_pop.txt', sim_pop, 'precision', 9);

```

### C.1.2.2 Subroutine quantile.m

The function `quantile.m` approximates the quantile function of a distribution corresponding to tax tabulation data received as an input. To do so, it calls the subroutine `pc.m` that provides an estimation of the Pareto curve.

### Description of the code

**Inputs** The following inputs have to be specified to the Matlab function `quantile.m`.

- A vector **thr** of incomes corresponding to the thresholds in the tax tabulation.
- A vector **pp** of percentiles (in  $[0, 1]$ ) corresponding to the thresholds in the tax tabulation.
- A vector **bb** of (inverted) Pareto coefficients corresponding to the thresholds in the tax tabulation.
- The average income of the taxpaying population **y\_av**.
- A table **N** of nodes in  $[0, 1]$ .

## Outputs

- A vector **Q** of approximative values taken by the quantile function  $Q$  at the points of **N**, that is,  $Q=Q(N)$ .
- The approximation **p0** of the share of the population with no income.

## Steps of the program

1. Definition of the Matlab function **quantile**. **T** is the number of thresholds in the tax tabulation. **Xn** is the number of points in the input vector **N**. **a** and **b** are the two extremities of the interval defined by the points of **N**.

**Ym** is the number of points for the were the integral appearing in the expression 3.5 of  $Q$  will be calculated. It determines the precision of the results. **mesh** is a mesh of **Ym** points of the interval  $[a, b]$ . **mid** is the vector of midpoints of the subintervals bracketed by the points of **mesh**.

2. **quantile.m** calls the function **pc.m** to obtain the approximation of the Pareto curve at the points of vector **mid**. These values are stored in vector **B**.
3. Numerical integration.

The vector **integrand** contains the values at the points of **mid** of the application  $p \mapsto \frac{1}{(1-p)b(p)}$  that is integrated in the expression 3.5 of  $Q$ .

The next step is to numerically integrate the function defined by **integrand** using the rectangle method. **h** is the vector containing the intervals between the consecutive nodes of **mesh**. The vector **int** which contains the approximated values of the integral at midpoints is computed as the cumulative sum:

$$\mathbf{int}_k = \sum_{i=1}^k h_i \cdot \mathbf{integrand}(\mathbf{mid}_i). \quad (\text{C.2})$$

4. The vector **temp0** gives, up to scalar multiplication, the values of the  $Q$  at the midpoints of the intervals (i.e. at points of vector **mid**). These values are interpolated using the Matlab function **interp1** at the points of vector **N**.

`q0` is the vector of values taken by `temp` at the percentiles of the tax scale (which are stored in `pp`). They are obtained by interpolating with the Matlab routine `interp1`.

The columns of matrix `Q_mat` correspond to numerical approximations of the quantile function  $Q$  respectively starting from each threshold of the tabulation. That is, for the column  $i$ , the interpolated value of  $Q$  at  $p_i$  is exactly the  $i$ th threshold of the tax scale `thr_i`.

5. The last step is the approximation of the quantile function  $Q$  based on the formula 3.5. The values calculated will be stored in the table `Q`.

We compute the vectors `Q`, which gives the approximated values of  $Q$  at the points of `mid`, as the weighted sum of the two approximations of  $Q$  passing through the thresholds bounding the bracket where each point lies. If the point is below the lower threshold (respectively above the higher threshold), we use the quantile function starting from the first threshold (respectively the last threshold).

**Code** Here is the code of the program `quantile.m`.

```
%% Estimation of the quantile function Q(p) using income tax data
%
% Inputs:
% * thr: table of incomes corresponding to the thresholds in the tax
% tabulation
% * pp: table of percentiles (in [0,1]) corresponding to the thresholds in the tax
% tabulation
% * bb: table of (inverted) Pareto coefficients corresponding to the
% thresholds in the tax tabulation
% * y_av: average income of the taxpaying population
% * N: table of nodes p (in [0,1])
%
% Outputs:
% * Q: table of approximative values taken by the quantile function Q at
% the points of table P, that is, Q=Q(N)
% * p0: approximation of the share of the population with no income
%
% Steps of the program:
% 1. Preliminaries
% 2. Calls function pc.m to estimate the Pareto curve
% 3. Numerically integrates the integral that appears in the formula of Q
% 4. Computes a temporary estimation of Q up to a constant
% 5. Adjusts the estimations of Q and clips it to the tax thresholds
%
%% 1. Preliminaries
% Defines function inputs and outputs.

function [Q,p0] = quantile(thr,pp,bb,y_av,N)

T = length(thr);
Xn = length(N);
a = min(N);
b = max(N);

Ym = floor((b-a)*200000000);
```

```

mesh = linspace(a,b,Ym)';
mid = (mesh(1:Ym-1,1)+mesh(2:Ym,1))/2;

%% 2. Estimates the Pareto curve
[B,p0] = pc(thr,pp,bb,y_av,mid);

%% 3. Numerical integration
h = mesh(2:Ym)-mesh(1:Ym-1);

integrand = 1./((1-mid).*B);
int = cumsum(h.*integrand);

%% 4. Computes a temporary estimation of Q up to a constant
temp0 = (1./(B.*(1.-mid))).*exp(-int);

temp = interp1(mid,temp0,N,'spline','extrap');
clear temp0

q0(1,:) = interp1(N,temp,pp(:,1),'spline','extrap');

Q_mat = temp*(thr'./q0);
% Matrix which columns corresponds to numerical approximations of the
% quantile function Q respectively starting from each threshold of the
% tabulation

%% 5. Adjusts the estimations of Q and clips it to the tax thresholds
Q = zeros(Xn,1);
% Vector which will contain the numerical approximation of the final
% quantile function Q

%within each bracket, the final value Qfin is a weighted average of the
%quantile functions starting from the two thresholds bounding the bracket
br=0; %bracket
for i=1:Xn
    if (br<T)&&(N(i,1)>pp(br+1,1))
        br = br+1;
    end
    if N(i,1) <= p0
        Q(i,1) = 0;
    elseif br==0
        Q(i,1) = Q_mat(i,1);
        %if below the first threshold, use the quantile function starting
        %from the first threshold
    elseif br==T
        Q(i,1) = Q_mat(i,T);
        %if after the last threshold, use the quantile function starting
        %from the last threshold
    else
        w1 = (pp(br+1,1)-N(i,1))/(pp(br+1,1)-pp(br,1));
        w2 = (N(i,1)-pp(br,1))/(pp(br+1,1)-pp(br,1));
        Q(i,1) = w1*Q_mat(i,br)+w2*Q_mat(i,br+1);
        %else, use a weighted average of the two surrounding thresholds
    end
end

clear Q_mat

disp('Quantile function estimated')

```

### C.1.2.3 Subroutine `pc.m`

The Matlab function `pc.m` approximates the Pareto curve of a distribution corresponding to tax tabulation data received as an input.

#### Description of the code

**Inputs** The following inputs have to be specified to the Matlab function `pc.m`.

- A vector `thr` of incomes corresponding to the thresholds in the tax tabulation.
- A vector `pp` of percentiles (in  $[0, 1]$ ) corresponding to the thresholds in the tax tabulation.
- A vector `bb` of (inverted) Pareto coefficients corresponding to the thresholds in the tax tabulation.
- The average income of the taxpaying population `y_av`.
- A table `P` of points in  $[0, 1]$  of nodes.

#### Outputs

- A vector `B` of approximative values taken by the Pareto curve  $b$  at the points of `P`, that is,  $B=b(P)$ .
- The approximation `p0` of the share of the population with no income.

#### Steps of the program

1. Definition of the Matlab function `pc`.
2. Interpolation of the Pareto curve from vectors inputs `pp` (percentiles) and `bb` (Pareto coefficients) using the PCHIP method. The spline approximating  $b(p)$  is stored in `interp_b`. The values that takes  $b$  at the points of vector `P` are stored in vector `B_interp`. The Matlab function `ppval` returns the value of the piecewise polynomial at desired entries.
3. The program then evaluate the fraction `p0` of the population with zero income and the Pareto curve below the lower threshold in the tax scale `p1`. It applies the extrapolation method detailed in appendix B. That is, it extrapolates the Lorenz curve with a cubic Hermite polynomial `L_po1` using the information on the shares obtained with the tabulation and the average income. `p0` is defined as  $p_1 - 2\frac{L_1}{L_1'}$ . The parameter 2 can be adjusted to obtain a visually pleasing extrapolation of  $b$ . It has to lie in the interval  $[\frac{3}{2}, \frac{5}{2}]$ .

The approximations of values taken by  $b$  below  $p_1$  are stored in vector `b_extrap`.

After extrapolating the Lorenz curve and the Pareto curve with this temporary value of  $p_0$ , `p0` is set equal to 0 if its first estimation was negative.

4. The program merges the two tables `B_interp` and `B_extrap` to obtain the entire approximation of  $b$  stored in vector `B`.
5. Last, the program plots the interpolated Pareto curve.

**Code** Here is the code of the program `pc.m`.

```

%% Estimation of the generalized Pareto curve b(p) using income tax data
%
% Method: PCHIP interpolation
% This routine extrapolates the lower part of the distribution.
%
% Inputs:
% * thr: table of incomes corresponding to the thresholds in the tax
% tabulation
% * pp: table of percentiles (in [0,1]) corresponding to the thresholds in the tax
% tabulation
% * bb: table of (inverted) Pareto coefficients corresponding to the
% thresholds in the tax tabulation
% * y_av: average income of the taxpaying population
% * P: table of points p (in [0,1]) where b(p) has to be calculated
%
% Outputs:
% * B: table of approximative values taken by b at points of the input
% table P, that is, B=b(P)
% * p0: estimation of the share of the population with no income
%
% Steps of the program:
% 1. Preliminaries
% 2. Interpolation of the Pareto curve above the lower threshold using
% PCHIP method
% 3. Extrapolation of the Pareto curve below the lower threshold
% 4. Merges the tables B_interp and B_extrap
% 5. Plots the Pareto curve
%
%% 1. Preliminaries
% Defines function inputs and outputs.

function [B,p0] = pc(thr,pp,bb,y_av,P)

%% 2. Interpolation

interp_b = pchip(pp,bb);

B_interp = ppval(interp_b,P);

%% 3. Extrapolation of b(p) for lower incomes

p1 = pp(1,1);
t1 = thr(1,1);
b1 = bb(1,1);

l1 = 1-((1-p1)*b1*t1)/y_av;
l1_pr = (1-l1)/((1-p1)*b1);

p0=p1-2*(l1/l1_pr);

```

```

syms x
x1 = (x-p0)/(p1-p0);
x2 = (x-p0)/(p1-p0);

syms y
phi = 3*y^2-2*y^3;
psi = y^3-y^2;

syms z
L_pol = l1*compose(phi,x1,y,x,z)+l1_pr*(p1-p0)*compose(psi,x2,y,x,z);
L = sym2poly(L_pol);

L_pr = polyder(L);

B_extrap = (1-polyval(L,P))./((1-P).*(polyval(L_pr,P)));

if p0<0
    p0=0;
end

%% 4. Merges the tables B_interp and B_extrap
B = B_interp;
i=1;

while P(i,1)<=p0
    B(i,1) = 1;
    i=i+1;
end

while P(i,1)<=p1
    B(i,1) = B_extrap(i,1);
    i=i+1;
end

disp('Pareto curve estimated')

%% 5. Plots the Pareto curve

xx = linspace(0,1,1000);
yy = interp1(P,B,xx,'linear');

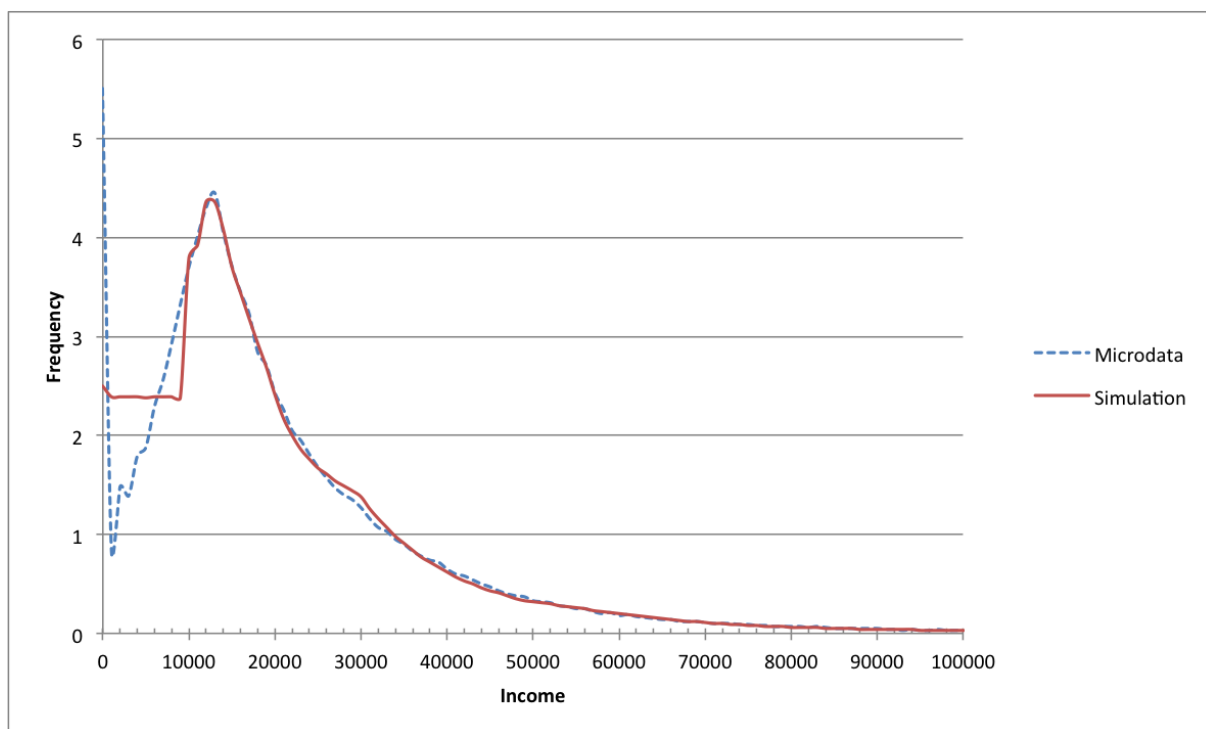
plot(xx,yy)
hold on
scatter(pp,bb)
title('Interpolated Pareto curve');

```

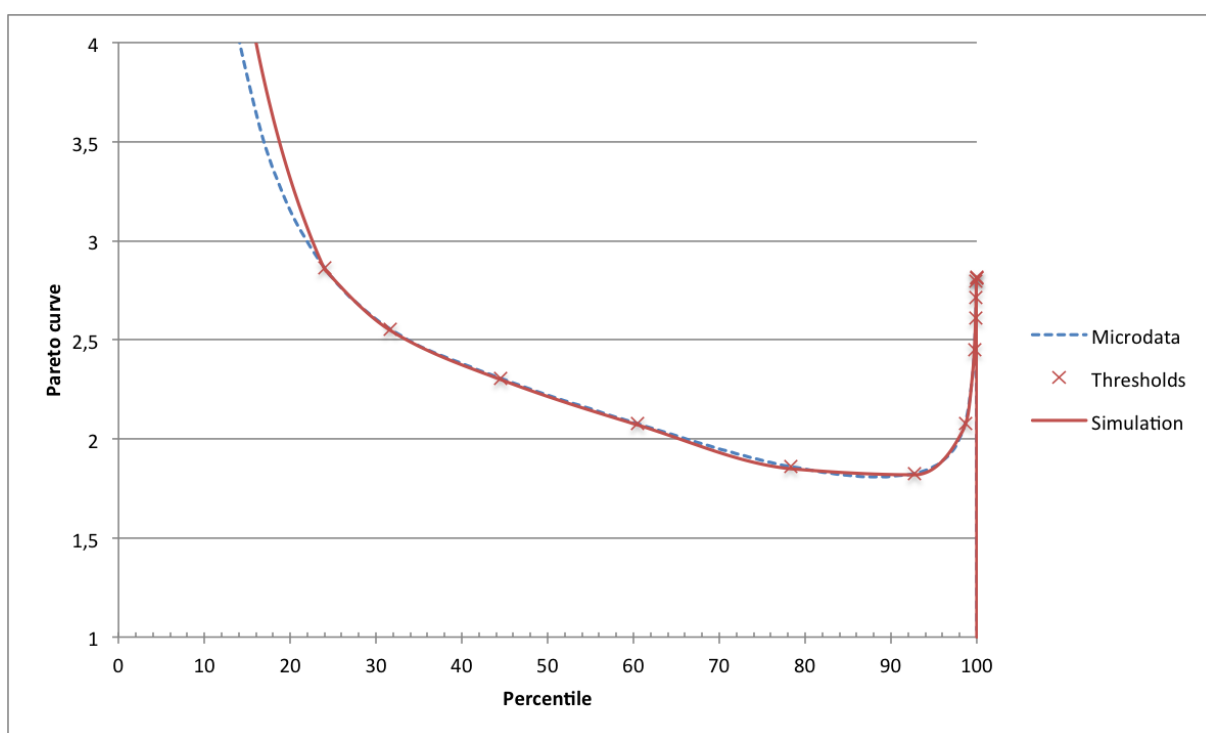
## C.2 Comparison of the results with microdata

The graphs below compare frequency distributions, Pareto curves and Lorenz curves for microdata and simulated population. The population generated with our program is very similar to microdata. The mean income of the simulated population is 22620 euros, while the mean income in microdata is 22974 euros.

As we can see in graph C.1a, the distribution is well-approximated, except for the lower incomes. That makes sense, as we do not have information about this part of the population in tax data.



(a) Frequencies



(b) Pareto curve

Figure C.1: Comparison of simulated population and microdata

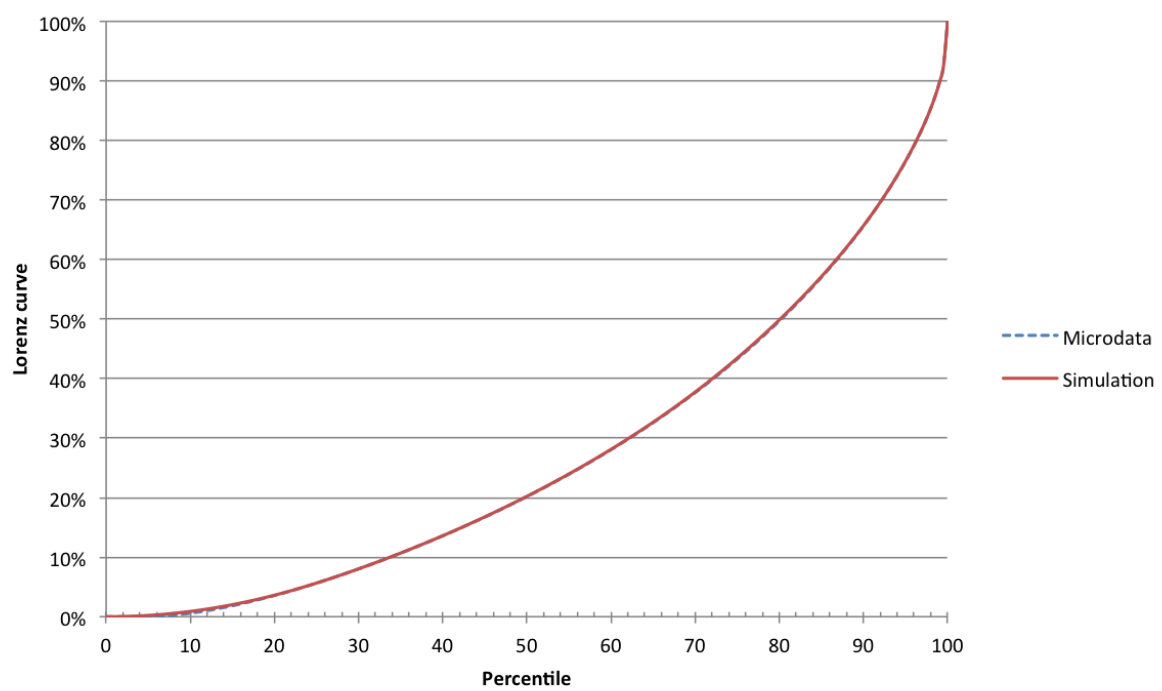


Figure C.2: Comparison of simulated population and microdata - Lorenz curve

## Appendix D

# From households to individuals: correcting for the variations in tax units

### D.1 Homogenization of series across countries: the problem of the changes in tax units

As underlined by Atkinson and Piketty [2007, Chapter 2, "Measuring Top Incomes: Methodological Issues", pp. 18-42] and by Atkinson et al. [2011], one measurement issue that hinders comparability across countries is the variability in the definition of the tax unit. If in some countries such as Australia, Canada, and New Zealand or United Kingdom in recent years, the tax unit is the individual, in many others the definition is based on the household. If, in order to compare carefully the distributions across countries, we want to move from a family-based tax system to an individual-based one, we have to know the joint distribution of income for couples.

#### D.1.1 The problem

Even if we arbitrarily decide to split incomes of households half-and-half and to assign half of the total income earned by the couple each spouse, estimation methods used until now in the literature could not provide approximations of resulting distribution of individual earnings. Indeed, some individuals would move to lower tax brackets after this sharing out of household income. These changes are unmanageable with piecewise estimation methods.

#### D.1.2 Method to correct for changes in tax units

A solution would be to run two simulations of the population one after the other: the first one with tax data corresponding to singles, the second one with tax data corresponding to couples. Indeed, we can generally find such informations, as in family-based tax systems households generally benefit from deductions depending on the size of the family.

Here, we estimate the quantile function on two vectors of points evenly spaced on the interval  $[0, 1]$  which respective sizes are exactly the number of singles and the number of persons in couple

in the population. Then, merging those two vectors, we obtain a new vector corresponding to the income distribution of the population of individuals.

Then, we can use Stata to obtain approximations of thresholds and average income of different groups of the population.

## D.2 Matlab code

### D.2.1 Description of the code

The program `tax_unit.m` imports tax tabulation data from two Excel spreadsheets, calls twice the Matlab function `quantile.m` to approximate the quantile function for singles and for individuals, and returns a vector describing the population of individuals. The program should take approximately 1 hour.

**Inputs** Two Excel spreadsheet workbook named `singles.xls` and `couples.xls` have to be placed in a subfolder, entitled "data", of the folder containing the Matlab program. The tabulated income data has to be presented as the data in `taxdata.xls`, the input of the program generating micro-files `gen_pop.m`.

In the worksheet named "Tabulation", a 3-column table has to give respectively for singles and couples:

- the thresholds in a first column "thr";
- the corresponding percentiles in a second column "perc";
- and the corresponding coefficient  $b$  in a third column "b".

In another worksheet "Average income", the average income of the singles and the couples has to be written.

Finally, in a worksheet "Number", the respective numbers of singles and couples in the population have to be specified.

**Output** A file named `indiv.txt` that is created in the folder where the Matlab program is. Each line corresponds to an *individual* taxpayer. To import it in Stata, click **File, Import, Text data (delimited, \*.csv, ...)**, and select the Excel file. Then, compute the thresholds, average income etc.

### Steps of the program

1. First, the path to access tabulated data is indicated to the program.
2. Then, the program loads the tabulated data from the files `singles.xls` and `couples.xls`. The shortcut `loadpath` gives the emplacement of this spreadsheet. The thresholds of singles and couples are respectively stored in the tables `tab1.thr` and `tab2.thr`, the percentiles in `tab1.p` and `tab2.p` and the Pareto coefficients in `tab1.b` and `tab2.b`.  $T$  is the total

number of thresholds in the tax scale considered. `y_av1` and `y_av2` are estimations of the mean income of singles and couples. `N1` and `N2` are the number of singles and the number of couples in the population.

3. The program calls the routine `quantile.m` to compute the values of the quantile function at the points of vectors `mid1` and `mid2`, which are the midpoints of the subintervals defined by the points of `nodes1` and `nodes2` (considering the midpoints avoids the problem of  $Q$  going to infinity in 1). `mid1` and `mid2` contains respectively `N1` and `N2` points. Values taken by the quantile function are stored in `Q1` and `Q2`.
4. The program merges the two vectors `Q1` and `Q2` after sharing out in 2 the incomes accruing to couples. The vector `Q` is sorted in an increasing order.
5. Exportation of vector `Q` in text file `indiv.txt`.

## D.2.2 Code

```
%% Estimation of shares of income accruing to different deciles and percentiles of the population
%
%
%% 1. Preliminaries
close all; clear all;
currentfolder=pwd;
% Path to load data
loadpath=[currentfolder, '/data'];

%% 2. Loads data
disp('Loads data');

cd(loadpath);
[tabulation_singles] = xlsread('singles', 'Tabulation'); %thresholds, singles
tab1.thr = tabulation_singles(:,1); %thresholds
tab1.p = tabulation_singles(:,2); %number of taxpayers
tab1.b = tabulation_singles(:,3); %inverted Pareto coefficients
T = length(tab1.thr); %number of thresholds in the tax tabulation
[y_av1] = xlsread('singles', 'Average income'); %average income
[N1] = xlsread('singles', 'Number'); %total number of singles

[tabulation_families] = xlsread('couples', 'Tabulation'); %thresholds, singles
tab2.thr = tabulation_families(:,1); %thresholds
tab2.p = tabulation_families(:,2); %number of taxpayers
tab2.b = tabulation_families(:,3); %inverted Pareto coefficients
T = length(tab2.thr); %number of thresholds in the tax tabulation
[y_av2] = xlsread('couples', 'Average income'); %average income
[N2] = xlsread('couples', 'Number'); %total number of couples

clear tabulation_singles tabulation_families;
cd(currentfolder);

%% 3.1. Estimates the quantile function - Singles
disp('Estimates the quantile function - Singles');

nodes1 = linspace(0,1,N1+1)';
```

```

mid1 = (nodes1(1:N1,1)+nodes1(2:N1+1,1))/2;

[Q1,p01] = quantile(tab1.thr,tab1.p,tab1.b,y_av1,mid1);

%% 3.2. Estimates the quantile function - Couples
disp('Estimates the quantile function - Couples');

nodes2 = linspace(0,1,2*N2+1)';
mid2 = (nodes2(1:2*N2,1)+nodes2(2:2*N2+1,1))/2;

[Q2,p02] = quantile(tab2.thr,tab2.p,tab2.b,y_av2,mid2);

%% 4. Merges the two parts of the population
disp('Merges the two parts of the population');

Q = [Q1
      Q2./2];
Q = sort(Q);

%% 5. Exports the population
disp('Exports the population');

dlmwrite('indiv.txt',Q,'precision',9);

```

### D.2.3 Evidence with microdata of France 2006

In the microdata provided by tax authorities, we split the households with two adults in 2 assigning the same income to the two spouses. We provided data about singles and couples to the program, using the thresholds of France 2012. The tables below compare the estimated thresholds and average incomes to the true values. They are always very close.

We also notice that both the thresholds and the average incomes are lower than for the distribution of income among households.

	Estimation	Microdata	Ratio
30%	9,915	9,925	0.9989
40%	11,816	11,825	0.9992
50%	13,671	13,658	1.0010
60%	15,655	15,691	0.9977
70%	18,194	18,268	0.9959
80%	21,860	21,957	0.9956
90%	29,064	29,117	0.9982
95%	37,681	38,046	0.9904
99%	71,581	72,577	0.9863
99.5%	98,859	99,112	0.9974
99.9%	230,255	228,425	1.0080
99.99%	1,015,445	1,018,528	0.9970

Table D.1: Comparison with microdata - Thresholds

	Estimation	Microdata	Ratio
30%	21,659	21,726	0,9969
40%	23,458	23,532	0,9968
50%	25,602	25,690	0,9966
60%	28,341	28,452	0,9961
70%	32,168	32,297	0,9960
80%	38,309	38,458	0,9961
90%	51,668	51,895	0,9956
95%	70,616	70,931	0,9956
99%	158,763	159,261	0,9969
99.5%	234,980	235,188	0,9991
99.9%	624,610	626,546	0,9969
99.99%	2,705,504	2,716,070	0,9961

Table D.2: Comparison with microdata - Average income

# Appendix E

## New series

We provide below tables with new estimations of the income and inheritance distributions. They have been constructed from tax tabulations using our new approximation method.

### E.1 Income distribution

#### E.1.1 Taxable income

	90%	95%	99%	99.5%	99.9%	99.99%
1919	4725	7016	22317	37248	111634	406346
1920	6659	9707	27663	43902	128202	441303
1921	7039	10287	27545	43024	119232	413922
1922	7559	11061	30491	47716	132999	449703
1923	8377	12680	35096	55168	157573	524048
1924	9423	14443	39765	61730	164373	491426
1925	10979	16603	43330	66412	172562	534253
1926	12728	18812	48951	77300	206441	656315
1927	13702	20196	48932	75694	204244	685857
1928	15118	21555	51845	80767	219254	729030
1929	16393	23224	54106	82299	216187	706349
1930	17273	24270	54516	81588	206935	663447
1931	16496	23336	50338	73468	181103	568131
1932	15923	22664	48337	69459	165540	510451
1933	15939	22780	47985	68170	158393	493193
1934	15249	21727	45324	64503	150006	466215
1935	14811	21128	44516	63735	146930	462789
1936	15980	21982	47049	67226	156727	492406
1937	18511	25404	53764	77434	180921	555177
1938	20328	28103	58439	83484	190996	556768
1939	18768	26259	54280	78586	192900	613748
1940	17505	24549	51821	73352	167324	504595
1941	22038	31251	68718	100219	226319	648004
1942	26981	38156	83906	119383	251161	628451
1943	32101	44634	94768	130535	258935	619099
1944	38464	53010	103339	136427	254534	566796

Table E.1: Estimations of taxable income 1919-1944 - Thresholds

Source: Results of estimations using raw inheritance tax tabulations.

	90%	95%	99%	99.5%	99.9%	99.99%
1945	77213	103169	189520	249333	470171	1193151
1946	124837	169964	341294	466414	987959	2804456
1947	180536	238716	442376	592004	1213396	3268827
1948	284610	374540	715799	978856	2066249	5603946
1949	342241	462466	899032	1230869	2668574	7701735
1950	399680	537713	1034770	1425914	3056073	8731980
1951	527136	695853	1324583	1813310	3779836	10573516
1952	608685	811487	1594332	2176338	4493251	12321283
1953	584206	781409	1563001	2149317	4439584	12305183
1954	592357	793486	1620938	2240432	4605779	12672873
1955	649159	879901	1801953	2480993	5059781	13810194
1956	708797	962745	1986190	2734622	5530090	14835185
1957	805738	1085970	2239265	3058441	6145172	16734017
1958	909195	1218201	2488749	3351698	6601455	17851310
1959	939194	1291110	2670632	3634393	7200974	19261450
1960	10192	14011	29577	40634	81919	220454
1961	11204	15586	32947	45420	89986	241869
1962	12459	17260	35930	49127	97631	256591
1963	13945	19327	39779	54092	105876	275621
1964	15293	21191	43924	60088	117225	301770
1965	16487	22942	47463	64734	124895	324953
1966	17538	24259	49825	67819	131679	346262
1967	18812	26057	53759	73357	143543	386147
1968	20303	27826	56082	75680	147850	401644
1969	22243	30320	60976	82642	160810	432342
1970	24516	33411	67544	91665	179435	467711
1971	26996	36655	74831	102136	202775	531820
1972	29436	39890	81997	112461	225097	603452
1973	33454	45367	94310	129748	260144	733133
1974	38981	52375	107292	147424	290116	773550
1975	45231	60394	123678	168671	328952	884051
1976	51628	68852	141327	191557	377359	1017130
1977	56774	75642	149079	200406	392845	1076140
1978	63927	84546	169227	227232	448643	1208709
1979	70967	93697	189185	255439	512155	1399908
1980	80455	106230	210075	283129	565346	1533630
1981	91956	121978	236503	316621	631554	1726091
1982	103003	135075	257607	341745	665884	1758205
1983	116884	152198	284758	375144	712379	1799882
1984	124676	162350	302757	397479	757403	1940849
1985	131711	172245	323696	427486	818702	2098697
1986	137041	179827	342166	455128	886463	2328979
1987	139593	183999	353789	475316	955561	2644979
1988	144759	191684	384020	517837	1013801	2651413
1989	150623	199670	390020	529613	1098915	3198988
1990	157359	209841	407092	552101	1144908	3333739
1991	162358	216457	413200	556159	1131865	3204048
1992	166112	220485	414625	554516	1111707	3078645
1993	167989	222502	413306	551019	1100699	3038431
1994	169851	224765	415617	555078	1117300	3125306
1995	172901	228639	423079	564354	1129246	3125872
1996	175330	231675	425165	565869	1128644	3112286
1997	177869	235287	431627	576318	1160282	3247300
1998	182733	241685	444484	593375	1189661	3302297

Table E.2: Estimations of taxable income 1945-1998 - Thresholds  
Source: Results of estimations using raw inheritance tax tabulations.

	90%	95%	99%	99.5%	99.9%	99.99%
1919	14230	22234	62487	96664	254781	854730
1920	18096	27315	74272	113822	285719	997739
1921	17877	26695	69729	104916	256794	825445
1922	19464	29334	76778	115965	283583	890412
1923	22247	33817	89970	136167	331018	1047309
1924	24200	36310	93590	138550	321893	955950
1925	26994	40376	100736	148167	344789	1038744
1926	31104	46158	119735	178917	428536	1399055
1927	32127	47162	120496	181109	444821	1451292
1928	34562	50132	128276	192403	469993	1542516
1929	35800	51423	126963	189141	450895	1479195
1930	36365	51737	124091	182538	430681	1339588
1931	32858	46688	108391	157307	363132	1116274
1932	31351	44251	99848	143084	322622	957091
1933	31132	43786	97386	138846	312756	956909
1934	29546	41422	92097	131244	294318	876589
1935	29455	40767	91216	129954	294244	931951
1936	31845	43927	98255	140421	319368	1016105
1937	36478	50492	113477	162878	370033	1213786
1938	39339	54050	119699	169296	367755	1146235
1939	37480	51998	117359	170113	393590	1265018
1940	34769	47928	104152	146958	319269	907769
1941	44392	62078	139076	195989	415213	1157057
1942	52258	72698	156594	214176	422785	1065745
1943	59104	80724	166497	222972	423974	1072364
1944	66603	88479	167828	218318	391766	888283

Table E.3: Estimations of taxable income 1919-1944 - Average income

Source: Results of estimations using raw inheritance tax tabulations.

	90%	95%	99%	99.5%	99.9%	99.99%
1945	128962	168433	315045	415250	797345	2068957
1946	225342	305368	628781	865306	1807203	5032246
1947	296979	398086	779056	1052867	2122554	5422727
1948	489842	649928	1306690	1787715	3648920	9593709
1949	608049	819482	1679515	2322728	4874111	13162622
1950	700616	942833	1929777	2663329	5575460	15272759
1951	902018	1201046	2418681	3311610	6811993	18347230
1952	1059616	1421513	2869097	3903073	7869499	20408654
1953	1030609	1390260	2843698	3878349	7851808	20835884
1954	1057229	1433103	2960824	4037398	8148318	21888888
1955	1168066	1585566	3261466	4436564	8919618	24057658
1956	1276786	1734897	3576797	4853715	9677307	26281914
1957	1440527	1952747	3993355	5402226	10728072	28877664
1958	1605301	2165026	4366458	5876621	11667623	30639456
1959	1708334	2323905	4700790	6332100	12458932	32978808
1960	18766	25683	52951	71715	142129	378738
1961	20812	28550	58714	79314	156627	423218
1962	22850	31193	63405	85301	166722	436235
1963	25394	34555	69345	92813	179558	465890
1964	27941	38094	76728	102670	197688	506540
1965	30186	41139	82556	110311	211974	547598
1966	31877	43340	87017	116584	226198	595396
1967	34397	46877	94798	127496	249841	672157
1968	36329	49077	97902	131419	258919	705734
1969	39583	53429	106881	143618	282858	783515
1970	43702	59035	118104	158458	308971	830550
1971	48295	65429	132339	178231	349401	941043
1972	52869	71819	146843	198759	393789	1088731
1973	60695	82788	171380	233325	472999	1367417
1974	69138	93606	190169	256290	504110	1349538
1975	79592	107379	217433	292237	574987	1553118
1976	90685	122337	248085	333815	660756	1784053
1977	97976	130969	260194	349645	696448	1914694
1978	110160	147492	294968	396187	785199	2116993
1979	123191	165621	334778	451998	906229	2477040
1980	138117	184926	370197	498550	994562	2697416
1981	156870	209489	414896	558119	1117484	3056574
1982	172313	227898	440564	586964	1147918	3033552
1983	191402	250852	474765	625872	1188945	3004161
1984	203967	267191	505651	667160	1277173	3276315
1985	217311	285554	544326	720054	1380851	3540809
1986	228932	302451	585479	780387	1522462	4001389
1987	236952	315297	625132	844971	1707448	4731862
1988	247480	330625	663254	890227	1735118	4532914
1989	260960	350492	712976	977165	2043384	5959106
1990	273276	366900	743295	1018282	2129383	6212271
1991	278525	371597	738669	1003603	2058305	5836887
1992	280978	372524	728892	983820	1987284	5512831
1993	282002	372597	723014	974183	1962791	5428774
1994	284929	376373	732588	990323	2013001	5643259
1995	289514	382090	741355	999720	2018016	5597176
1996	292294	385134	742015	999029	2011005	5556907
1997	297431	392268	760734	1028507	2091329	5866105
1998	305345	402654	780412	1053419	2130568	5925698

Table E.4: Estimations of taxable income 1945-1998 - Average income  
Source: Results of estimations using raw inheritance tax tabulations.

### E.1.2 Fiscal income

	90%	95%	99%	99.5%	99.9%	99.99%
1919	5559	8158	27274	46932	144793	548439
1920	7894	11603	33878	55835	178621	698989
1921	8364	12397	34359	56882	173779	747777
1922	8982	13300	37782	61625	184925	721274
1923	9950	15239	43434	71133	219485	833264
1924	11201	17437	49801	82126	245530	900152
1925	13073	20028	54810	90605	263279	955649
1926	15174	22759	61478	102975	297381	1055861
1927	16312	24352	61314	99765	283207	1022967
1928	18007	25983	64538	105335	302606	1089207
1929	19505	28002	68004	108947	308786	1109499
1930	20554	29272	68221	108248	292680	1055169
1931	19656	28262	63552	98332	263066	919896
1932	18971	27386	60628	91788	234254	816584
1933	18981	27502	60074	89546	222336	763256
1934	18169	26288	56900	84810	209983	739021
1935	17631	25451	55267	82238	201100	673635
1936	19005	26402	58138	86109	213008	738549
1937	21985	30501	66527	99699	256256	875672
1938	24195	33848	72924	110152	284626	987244
1939	22430	31874	68915	106345	288451	1036258
1940	20897	29751	65276	99035	264088	1028531
1941	26185	37425	84402	127661	315801	1033289
1942	32162	45994	105113	161011	388414	1191306
1943	38368	54207	121678	182812	419159	1143091
1944	45874	64334	133753	192764	416848	1137776

Table E.5: Estimations of fiscal income 1919-1944 - Thresholds

	90%	95%	99%	99.5%	99.9%	99.99%
1945	91545	122818	231496	313676	602847	1538093
1946	147775	201111	407707	561599	1168566	3243210
1947	215787	287280	544490	747864	1557727	4379668
1948	336674	443832	857776	1185574	2471585	6578605
1949	405947	550595	1087088	1509848	3239083	9177718
1950	474325	641517	1255240	1759200	3747233	10539281
1951	624899	827966	1600528	2226114	4607695	12730885
1952	722249	967590	1931257	2678092	5499669	14920041
1953	737849	982465	1955118	2675556	5504041	15011724
1954	797589	1060701	2123271	2915555	5822782	15421696
1955	873391	1173568	2354102	3217971	6368193	16740557
1956	953688	1284752	2593737	3547210	6965067	17994462
1957	1083879	1447620	2924592	3963834	7725461	20272126
1958	1222976	1624583	3252418	4350185	8303210	21717576
1959	1335478	1818078	3677424	4897654	9189419	23391306
1960	14684	19950	41070	55047	104587	264489
1961	16126	22132	45567	61306	114455	288794
1962	17920	24463	49550	66039	123512	305832
1963	20041	27323	54639	72307	133042	325381
1964	21964	29900	60112	79868	146187	353279
1965	23662	32304	64748	85698	154922	377751
1966	25153	34093	67753	89355	162334	399503
1967	26958	36527	72783	96067	175564	441686
1968	29071	38921	75677	98679	179849	456904
1969	31823	42306	81923	107082	194084	487600
1970	35049	46510	90397	118138	215080	523630
1971	38566	50910	99775	130944	241399	590911
1972	42052	55403	109329	144182	267972	670502
1973	47792	63010	125746	166343	309696	814592
1974	55686	72743	143056	189005	345377	859499
1975	64616	83880	164903	216244	391609	982279
1976	73754	95627	188435	245586	449237	1130145
1977	81106	105058	198773	256930	467672	1195712
1978	91324	117425	225635	291324	534099	1343010
1979	101382	130135	252247	327486	609709	1555454
1980	114936	147542	280100	362986	673031	1704034
1981	131365	169414	315337	405925	751850	1917878
1982	147148	187604	343475	438135	792719	1953561
1983	166977	211386	379677	480955	848070	1999869
1984	178109	225487	403677	509588	901670	2156499
1985	188158	239229	431594	548059	974645	2331886
1986	195773	249761	456221	583497	1055313	2587754
1987	199419	255555	471718	609379	1137572	2938865
1988	206799	266229	512027	663894	1206906	2946014
1989	215177	277319	520026	678992	1308232	3554431
1990	224799	291446	542790	707822	1362986	3704154
1991	231940	300635	550933	713024	1347458	3560053
1992	237303	306229	552833	710918	1323461	3420717
1993	239984	309029	551076	706435	1310355	3376034
1994	242643	312174	554155	711639	1330119	3472562
1995	247001	317555	564106	723531	1344340	3473191
1996	250473	321770	566886	725472	1343623	3458095
1997	254099	326788	575502	738869	1381288	3608111
1998	261047	335674	592645	760738	1416262	3669218

Table E.6: Estimations of fiscal income 1945-1998 - Thresholds

	90%	95%	99%	99.5%	99.9%	99.99%
1919	17283	27301	79538	124487	334719	1151417
1920	22685	34766	99052	155522	416018	1571694
1921	22858	34876	97252	151280	401346	1476375
1922	24440	37425	102842	159085	412440	1418258
1923	27928	43127	120358	186472	480417	1654693
1924	30947	47455	131017	201141	511999	1744133
1925	34676	53110	143012	218493	551643	1843693
1926	39458	59708	164039	251994	639031	2239459
1927	40256	60113	161426	247789	631321	2157239
1928	43182	63662	170687	261449	665158	2296070
1929	45004	65930	172601	263865	663563	2313118
1930	45595	66153	168245	254329	631443	2118551
1931	41356	60040	148736	222367	544730	1797737
1932	39137	56292	134743	198465	472930	1521429
1933	38713	55421	130356	190682	452035	1476111
1934	36800	52546	123561	180637	427077	1380639
1935	36326	51034	119498	173835	410223	1352404
1936	39234	54898	128511	187649	447509	1516916
1937	45208	63672	151240	223124	542491	1905309
1938	49351	69309	165020	242344	577910	2016301
1939	47517	67502	164077	246064	611274	2123711
1940	44340	62768	149256	221099	542752	1826545
1941	54675	77579	182566	264155	600599	1831818
1942	65739	93550	218323	312572	687955	1997764
1943	75035	105269	238335	335368	708503	1965541
1944	84202	115063	242056	332208	671451	1761917

Table E.7: Estimations of fiscal income 1919-1944 - Average income

	90%	95%	99%	99.5%	99.9%	99.99%
1945	155278	204435	394482	526534	1023782	2667088
1946	267282	362545	750066	1031546	2124265	5819626
1947	361044	488197	983395	1344891	2755388	7265309
1948	582298	774474	1569819	2150215	4345067	11275039
1949	727339	984190	2042067	2831070	5889327	15702604
1950	840357	1136703	2361456	3271832	6809339	18450992
1951	1078812	1442758	2946003	4048437	8278331	22106934
1952	1269907	1711865	3505190	4787496	9607705	24729634
1953	1293886	1740390	3537548	4812225	9696234	25443710
1954	1402810	1888305	3820989	5173890	10206563	26698112
1955	1546731	2083766	4195209	5664731	11124136	29228016
1956	1691160	2280867	4601651	6199869	12076265	31949558
1957	1906942	2565217	5134316	6893172	13367284	35061612
1958	2126515	2846823	5620548	7508452	14555833	37350684
1959	2378172	3202757	6271519	8316346	15713106	40155272
1960	26372	35678	70928	94386	178806	456127
1961	29177	39538	78355	103993	196223	507256
1962	31983	43103	84334	111427	208020	521873
1963	35462	47596	91764	120520	222415	552147
1964	38919	52294	100995	132494	243032	595358
1965	41960	56318	108218	141684	259063	639162
1966	44210	59146	113495	148874	274579	689744
1967	47552	63699	122846	161624	300745	771965
1968	50139	66520	126322	165768	309911	806076
1969	54473	72127	137049	179869	335735	887162
1970	59992	79433	150712	197376	364364	933636
1971	66104	87697	167962	220674	409157	1049925
1972	72325	96202	186254	245960	460923	1214608
1973	82968	110797	217132	288381	553159	1525319
1974	94629	125451	241361	317314	590406	1505771
1975	108966	143934	275976	361773	673308	1732870
1976	124143	163969	314827	413170	773754	1990546
1977	134228	175641	330121	432600	815280	2136186
1978	150916	197800	374340	490365	919503	2362036
1979	168670	221979	424590	559161	1060969	2763645
1980	189185	247937	469627	616895	1164552	3009592
1981	214958	280955	526281	690471	1308280	3410223
1982	236376	306017	559530	726949	1344748	3384893
1983	262875	337251	603808	776148	1393894	3352560
1984	280118	359183	642880	827037	1496944	3656102
1985	298354	383756	691961	892566	1618473	3951266
1986	314071	406131	743651	966632	1783686	4464901
1987	324669	422760	792622	1044929	1998487	5279129
1988	339227	443668	842383	1102771	2033046	5058099
1989	357034	469182	902347	1206387	2389345	6647269
1990	373914	491199	940729	1257144	2489885	6929660
1991	381471	498024	935878	1240209	2408101	6511494
1992	385152	499694	924250	1216646	2325978	6150399
1993	386692	499951	916944	1204867	2297438	6056674
1994	390638	504870	928667	1224311	2355612	6295705
1995	397002	512670	940113	1236341	2361954	6244499
1996	400906	516881	941100	1235647	2353914	6199645
1997	407823	526235	964333	1271495	2447245	6544304
1998	418714	540246	989539	1302630	2493561	6610961

Table E.8: Estimations of fiscal income 1945-1998 - Average income

	90%	95%	99%	99.5%	99.9%	99.99%
1919	42.25	33.38	19.49	15.28	8.34	2.93
1920	41.13	31.53	18.00	14.16	7.72	3.13
1921	40.71	31.06	17.35	13.51	7.26	2.77
1922	42.33	32.41	17.85	13.83	7.28	2.61
1923	43.80	33.83	18.91	14.68	7.68	2.73
1924	42.25	32.40	17.92	13.79	7.13	2.59
1925	44.05	33.74	18.20	13.94	7.14	2.45
1926	42.81	32.39	17.82	13.71	7.02	2.47
1927	43.49	32.48	17.46	13.43	6.91	2.45
1928	43.64	32.17	17.27	13.25	6.80	2.37
1929	42.11	30.85	16.18	12.39	6.31	2.34
1930	41.46	30.08	15.33	11.61	5.83	1.97
1931	40.50	29.41	14.60	10.94	5.44	1.92
1932	42.75	30.76	14.75	10.89	5.25	1.76
1933	44.15	31.61	14.88	10.90	5.21	1.71
1934	45.28	32.34	15.25	11.18	5.37	1.84
1935	45.84	32.75	15.35	11.17	5.30	1.76
1936	45.00	31.49	14.76	10.80	5.18	1.81
1937	43.19	30.42	14.48	10.70	5.27	1.87
1938	42.53	29.87	14.26	10.49	5.09	1.90
1939	38.48	27.34	13.32	10.00	5.05	1.85
1940	39.60	28.04	13.36	9.91	4.92	1.73
1941	38.56	27.37	12.91	9.36	4.31	1.36
1942	34.59	24.64	11.56	8.32	3.75	1.18
1943	31.73	22.27	10.13	7.15	3.09	0.91
1944	28.98	19.82	8.39	5.79	2.41	0.67

Table E.9: Estimation of fiscal income 1919-1944 - Shares

	90%	95%	99%	99.5%	99.9%	99.99%
1945	29.72	19.57	7.58	5.07	2.00	0.55
1946	32.90	22.32	9.25	6.36	2.62	0.72
1947	33.88	22.91	9.24	6.33	2.60	0.69
1948	32.49	21.61	8.77	6.02	2.45	0.63
1949	32.11	21.72	9.03	6.26	2.63	0.71
1950	31.98	21.63	9.00	6.24	2.62	0.71
1951	32.98	22.05	9.02	6.20	2.56	0.69
1952	33.19	22.37	9.18	6.28	2.54	0.66
1953	32.90	22.13	9.00	6.13	2.48	0.66
1954	33.54	22.58	9.15	6.20	2.46	0.65
1955	34.39	23.17	9.34	6.31	2.49	0.65
1956	34.28	23.12	9.35	6.30	2.48	0.66
1957	34.75	23.38	9.38	6.30	2.47	0.67
1958	34.06	22.80	9.03	6.05	2.39	0.66
1959	35.87	24.16	9.48	6.31	2.42	0.65
1960	36.11	24.43	9.74	6.51	2.51	0.68
1961	36.80	24.94	9.91	6.59	2.51	0.69
1962	35.86	24.17	9.49	6.28	2.37	0.63
1963	36.42	24.44	9.45	6.21	2.33	0.61
1964	36.85	24.76	9.58	6.30	2.35	0.60
1965	37.14	24.93	9.60	6.29	2.35	0.60
1966	36.45	24.39	9.38	6.17	2.32	0.59
1967	36.21	24.26	9.37	6.17	2.32	0.61
1968	34.80	23.09	8.78	5.76	2.17	0.56
1969	33.96	22.49	8.56	5.62	2.12	0.58
1970	33.15	21.95	8.34	5.47	2.03	0.52
1971	33.34	22.12	8.48	5.58	2.09	0.54
1972	33.04	21.98	8.52	5.63	2.13	0.56
1973	33.87	22.62	8.87	5.90	2.28	0.63
1974	33.33	22.09	8.51	5.60	2.09	0.53
1975	33.42	22.08	8.48	5.56	2.08	0.53
1976	33.18	21.92	8.43	5.53	2.08	0.53
1977	31.64	20.70	7.79	5.11	1.94	0.51
1978	31.37	20.56	7.79	5.10	1.92	0.49
1979	31.03	20.42	7.82	5.15	1.96	0.51
1980	30.69	20.11	7.63	5.02	1.90	0.49
1981	30.73	20.09	7.53	4.94	1.88	0.49
1982	29.92	19.37	7.09	4.61	1.71	0.43
1983	30.43	19.52	7.00	4.50	1.62	0.39
1984	30.51	19.56	7.00	4.51	1.64	0.40
1985	31.03	19.96	7.21	4.65	1.69	0.41
1986	31.38	20.29	7.43	4.83	1.79	0.45
1987	31.71	20.65	7.75	5.11	1.96	0.52
1988	32.05	20.96	7.96	5.22	1.93	0.48
1989	32.39	21.28	8.19	5.48	2.18	0.60
1990	32.60	21.41	8.21	5.49	2.18	0.60
1991	32.39	21.15	7.95	5.27	2.05	0.55
1992	32.17	20.87	7.72	5.09	1.95	0.51
1993	32.15	20.79	7.63	5.02	1.92	0.50
1994	32.29	20.87	7.68	5.07	1.96	0.52
1995	32.35	20.89	7.67	5.04	1.93	0.51
1996	32.19	20.75	7.56	4.97	1.90	0.50
1997	32.32	20.86	7.65	5.05	1.95	0.52
1998	32.44	20.93	7.68	5.05	1.94	0.51

Table E.10: Estimation of fiscal income 1945-1998 - Shares

### E.1.3 Estimations for the years 2001-2012

These estimations are obtained from raw data in tax tabulations for the years 2001-2012. For the recent years, tax tabulations provide data for nontaxable households and for fiscal incomes.

	90%	95%	99%	99.5%	99.9%	99.99%
2001	56205	76736	177305	260067	636551	2330794
2002	57402	78297	179074	260776	627154	2237582
2003	59315	81847	184951	268183	640739	2269347
2004	61150	84907	196950	289265	712408	2637346
2005	61747	85419	196531	287451	700850	2556411
2006	78110	107389	241451	348514	822781	2861766
2007	81641	113006	258393	375090	896560	3173227
2008	82098	112490	248303	353882	809282	2684245
2009	80157	108060	225514	312609	669335	2016697
2010	83167	112830	241830	339522	748420	2350825
2011	88570	121297	267179	385839	948299	3608228
2012	89568	121979	261793	371378	867033	3091851

Table E.11: Estimations of income distribution 2001-2012 - Threshold

Source: Results of estimations using raw inheritance tax tabulations.

	90%	95%	99%	99.5%	99.9%	99.99%
2001	56205	76736	177305	260067	636551	2330794
2002	57402	78297	179074	260776	627154	2237582
2003	59315	81847	184951	268183	640739	2269347
2004	61150	84907	196950	289265	712408	2637346
2005	61747	85419	196531	287451	700850	2556411
2006	78110	107389	241451	348514	822781	2861766
2007	81641	113006	258393	375090	896560	3173227
2008	82098	112490	248303	353882	809282	2684245
2009	80157	108060	225514	312609	669335	2016697
2010	83167	112830	241830	339522	748420	2350825
2011	88570	121297	267179	385839	948299	3608228
2012	89568	121979	261793	371378	867033	3091851

Table E.12: Estimations of income distribution 2001-2012 - Average income

Source: Results of estimations using raw inheritance tax tabulations.

	90%	95%	99%	99.5%	99.9%	99.99%
2001	36,17	24,69	11,42	8,39	4,14	1,50
2002	35,92	24,50	11,22	8,18	3,96	1,40
2003	36,08	24,89	11,26	8,17	3,94	1,38
2004	36,34	25,23	11,72	8,61	4,28	1,57
2005	35,97	24,88	11,46	8,39	4,12	1,49
2006	35,62	24,48	11,02	7,96	3,79	1,31
2007	35,95	24,88	11,39	8,28	3,99	1,40
2008	35,38	24,24	10,71	7,64	3,52	1,16
2009	34,58	23,31	9,74	6,76	2,92	0,87
2010	35,04	23,77	10,20	7,17	3,18	0,99
2011	35,30	24,17	10,66	7,70	3,82	1,44
2012	35,11	23,91	10,27	7,29	3,43	1,21

Table E.13: Estimations of income distribution 2001-2012 - Share

Source: Results of estimations using raw inheritance tax tabulations.

## E.2 Inheritance distribution

	90%	95%	99%	99,5%	99,9%	99,99%
1902	9258	22675	120910	231278	798511	3464811
1903	10123	23925	128002	245483	926245	3239522
1904	10022	23389	128191	245705	890946	4444482
1905	10171	23612	131024	254228	925213	4472607
1907	10565	24269	131143	249523	903462	3957299
1909	11140	25096	133647	259756	971165	4935991
1910	11509	26563	141071	274361	1010759	4127042
1911	11060	25490	135788	261765	974457	4074758
1912	11644	26809	137557	265844	980378	4521541
1913	12121	27431	146264	274825	990326	4485090
1925	27793	55905	227814	401948	1297756	5231382
1926	32084	62950	265363	469113	1585633	6927137
1927	34513	68269	291002	526524	1802389	8843049
1929	38566	75441	327958	608617	2216410	11685246
1930	43787	86228	368966	678663	2663158	12003789
1931	44832	88058	353781	639139	2219694	9589354
1932	43650	84723	352605	625968	2079418	9092005
1933	42544	82410	338452	597383	1972250	10791831
1935	41460	80455	326807	571442	1890222	7849900
1936	41623	79868	318685	553537	1758376	7477559
1937	45112	85974	351894	632561	1927869	8520936
1938	52451	97334	385211	637667	1961731	7846120
1939	53793	101062	385106	668722	2000160	8367702
1940	48154	86617	309449	491922	1344669	5135244
1941	71047	133961	462756	733739	2112697	8118220
1942	92590	180850	651329	1084022	3181334	10585030
1943	126836	247720	918720	1543554	4443974	15443597
1944	125679	245473	896709	1471682	4427705	17086738
1945	167660	319370	1118157	1785367	5078724	19121070
1946	241183	430735	1377151	2177011	5875611	21822920
1947	331661	588114	1859529	2949725	7881845	27724814
1948	396821	716021	2270416	3566617	9410562	34599948
1949	453410	824249	2567219	4113315	11203541	38650340
1950	550680	1032805	3419171	5503987	15706564	53822012
1951	627382	1177984	3991036	6490455	18347724	58440164
1952	1001310	1912657	6387748	9898641	25885702	86407208
1953	992216	1923016	6564267	10329990	24692960	88673464
1954	1367198	2592613	8297632	12719076	32596666	116995312
1955	1292198	2511014	8309039	12904740	34730576	114893248
1956	1306761	2585862	8795604	13548104	36352540	125927672
1957	1390941	2698800	9668779	15396905	42324480	136847056
1958	1839187	3477004	11579041	18435070	50874656	155643744
1959	2011109	3872347	12997238	20674650	54801696	211810592
1960	21323	41417	134813	212605	590122	1899978
1962	26274	52203	178027	273895	756390	2732173
1964	39715	76418	259799	409295	1040133	3509496
1984	429083	685940	1740691	2449851	5252545	16130809
1994	829518	1308901	3146175	4431961	10185302	27885826

Table E.14: New estimations of inheritance distribution - Threshold  
Source: Results of estimations using raw inheritance tax tabulations.

	90%	95%	99%	99,5%	99,9%	99,99%
1902	71604	127351	438781	709685	1949726	6608619
1903	76300	137009	480402	783640	2150237	7448979
1904	82391	149792	544755	910509	2784056	13156384
1905	87212	158909	589306	995702	3112412	15583179
1907	76944	138170	483705	787386	2123706	7082355
1909	83969	151588	546197	904252	2648118	10683374
1910	83332	149488	523350	849934	2307788	8453069
1911	86601	156825	567167	945257	2792709	12427024
1912	88776	160214	579869	969166	2928916	13327304
1913	86601	155564	548638	900991	2596662	10346299
1925	133098	228244	745696	1199451	3224201	11331505
1926	150463	258850	857210	1396365	3970083	15461063
1927	173905	301826	1028313	1685479	4894829	19756828
1929	210860	369609	1288534	2130661	6330680	26575850
1930	238515	418376	1465620	2436281	7270435	30104414
1931-32	219990	380605	1274018	2081728	5869598	21892006
1932	212000	364853	1197480	1932003	5404150	22173680
1933	201291	346303	1131159	1825413	5086092	17700242
1935	203461	351967	1179398	1932171	5756772	28750960
1936	205599	355305	1191273	1968478	5980845	31591142
1937	206978	354160	1134608	1812911	5009341	21066998
1938	229428	390673	1250968	2031347	5931312	28772394
1939	235315	399994	1287463	2081687	6110628	30017448
1940	158167	261505	758179	1173574	3067032	12210908
1941	260971	426862	1215610	1850172	4604528	14265910
1942	368651	608527	1777269	2718539	6709060	22308678
1943	507870	841018	2446467	3723612	8908560	25443574
1944	507479	841878	2484411	3836515	9743567	31700598
1945	628367	1027849	2935315	4475374	11111664	37674528
1946	783347	1250580	3382195	5048128	12151263	35588840
1947	1069161	1699774	4599540	6918688	16505401	48907928
1948	1295241	2059546	5534244	8246667	19647112	58546376
1949	1581024	2549524	7055769	10486333	24111958	73513592
1950	1993559	3224758	9003571	13698118	33968868	110835320
1951	2335600	3792334	10706515	16229856	39673592	128878544
1952	3549330	5714353	15359013	22765700	54078120	152144416
1953	3542334	5688631	15216540	22423482	52249172	172641408
1954	4626221	7375033	19432038	28668360	68944192	216501824
1955	4571613	7367668	19700782	29134778	68177496	199842832
1956	4801859	7782297	21048170	31277140	75383624	226954432
1957	5462126	8936713	25196552	37946800	91583008	305812448
1958	6663103	10710831	29196732	43475772	99096664	284072320
1959	7601236	12293109	34111896	51448628	129560296	446731776
1960	77653	124534	335052	499146	1170005	3206106
1962	99963	162016	445125	669049	1679370	5543180
1964	143866	232202	626652	927913	2184135	7219659
1984	1033539	1530501	3427670	4763135	10315593	32673196
1994	1903161	2757693	6098757	8534659	18016304	47899512

Table E.15: New estimations of inheritance distribution - Average income above  
Source: Results of estimations using raw inheritance tax tabulations.

	90%	95%	99%	99,5%	99,9%	99,99%
1902	83,94	75,47	51,64	41,68	23,12	7,78
1903	86,13	78,28	54,44	44,29	24,40	8,28
1904	87,54	80,52	58,12	48,44	29,79	13,99
1905	88,09	81,03	59,75	50,37	31,65	15,73
1907	86,26	78,43	54,45	44,19	23,93	8,01
1909	86,91	79,24	56,84	46,88	27,58	11,06
1910	86,08	77,94	54,39	43,96	23,94	8,73
1911	87,64	80,14	57,69	47,92	28,44	12,54
1912	87,05	79,17	57,08	47,58	28,90	13,05
1913	86,27	78,20	54,91	44,89	26,03	10,31
1925	80,35	69,18	44,60	35,96	19,26	6,76
1926	80,01	69,06	45,12	36,93	21,19	7,68
1927	81,03	70,40	47,59	39,07	22,84	9,00
1929	82,01	71,93	50,24	41,69	24,73	10,59
1930	81,59	71,61	50,30	41,81	24,88	10,13
1931-32	79,99	69,26	46,47	38,16	21,56	7,95
1932	79,06	68,07	44,76	36,39	20,30	8,26
1933	79,51	68,48	44,93	36,47	20,19	6,77
1935	79,31	68,68	46,09	37,88	22,56	11,13
1936	78,83	68,18	45,77	37,97	23,04	12,06
1937	77,47	66,34	42,63	34,39	18,86	7,88
1938	77,62	65,26	42,02	34,09	19,94	9,59
1939	79,11	66,42	42,89	34,73	20,38	9,94
1940	81,73	64,66	38,65	28,90	15,29	5,98
1941	75,26	62,03	34,88	26,58	13,28	4,13
1942	75,95	62,86	36,85	28,04	14,09	4,52
1943	76,54	63,26	36,84	28,16	13,48	3,88
1944	79,37	64,80	38,28	29,67	15,09	4,90
1945	75,85	62,09	35,26	26,89	13,42	4,52
1946	70,95	56,43	30,67	22,81	10,94	3,15
1947	69,86	55,37	29,92	22,84	10,98	3,27
1948	71,12	56,85	30,38	22,70	11,01	3,36
1949	73,66	59,40	34,01	24,28	12,68	3,42
1950	74,34	60,07	33,61	25,52	12,74	4,13
1951	72,37	58,72	33,04	25,31	12,30	4,02
1952	74,20	59,73	32,32	23,82	11,45	3,16
1953	75,84	60,90	32,56	24,23	11,33	3,70
1954	72,99	58,11	30,52	22,57	10,98	3,40
1955	73,16	58,88	31,48	23,35	11,04	3,19
1956	69,44	56,26	30,38	22,65	10,99	3,28
1957	70,00	57,65	32,27	24,30	11,76	3,91
1958	68,14	54,88	30,07	22,25	10,20	2,92
1959	70,27	56,79	31,88	23,83	12,09	4,18
1960	67,60	54,12	29,46	21,75	10,25	2,86
1962	68,54	55,24	30,33	22,92	11,67	3,89
1964	71,61	57,79	31,34	23,09	10,91	3,65
1984	61,70	44,50	19,63	13,74	5,99	1,89
1994	55,42	40,53	17,73	12,43	5,26	1,43

Table E.16: New estimations of inheritance distribution - Share  
Source: Results of estimations using raw inheritance tax tabulations.

## Appendix F

# Empirical Pareto curves of the income and inheritance distributions

### F.1 Pareto curves of the income distribution 1915-2012

The links below point to the Pareto curves obtained with French income tax data for the incomes of the years 1919-2012.

- Whole Pareto curve  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc.pdf)
- Zoom on the top 10%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc\\_top10.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc_top10.pdf)
- Zoom on the top 1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc\\_top1.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc_top1.pdf)
- Zoom on the top 0.1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc\\_top01.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc_top01.pdf)

The following graphs represent both the Pareto curve interpolated from tax tabulations and the final Pareto curve corresponding to the income estimations obtained using our new method.

- Whole Pareto curve  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc\\_fin.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc_fin.pdf)
- Zoom on the top 10%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc\\_fin\\_top10.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc_fin_top10.pdf)
- Zoom on the top 1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc\\_fin\\_top1.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc_fin_top1.pdf)
- Zoom on the top 0.1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_pc\\_fin\\_top01.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_pc_fin_top01.pdf)

Finally, here are the graphs of the quantile functions, that is, the estimates of incomes corresponding to the different percentiles of the distribution.

- Whole quantile function  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_q.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_q.pdf)
- Zoom on the top 10%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_q\\_top10.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_q_top10.pdf)
- Zoom on the top 1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_q\\_top1.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_q_top1.pdf)
- Zoom on the top 0.1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inc\\_q\\_top01.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inc_q_top01.pdf)

## F.2 Pareto curves of the inheritance distribution 1902-1994

The graphs below represent the Pareto curves obtained with French inheritance tax data for the years 1902-1994.

- Whole Pareto curve  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_pc.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_pc.pdf)
- Zoom on the top 10%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_pc\\_top10.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_pc_top10.pdf)
- Zoom on the top 1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_pc\\_top1.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_pc_top1.pdf)
- Zoom on the top 0.1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_pc\\_top01.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_pc_top01.pdf)

The graphs below depict both the Pareto curve interpolated from tax tabulations and the final Pareto curve corresponding to the inheritance estimations obtained using our new method.

- Whole Pareto curve  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_pc\\_fin.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_pc_fin.pdf)
- Zoom on the top 10%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_pc\\_fin\\_top10.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_pc_fin_top10.pdf)
- Zoom on the top 1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_pc\\_fin\\_top1.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_pc_fin_top1.pdf)

At last, these links point to the graphs of the quantile functions, that is, the estimates of inheritance corresponding to the different percentiles of the distribution.

- Whole quantile function  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_q.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_q.pdf)
- Zoom on the top 10%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_q\\_top10.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_q_top10.pdf)
- Zoom on the top 1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_q\\_top1.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_q_top1.pdf)
- Zoom on the top 0.1%  
[http://piketty.pse.ens.fr/files/Fournier2015\\_inh\\_q\\_top01.pdf](http://piketty.pse.ens.fr/files/Fournier2015_inh_q_top01.pdf)

# Bibliography

- Aaberge, R., Atkinson, A. B., Königs, S., and Lakner, C. (2015). From Classes to Copulas: Wages, Capital, and Top Incomes. Forthcoming.
- Acemoglu, D. (2015a). Topics in Inequality, Lecture 5, Superstars and Top Inequality. MIT teaching slides.
- Acemoglu, D. (2015b). Topics in Inequality, Lecture 8, Pareto Income and Wealth Distributions. MIT teaching slides.
- Aigner, D. J. and Goldberger, A. S. (1970). Estimation of Pareto's Law from Grouped Observations. *Journal of the American Statistical Association*, 65(330):712–723.
- Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*. London: Cambridge University Press.
- Alvaredo, F., Atkinson, A. B., Piketty, T., and Saez, E. (2013). The Top 1 Percent in International and Historical Perspective. *Journal of Economic Perspectives*, 27(3):3–20.
- Alvaredo, F., Atkinson, A. B., Piketty, T., and Saez, E. (2015). The World Top Incomes Database. <http://topincomes.g-mond.parisschoolofeconomics.eu/>.
- Aoki, S. and Nirei, M. (2014). Zipf's Law, Pareto's Law, and the Evolution of Top Incomes in the U.S. UTokyo Price Project Working Paper Series 023, University of Tokyo, Graduate School of Economics.
- Arnold, B. C. (2015). *Pareto Distributions*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC Press, 2nd edition.
- Atkinson, A. B. and Harrison, A. J. (1978). *Distribution of Personal Wealth in Britain*. Cambridge University Press.
- Atkinson, A. B. and Piketty, T., editors (2007). *Top Incomes over the Twentieth Century: A Contrast Between European and English-Speaking Countries*. Oxford University Press.
- Atkinson, A. B. and Piketty, T., editors (2010). *Top Incomes: A Global Perspective*. Oxford University Press.

- Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top Incomes in the Long Run of History. *Journal of Economic Literature*, 49(1):3–71.
- Benhabib, J. (2014). Wealth Distribution Overview. NYU teaching slides.
- Benhabib, J., Bisin, A., and Zhu, S. (2011). The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents. *Econometrica*, 79(1):123–157.
- Benhabib, J., Bisin, A., and Zhu, S. (2014). The Wealth Distribution in Bewley Models with Investment Risk. NBER Working Papers 20157, National Bureau of Economic Research.
- Benhabib, J. and Zhu, S. (2008). Age, Luck, and Inheritance. NBER Working Papers 14128, National Bureau of Economic Research.
- Cantelli, F. P. (1921). Sulla deduzione delle leggi di frequenza da considerazioni di probabilità. *Metron*, 1(3):83–91.
- Cantelli, F. P. (1929). Sulla legge di distribuzione dei redditi. *Giornale degli Economisti e Rivista di Statistica*, 69(11):850–852.
- Champernowne, D. G. (1953). A Model of Income Distribution. *The Economic Journal*, 63(250):318–351.
- Cho, J. S., Park, M.-H., and Phillips, P. C. B. (2015). Minimum Distance Testing and Top Income Shares in Korea. Cowles foundation discussion paper no. 2007, Yale University.
- Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51:661–703.
- Clementi, F. and Gallegati, M. (2005). Pareto’s Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States. Papers physics/0504217, arXiv.org.
- Cowell, F. A. (1998). Inheritance and the Distribution of Wealth. LSE Research Online Documents on Economics 2124, London School of Economics and Political Science, LSE Library.
- Cowell, F. A. (2009). *Measuring Inequality*. LSE Perspectives in Economic Analysis. Oxford University Press, <http://darp.lse.ac.uk/MI3>.
- Cowell, F. A. and Mehta, F. (1982). The Estimation and Interpolation of Inequality Measures. *The Review of Economic Studies*, 49(2):273–290.
- der Wijk, J. V. (1939). *Inkomens-en Vermogens-Verdeling (The Distribution of Income and Property)*. Netherlands Economic Institute.
- Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag New York.
- Dougherty, R. L., Edelman, A., and Hyman, J. M. (1989). Nonnegativity-, Monotonicity-, or Convexity-Preserving Cubic and Quintic Hermite Interpolation. *Mathematics of Computation*, 52(186):471–494.

- Esteban, J. (1986). Income-Share Elasticity and the Size Distribution of Income. *International Economic Review*, 27(2):439–444.
- Feenberg, D. R. and Poterba, J. M. (1993). Income Inequality and the Incomes of Very High-Income Taxpayers: Evidence from Tax Returns. In *Tax Policy and the Economy*, volume 7, pages 145–177. MIT Press.
- Fisk, P. R. (1961). The Graduation of Income Distributions. *Econometrica*, 29(2):171–185.
- Fritsch, F. N. and Carlson, R. E. (1980). Monotone Piecewise Cubic Interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246.
- Gabaix, X. (2009). Power Laws in Economics and Finance. *Annual Review of Economics*, 1(1):255–294.
- Gabaix, X. (2014). Power Laws in Economics: An Introduction. Prepared for the *Journal of Economic Perspectives*.
- Gabaix, X. and Landier, A. (2008). Why has CEO Pay Increased So Much? *The Quarterly Journal of Economics*, 123(1):49–100.
- Gibrat, R. (1931). *Les inégalités économiques*. Paris: Librairie du Recueil Sirey.
- Harrison, A. (1979). The Upper Tail of the Earnings Distribution: Pareto or Lognormal? *Economic Letters*, 2:191–195.
- Harrison, A. (1981). Earnings by Size: A Tale of Two Distributions. *The Review of Economic Studies*, 48(4):621–631.
- Herriot, J. G. and Reinsch, C. H. (1973). Algorithm 472: Procedures for Natural Spline Interpolation. *Communications of the Association for Computing Machinery*, 16(12):763–768.
- Hyman, J. M. (1983). Accurate Monotonicity Preserving Cubic. *SIAM Journal on Scientific and Statistical Computing*, 4(4):645–654.
- Hyman, J. M. and Larrouturou, B. (1982). The numerical differentiation of discrete functions using polynomial interpolation methods. In Thompson, J. F., editor, *Numerical Grid Generation for Numerical Solution of Partial Differential Equations*, pages 487–506. Elsevier North-Holland, New York.
- Johnson, N. O. (1937). The Pareto Law. *The Review of Economics and Statistics*, 19(1):20–26.
- Jones, C. I. (2014). Simple Models of Pareto Income and Wealth Inequality. Technical report, Stanford GSB and NBER.
- Jones, C. I. (2015). Pareto and Piketty: The Macroeconomics of Top Income and Wealth Inequality. *Journal of Economic Perspectives*, 29(1):29–46.

- Jones, C. I. and Kim, J. (2014). A Schumpeterian Model of Top Income Inequality. NBER Working Papers 20637, National Bureau of Economic Research.
- Kaldor, N. (1961). Capital Accumulation and Economic Growth. In Lutz, F. A. and Hague, D. C., editors, *The Theory of Capital*, pages 177–222. St. Martin’s Press.
- Kesten, H. (1973). Random difference equations and Renewal theory for products of random matrices. *Acta Mathematica*, 131(1):207–248.
- Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London: Macmillan.
- Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons.
- Kremer, M. (1993). The O-Ring Theory of Economic Development. *The Quarterly Journal of Economics*, 108(3):551–575.
- Kuznets, S. (1953). Shares of Upper Income Groups in Income and Savings. *New York: National Bureau of Economic Research*.
- Kvasov, B. I. (2000). *Methods of Shape-Preserving Spline Approximation*. World Scientific.
- Landais, C., Piketty, T., and Saez, E. (2011). *Pour une révolution fiscale: un impôt sur le revenu pour le XXI<sup>e</sup> siècle*. La république des idées. Seuil.
- Lorenz, M. O. (1905). Methods of Measuring the Concentration of Wealth. *American Statistical Association*, 9:209–219.
- Lydall, H. F. (1959). The Distribution of Employment Incomes. *Econometrica*, 27(1):110–115.
- Mandelbrot, B. (1960). The Pareto-Lévy Law and the Distribution of Income. *International Economic Review*, 1(2):79–106.
- Mandelbrot, B. (1961). Stable Paretian Random Functions and the Multiplicative Variation of Income. *Econometrica*, 29(4):517–543.
- McAllister, D. F., Passow, E., and Roulier, J. A. (1977). Algorithms for computing shape preserving spline interpolations to data. *Mathematics of Computation*, 31:717–725.
- McDonald, J. B. (1984). Some Generalized Functions for the Size Distribution of Income. *Econometrica*, 52(3):647–665.
- McDonald, J. B. and Ransom, M. R. (1979). Functional Forms, Estimation Techniques and the Distribution of Income. *Econometrica*, 47(6):1513–1525.
- McLure, M. and Wood, J. C. (1999). *Vilfredo Pareto: Critical Assessments*. Critical Assessments of Leading Economists. Routledge.

- Metcalfe, C. E. (1969). The Size Distribution of Personal Income During the Business Cycle. *The American Economic Review*, 59(4):657–668.
- Mitzenmacher, M. (2003). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251.
- Modigliani, F. (1986). Life Cycle, Individual Thrift, and the Wealth of Nations. *The American Economic Review*, 76(3):297–313.
- Moll, B. (2012a). Inequality and Financial Development: A Power-Law Kuznets Curve. Technical report, Princeton University.
- Moll, B. (2012b). Why Piketty Says  $r - g$  Matters for Inequality. Princeton University supplementary lecture notes.
- Moll, B. (2014). Lecture 6: Income and Wealth Distribution. Princeton University teaching slides.
- Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Series in Statistics. Springer-Verlag New York, 2nd edition.
- Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46:323–351.
- Nirei, M. (2009). Pareto Distributions in Economic Growth Models. IIR Working Paper 09-05, Institute of Innovation Research, Hitotsubashi University.
- Pareto, V. (1967 (originally published in 1896)). Écrits sur la courbe de la répartition de la richesse. In *Œuvres complètes de Vilfredo Pareto*. Giovanni Busino, Librairie Droz, Genève.
- Persky, J. (1992). Retrospectives: Pareto’s Law. *Journal of Economic Perspectives*, 6(2):181–192.
- Piketty, T. (1998). Les hauts revenus face aux modifications des taux marginaux supérieurs de l’impôt sur le revenu en France, 1970-1996. Working Paper, CEPREMAP.
- Piketty, T. (2001). *Les Hauts revenus en France au 20e siècle : inégalités et redistribution, 1901-1998*. Paris: B. Grasset.
- Piketty, T. (2011). On the Long-Run Evolution of Inheritance: France 1820–2050. *The Quarterly Journal of Economics*, 126(3):1071–1131.
- Piketty, T. (2013). *Le capital au XXI<sup>e</sup> siècle*. Seuil.
- Piketty, T. and Zucman, G. (2014). Capital is Back: Wealth-Income Ratios in Rich Countries 1700-2010. *The Quarterly Journal of Economics*, 129(3):1255–1310.
- Piketty, T. and Zucman, G. (2015). Wealth and inheritance in the long run. In *Handbook of Income Distribution*, volume 2B, pages 1303–1368. North Holland.

- Reed, W. J. (2001). The Pareto, Zipf and other power laws. *Economic Letters*, 74:15–19.
- Rodriguez, A. (2014). Does growth promote Equality? A Note on Piketty’s capital on the twenty-first century. *Economics Bulletin*, 34(3):2044–2050.
- Rosen, S. (1981). The Economics of Superstars. *The American Economic Review*, 71(5):845–858.
- Roy, A. D. (1950). The Distribution of Earnings and of Individual Output. *Economic Journal*, 60:489–505.
- Runge, C. (1901). Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Zeitschrift für Mathematik und Physik*, 46:224–243.
- Saez, E. and Zucman, G. (2014). Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data. NBER Working Papers 20625, National Bureau of Economic Research.
- Salem, A. B. Z. and Mount, T. D. (1974). A Convenient Descriptive Model of Income Distribution: The Gamma Density. *Econometrica*, 42(6):1115–1127.
- Simon, H. A. (1955). On a Class of Skew Distribution Functions. *Biometrika*, 42(3-4):425–440.
- Singh, S. K. and Maddala, G. S. (1976). A Function for Size Distribution of Incomes. *Econometrica*, 44(5):963–970.
- Slottje, D. J. (1984). A measure of income inequality in the U.S. for the years 1952–1980 based on the beta distribution of the second kind. *Economic Letters*, 15:369–375.
- Späth, H. (1969). Exponential Spline Interpolation. *Computing*, 4:225–233.
- Stiglitz, J. E. (1969). Distribution of Income and Wealth Among Individuals. *Econometrica*, 37(3):382–397.
- Stiglitz, J. E. (2015). New Theoretical Perspectives on the Distribution of Income and Wealth among Individuals: Part II. Equilibrium Wealth Distributions. NBER Working Papers 21190, National Bureau of Economic Research.
- Thatcher, A. R. (1968). The Distribution of Earnings of Employees in Great Britain. *Journal of the Royal Statistical Society. Series A (General)*, 131(2):133–181.
- Thurow, L. C. (1970). Analyzing the American Income Distribution. *The American Economic Review*, 60(2):261–269.
- Wold, H. O. A. and Whittle, P. (1957). A Model Explaining the Pareto Distribution of Wealth. *Econometrica*, 25(4):591–595.
- Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 213(402-410):21–87.