the review of income and wealth

Review of Income and Wealth
Series 68, Number 1, March 2022
DOI: 10.1111/roiw.12510

GENERALIZED PARETO CURVES: THEORY AND APPLICATIONS

BY THOMAS BLANCHET* and THOMAS PIKETTY

Paris School of Economics
AND
JULIETTE FOURNIER

Massachusetts Institute of Technology

We define *generalized Pareto curves* as the curve of inverted Pareto coefficients $b(p)$, where $b(p)$ is the ratio between average income above rank p and the p-th quantile $Q(p)$ (i.e., $b(p) = \mathbb{E}[X|X > Q(p)]/Q(p)$). We use them to characterize income distributions. We develop a method to flexibly recover a continuous distribution based on tabulated income data as is generally available from tax authorities, which produces smooth and realistic shapes of generalized Pareto curves. Using detailed tabulations from quasi-exhaustive tax data, we show the precision of our method. It gives better results than the most commonly used interpolation techniques for the top half of the distribution.

JEL Codes: C14, D31

Keywords: income, inequality, Pareto, power law, interpolation

1. INTRODUCTION

It has long been known that the upper tail of the distribution of income and wealth can be approximated by a Pareto distribution, or power law (Pareto, 1896). This fact has been widely used in the empirical literature on inequality to overcome certain limitations of the data. In particular, Pareto interpolation methods have been used by Kuznets (1953), Atkinson and Harrison (1978), Piketty (2001, 2003), Piketty and Saez (2003) and the subsequent literature exploiting historical tax tabulations to construct long-run series on income and wealth inequality. The widespread applicability of this functional form is often justified using models where income and wealth evolve according to random multiplicative shocks (Champernowne, 1953; Simon, 1955; Wold and Whittle, 1957). Recent contributions have shown how such models can account for both the levels and the changes

Note: We thank the editor and two referees for their comments. The author gratefully acknowledges funding from the European Research Council (ERC Grant 856455) and from the French National Research Agency (EUR Grant ANR-17-EURE-0001). All R programs developed in this article are available at http://wid.world/gpinter, where we also provide a set of online tools to estimate and manipulate distributions of income and wealth on the basis of simple tabulated data files (such as those provided by tax administrations and statistical institutes) and generalized Pareto interpolation methods.

*Correspondence to: Thomas Piketty, Paris School of Economics, 48 Boulevard Jourdan, 75014 Paris, France (thomas.blanchet@wid.world, piketty@psemail.eu).

in inequality (Nirei, 2009; Benhabib *et al*., 2011; Piketty and Zucman, 2015; Jones and Kim, 2018; Jones, 2015; Benhabib and Bisin, 2016; Gabaix *et al*., 2016).

However, although the Pareto approximation is acceptable for some purposes, it is not entirely correct, not even at the top. As a result, empirical methods that strictly rely on it can miss important features of the distribution (Atkinson, 2017; Jenkins, 2017). If we want to better exploit the data at our disposal, and also to better understand the economic mechanisms giving rise to the observed distributions of income and wealth, we need to move beyond standard Pareto distributions.

In this article, we develop the flexible notion of *generalized Pareto curve* to characterize and estimate income and wealth distributions. A generalized Pareto curve is defined as the curve of inverted Pareto coefficients $b(p)$, where $0 \leq p < 1$ is the (normalized) rank, and $b(p)$ is the ratio between average income or wealth above rank p and the p-th quantile $Q(p)$ (i.e., $b(p) = \mathbb{E}[X|X > Q(p)]/Q(p)$). If the tail follows a standard Pareto distribution, the coefficient $b(p)$ is constant. For example, if $b(p) = 2$ at the top of the wealth distribution, then the average wealth of individuals above €1 million is €2 million, the average wealth of individuals above €10 million is €20 million, and so on. In practice, we find that $b(p)$ does vary within the upper tail of observed income and wealth distributions (including within the top 10 percent or the top 1 percent), but that the curves $b(p)$ are relatively similar (typically U-shaped).

Generalized Pareto curves are a particularly useful tool to describe distributions with a power-law tail. Looking at them reveals significant deviations of real distribution of income and wealth from strict Pareto behavior, even at the very top. We exploit this framework to develop an improved methodological approach for the estimation of income and wealth distributions using tax data, which is often available solely in the form of tabulations with a finite number of inverted Pareto coefficients $b_1, \ldots, b_K$ and thresholds $q_1, \ldots, q_K$ observed for ranks $p_1, \ldots, p_K$. We call it *generalized Pareto interpolation*. Existing methods typically rely on diverse Paretian assumptions (or even less realistic ones) that, by construction, blur or even erase deviations from the standard Pareto distribution. We show that considering how the Pareto coefficient $b(p)$ varies can dramatically improve the way we produce statistics on income and wealth inequality, especially for the top and with few data points. Using quasi-exhaustive (i.e., including the full population, at least at the top) annual micro files of income tax returns available in the US and France over the 1962–2014 period (a time of rapid and large transformation of the distribution of income, particularly in the US), we show the precision of the method. That is, based on the information for a small number of ranks (e.g., $p_1 = 10$ percent, $p_2 = 50$ percent, $p_3 = 90$ percent, and $p_4 = 99$ percent), we can recover the top half of the distribution with remarkable precision. The method also gives reasonably good results for the bottom (between $p = 10$ percent and $p = 50$ percent) and generates a consistent and smooth distribution with a continuous density. In fact, we find that the precision of the method is such that it is often preferable to use tabulations based on exhaustive data rather than microdata from a non-exhaustive subsample of the population, even for subsamples considered very large by statistical standards. For example, a subsample of 100000 observations can typically lead to a mean relative error of about 3 percent on the top 5 percent share, whereas a tabulation based on exhaustive data that includes the percentile

ranks p = 10 percent, 50 percent, 90 percent, and 99 percent gives a mean relative error of less than 0.5 percent. For the top 0.1 percent share, the same error can reach 20 percent with the same subsample, whereas the same tabulation yields an error below 4 percent.

We believe that the methodology developed in this article can help researchers avoid excessive reliance on restrictive assumptions when using tabulated data, which is still commonplace in some areas of research. To that end, we developed an R (R Core Team, 2016) package, named `gpinter`, that implements the methods described in this article and make them easily available to researchers. We also provide a web interface built on top of this package (Chang et al., 2017), available at http://wid.world/gpinter, to estimate and manipulate distributions of income and wealth on the basis of simple tabulated data files (such as those provided by tax administrations and statistical institutes) and generalized Pareto interpolation methods. These tools have successfully been used to estimate series of the income distribution in the Middle-East (Alvaredo *et al*., 2019), Poland (Bukowski and Novokmet, 2017), Brazil (Morgan, 2017), India (Chancel and Piketty, 2019), Russia (Novokmet *et al*., 2018), Ivory Coast (Czajka, 2017), China (Piketty *et al*., 2019), France (Garbinti *et al*., 2018), and India (Chancel and Piketty, 2019). Furthermore, we plan to use them to keep expanding the World Inequality Database (wid.world). However, the method is not limited to the production of specific inequality statistics: it outputs a continuous and consistent distribution which, depending on what is most practical, can be characterized by its density, its cumulative distribution function, its quantile function, or its Lorenz curve. As such, it offers readily available tools for using tabulated data in a variety of contexts (see, e.g., Bierbrauer *et al*. (2021) in the field of optimal taxation).

The rest of the article is organized as follows. In Section 2, we provide the formal definition and the key properties of generalized Pareto curves $b(p)$. In Section 3, we present our generalized Pareto interpolation method, which is based on a transformation of $b(p)$. In Section 4, we test its precision and compare it to other interpolation methods using individual income data for the US and France covering the 1962–2014 period. In Section 5, we consider extensions of the framework that allows us to further discuss the level of precision that we can expect from our method in comparison to others.

2. Generalized Pareto Curves

2.1. *Definition and Properties*

We characterize the distribution of income or wealth by a random variable X with cumulative distribution function (CDF) F. We assume that X is integrable (i.e., $\mathbb{E}[\,|X|\,] < +\infty$) and that F is differentiable over a domain $D = [a, +\infty[$ or $D = \mathbb{R}$. We denote f the probability density function (PDF) and Q the quantile function. Our definition of the inverted Pareto coefficient follows the one first given by Fournier (2015).

Definition 1 (Inverted Pareto coefficient)For any income level $x > 0$, the inverted Pareto coefficient is $b^*(x) = \mathbb{E}[X \mid X > x]$, or:

$$b^*(x) = \frac{1}{(1-F(x))x} \int_x^{+\infty} z f(z)\, \mathrm{d}z.$$

We can express it as a function of the fractile p with $p = F(x)$ and $b(p) = b^*(x)$:

$$b(p) = \frac{1}{(1-p)Q(p)} \int_p^1 Q(u)\, \mathrm{d}u.$$

If X follows a Pareto distribution with coefficient α and lower bound $\overline{x}$, so that $F(x) = 1 - (\overline{x}/x)^{\alpha}$, then $b(p) = \alpha/(\alpha-1)$ is constant (a property also known as van der Wijk's (1939) law), and the top $100 \times (1-p)$ percent share is an increasing function of b and is equal to $(1-p)^{1/b}$. Otherwise, $b(p)$ will vary. We can view the inverted Pareto coefficient as an indicator of the tail's fatness, or similarly an indicator inequality at the top. It also naturally appears in some economic contexts, such as optimal taxation formulas (Saez, 2001). We favor looking at them as a function of the fractile p rather than the income x, because it avoids differences because of scaling, and make them more easily comparable over time and between countries. We call *generalized Pareto curve* the function $b{:}p \mapsto b(p)$ defined over $[\overline{p}, 1[$ with $\overline{p} = F(\overline{x})$.[1] (Where the notation $[x,y[$ means the interval containing all real numbers t such that $x \leq t < y$.)

Proposition 1 If X satisfies the properties stated above, then b is differentiable and for all $\overline{p} \leq p \in <1, 1 - b(p) + (1-p)b'(p) \leq 0$ and $b(p) \geq 1$.

The proof of that proposition—as well as all the others in this section—is available in Section A.3 in appendix. The definition of $b(p)$ directly implies $b(p) \geq 1$. The fact that the quantile function is increasing implies $1 - b(p) + (1-p)b'(p) \leq 0$. Conversely, for $0 \leq \overline{p} < 1$ and $\overline{x} > 0$, any function b: $[\overline{p}, 1[\to \mathbb{R}$ that satisfies property 1 uniquely defines the top $(1-\overline{p})$ fractiles of a distribution with $\overline{p} = F(\overline{x})$.

Proposition 2 If X is defined for $x > \overline{x}$ by $F(\overline{x}) = \overline{p}$ and the generalized Pareto curve b: $[\overline{p}, 1[\to \mathbb{R}$, then for $p \geq \overline{p}$, the p-th quantile is:

$$Q(p) = \overline{x} \frac{(1-\overline{p})b(\overline{p})}{(1-p)b(p)} \exp\left(- \int_{\overline{p}}^{p} \frac{1}{(1-u)b(u)}\, \mathrm{d}u \right).$$

The coefficient defined in 1 is only one of several "local" notion Pareto coefficients that may be defined using a similar logic. In Appendix A, we discuss other properties of generalized Pareto curves, how they relate to the theory of power laws, and the economics models of the distribution of income and wealth.

[1]We solely consider inverted Pareto coefficient above a strictly positive threshold $\overline{x} > 0$, because they have a singularity at zero and a less clear meaning below that. The threshold must thus correspond to a percentile above the share of people with negative or zero income, typically at least p = 10 percent.

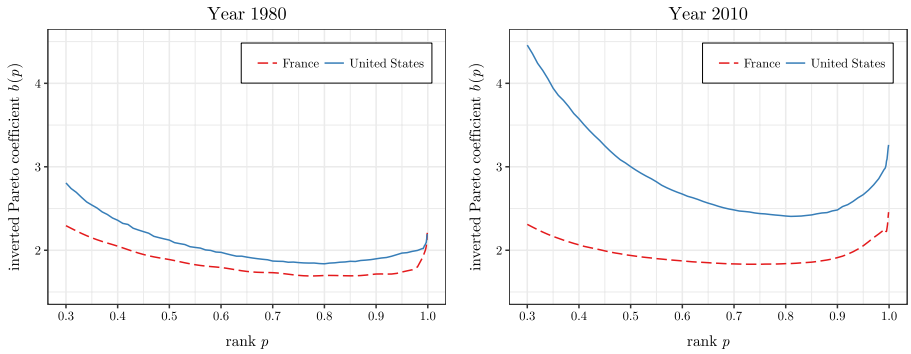

Figure 1. Generalized Pareto Curves of DINA Income
Sources: Piketty *et al*. (2018) (United States), Garbinti *et al*. (2018) (France). [Colour figure can be viewed at wileyonlinelibrary.com]

2.2. *Pareto Curves in Practice*

We now consider a sample $(X_1, \ldots, X_n)$ of n iid. realizations of X. We write $X_{(r)}$ the r-th order statistic (i.e., the r-th largest value). Let $x \mapsto \lfloor x \rfloor$ denote the floor function. The natural estimator of the inverted Pareto coefficient may be written:[2]

$$\widehat{b}_n(p) = \frac{1}{(n - \lfloor np \rfloor) X_{(\lfloor np \rfloor + 1)}} \sum_{k = \lfloor np \rfloor + 1}^{n} X_{(k)}.$$

Figure 1 depicts the empirical Pareto curves for the distribution of Distributional National Accounts (DINA) income in France and in the US in 1980 and 2010, based on quasi-exhaustive income tax data. The curve has changed a lot more in the US than in France, which reflects the well-known increase in inequality that the US has experienced over the period. In 2010, the inverted Pareto coefficients are much higher in the US than in France, which means that the tail is fatter, and the income distribution more unequal.

In both countries, $b(p)$ does appear to converge toward a value strictly above 1, which confirms that the distribution of income is an asymptotic power law. However, the coefficients vary significantly, even within the top decile group, so that the strict Pareto assumption will miss important patterns in the distribution. Because $b(p)$ rises within the top 10 percent of the distribution, inequality in both France and the US is in fact even more skewed toward the very top than what the standard Pareto model suggests, and the amount by which inverted Pareto coefficients vary is not negligible. For the US, in 2010, at its lowest point (near $p = 80$ percent, $b(p)$ is around 2.4. If it were a strict Pareto distribution, it would correspond to the top 1 percent owning 15 percent of the income. However, the asymptotic value is closer to 3.3, which would mean a top 1 percent share of 25 percent.

[2]Note that for $(n-1)/n \leq p < 1$, we have $\widehat{b}_n(p) = 1$ regardless of the distribution of X. This speaks to the impossibility of directly estimating asymptotic quantities from a finite sample. However, with fiscal data, for which samples are extremely large, we need not be concerned by the problem until extremely narrow top income groups.

Although empirical evidence leads us to reject the strict Pareto assumption, we can notice that the generalized Pareto curves are U-shaped. We observe that fact for all countries and time periods for which we have sufficient data.

3. Generalized Pareto Interpolation

The tabulations of income or wealth such as those provided by tax authorities and national statistical institutes typically take the form of K fractiles $0 \leq p_1 < \cdots < p_K < 1$ of the population, alongside their income quantiles $q_1 < \cdots < q_K$ and the income share of each bracket $[p_k, p_{k+1}]$. That last element may take diverse forms (top income shares, bottom income shares, average income in the brackets, average income above the bracket, etc.), all of which are just different ways of presenting the same information. The interpolation method that we now present uses the way inverted Pareto coefficients vary smoothly to estimate a complete distribution based solely on that information: we call it *generalized Pareto interpolation*. Note that we assume that we know both the thresholds and the shares of each bracket. In some cases, only one of these is available. Although a similar method could be extended to these settings (especially when we know the shares), we leave this for future research.

The first goal of the method is to be as flexible as we are allowed to be: i.e., we do not force the estimated distribution into a predetermined shape. We stress that a fully nonparametric approach is not possible here because of the lack of a suitable asymptotic framework.(The number of brackets would have to go to infinity, which is not a good approximation of real-life settings.) But we can still get a lot more flexibility than a strict Pareto model by introducing a large enough number of parameters. The second goal is to generate a solution with desirable properties. Indeed the interpolation problem is technically ill-posed as it has an infinite number of candidate solutions. Our method overcomes that issue by looking for a "regular" curve of Pareto coefficients.

Our method combines three components, which solve different aspects of the problem. First, we interpolate the generalized Pareto curve in a way that maximizes its smoothness while satisfying two sets of constraints: those related to the quantiles and those related to the means. Second, we enforce if necessary the constraint that the quantile function is increasing by finding an admissible solution that is as close as possible to the original one. Finally, we deal separately with last bracket, for which the interpolation is not possible because of the lack of an endpoint in the interval.

For the exposition of the method, we will set aside sampling-related issues and treat empirical quantities as equivalent to their theoretical counterpart. However, we come back to that issue in Section 5.

3.1. *Interpolation of the Pareto Coefficients*

The tabulations let us compute $b(p_1), \ldots, b(p_K)$ directly. However, interpolating the curve $b(p)$ based solely on those points offers no guarantee that the resulting function will be consistent with the input data on quantiles. To that end, the

interpolation needs to be constrained. To do so in a computationally efficient and analytically tractable way, we start from the following function:

$$\forall x \geq 0 \qquad \varphi(x) = -\log \int_{1-\mathrm{e}^{-x}}^{1} Q(p)\,\mathrm{d}p,$$

which is essentially a transform of the Lorenz curve $L(p)$:

$$\varphi(x) = -\log((1 - L(p))\mathbb{E}[X]),$$

with $p = 1 - \mathrm{e}^{-x}$. The value of φ at each point $x_k = -\log(1 - p_k)$ can therefore be estimated directly from the data in the tabulation. Moreover:

$$\forall x \geq 0 \qquad \varphi'(x) = \mathrm{e}^{\varphi(x)-x} Q(1 - \mathrm{e}^{-x}) = 1/b(1 - \mathrm{e}^{-x}),$$

which means that the generalized Pareto coefficient $b(p)$ is equal to $1/\varphi'(x)$. Therefore, the value of $\varphi'(x_k)$ for $k \in \{1,\ldots,K\}$ is also given by the tabulation.

Because of the bijection between $(p,b(p),Q(p))$ and $(x, \varphi(x), \varphi'(x))$, the problem of interpolating $b(p)$ in a way that is consistent with $Q(p)$ is identical to that of interpolating the function φ, whose value and first derivative are known at each point x_k.

We assume that we know a set of points $\{(x_k, y_k, s_k), 1 \leq k \leq K\}$ that correspond to the values of $\{(x_k, \varphi(x_k), \varphi'(x_k)), 1 \leq k \leq K\}$, and we seek a sufficiently smooth function $\widehat{\varphi}$ such that:

$$(1) \qquad \forall k \in \{1, \ldots, K\} \qquad \widehat{\varphi}(x_k) = \varphi(x_k) = y_k \qquad \widehat{\varphi}'(x_k) = \varphi'(x_k) = s_k.$$

By sufficiently smooth, we mean that φ should be at least twice continuously differentiable. That requirement is necessary for the estimated Pareto curve (and by extension the quantile function) to be once continuously differentiable, or, put differently, not to exhibit any asperity at the fractiles included in the tabulation.

To get an appropriate function, we rely on quintic splines—i.e., piecewise polynomials of degree 5 defined over each bracket. The quintic spline is fully determined by three quantities at each boundary: the value of the polynomial and its first and second derivatives. The value (x_k) and the first derivative (s_k) are already fixed by the problem. The value of the second derivative (a_k) is a free parameter to be set. To pick appropriate values for $a_1, \ldots, a_k$, we follow the usual approach of imposing additional regularity conditions at the boundaries. We have a system of $K-2$ equations, linear in $a_1, \ldots, a_k$, defined by:

$$\forall k \in \{2, \ldots, K-1\} \qquad \widehat{\varphi}'''_{k-1}(x_k) = \widehat{\varphi}'''_k(x_k).$$

Two additional equations are required for that system to have a unique solution. One solution is to use predetermined values for a_1 and a_K (known as the "clamped spline"). Another, known as the "natural spline," sets:

$$\widehat{\varphi}'''_1(x_1) = 0 \qquad \text{and} \qquad \widehat{\varphi}'''_{K-1}(x_K) = 0.$$

Both approaches are equivalent to the minimization of an irregularity criterion (e.g., Lyche and M⊘rken, 2002):

$$\min_{a_1,\ldots,a_K} \int_{x_1}^{x_k} \{\,\widehat{\varphi}'''(x)\,\}^2\,\mathrm{d}x$$

subject to fixed values for a_1 and a_K (clamped spline) or not (natural spline).

We adopt a hybrid approach, in which a_1 is determined through $\widehat{\varphi}_1'''(x_1) = 0$, but where a_K is estimated separately using the two-point finite difference:

$$a_K = \frac{s_K - s_{K-1}}{x_K - x_{K-1}}.$$

Because the function is close to linear near x_K, it yields results that are generally similar to traditional natural splines. However, that estimation of $\varphi''(x_K)$ is also more robust, so we get more satisfactory results when the data exhibit potentially troublesome features.

Finding the actual value of each parameter amounts to solving a linear system of equations. We provide the detailed algebraic expressions in Appendix C.

3.2. *Enforcing Admissibility Constraints*

The interpolation method presented above does not guarantee that the estimated generalized Pareto curve will satisfy property 1—or equivalently that the quantile will be an increasing function. In most situations that constraint need not be enforced, because it is not binding: the estimated function spontaneously satisfies it. However, it may occasionally not be the case, so that estimates of quantiles of averages at different points of the distribution may be mutually inconsistent. To solve that problem, we present an *ex post* adjustment procedure that constrains appropriately the interpolated function.

We can express the quantile as a function of φ:

$$\forall x \geq 0 \qquad Q(1 - \mathrm{e}^{-x}) = \mathrm{e}^{x - \varphi(x)}\varphi'(x).$$

Therefore:

$$\forall x \geq 0 \qquad Q'(1 - \mathrm{e}^{-x}) = \mathrm{e}^{2x - \varphi(x)}[\varphi''(x) + \varphi'(x)(1 - \varphi'(x))].$$

Therefore, the estimated quantile function is increasing if and only if:

$$\forall x \geq 0 \qquad \Phi(x) = \widehat{\varphi}''(x) + \widehat{\varphi}'(x)(1 - \widehat{\varphi}'(x)) \geq 0. \tag{2}$$

The polynomial Φ (of degree 8) needs to be positive. There are no simple necessary and sufficient conditions on the parameters of the spline that can ensure such a constraint. However, it is possible to derive conditions that are only sufficient, but general enough to be used in practice. We use conditions based on the Bernstein representation of polynomials, as derived by Cargo and Shisha (1966):

Theorem 1 (Cargo and Shisha (1966))Let $P(x) = c_0 + c_1x_1 + \cdots + c_nx^n$ be a polynomial of degree $n \geq 0$ with real coefficients. Then:

$$\forall x \in [0,1] \qquad \min_{0 \leq i \leq n} b_i \leq P(x) \leq \max_{0 \leq i \leq n} b_i,$$

where:

$$b_i = \sum_{r=0}^{n} c_r \binom{i}{r} \Big/ \binom{n}{r}.$$

To ensure that the quantile is increasing over $[x_k, x_{k+1}]$ $(1 \le k < K)$, it is therefore enough to enforce the constraint that $b_i \ge 0$ for all $0 \le i \le 8$, where b_i is defined as in Theorem 1 with respect to the polynomial $x \mapsto \Phi(x_k + x(x_{k+1} - x_k))$. Those nine conditions are all explicit quadratic forms in $(y_k, y_{k+1}, s_k, s_{k+1}, a_k, a_{k+1})$, so we can compute them and their derivatives easily.

To proceed, we start from the unconstrained estimate from the previous section. We set $a_k = -s_k(1 - s_k)$ for each $1 \le k \le K$ if $a_k + s_k(1 - s_k) < 0$, which ensures that condition (2) is satisfied at least at the interpolation points. Then, over each segment $[x_k, x_{k+1}]$, we check whether the condition $\Phi(x) \ge 0$ is satisfied for $x \in [x_k, x_{k+1}]$ using the Theorem 1, or more directly by calculating the values of Φ over a tight enough grid of $[x_k, x_{k+1}]$. If so, we move on to the next segment. If not, we consider $L \ge 1$ additional points $(x_1^*, \ldots, x_L^*)$ such that $x_k < x_1^* < \cdots < x_L^* < x_{k+1}$, and we redefine the function $\widehat{\varphi}_k$ over $[x_k, x_{k+1}]$ as:

$$\tilde{\varphi}_k(x) = \begin{cases} \varphi_0^*(x) & \text{if} \quad x_k \le x < x_1^* \\ \varphi_\ell^*(x) & \text{if} \quad x_\ell^* \le x < x_{\ell+1}^* \\ \varphi_L^*(x) & \text{if} \quad x_L^* \le x < x_{k+1}, \end{cases}$$

where the φ_ℓ^* $(0 \le \ell \le L)$ are quintic splines such that for all $1 \le \ell < L$:

$$\begin{array}{llllll} \varphi_0^*(x_k) & = y_k & (\varphi_0^*)'(x_k) & = s_k & (\varphi_0^*)''(x_k) & = a_k \\ \varphi_L^*(x_{k+1}) & = y_{k+1} & (\varphi_L^*)'(x_{k+1}) & = s_{k+1} & (\varphi_L^*)''(x_{k+1}) & = a_{k+1} \\ \varphi_\ell^*(x_\ell^*) & = y_\ell^* & (\varphi_\ell^*)'(x_\ell^*) & = s_\ell^* & (\varphi_\ell^*)''(x_\ell^*) & = a_\ell^* \\ \varphi_\ell^*(x_{\ell+1}^*) & = y_{\ell+1}^* & (\varphi_\ell^*)'(x_{\ell+1}^*) & = s_{\ell+1}^* & (\varphi_\ell^*)''(x_{\ell+1}^*) & = a_{\ell+1}^* \end{array}$$

and $y_\ell^*, s_\ell^*, a_\ell^*$ $(1 \le \ell \le L)$ are parameters to be adjusted. In simpler terms, we divided the original spline into several smaller ones, thus creating additional parameters that can be adjusted to enforce the constraint. We set the parameters $y_\ell^*, s_\ell^*, a_\ell^*$ $(1 \le \ell \le L)$ by minimizing the L^2 norm between the constrained and the unconstrained estimate, subject to the $9 \times (L+1)$ conditions that $b_i^\ell \ge 0$ for all $0 \le i \le 8$ and $0 \le \ell \le L$:

$$\min_{\substack{y_\ell^*, s_\ell^*, a_\ell^* \\ 1 \le \ell \le L}} \int_{x_k}^{x_{k+1}} \{\widehat{\varphi}_k(x) - \tilde{\varphi}_k(x)\}^2 \, \mathrm{d}x \qquad \text{st.} \qquad b_i^\ell \ge 0 \quad (0 \le i \le 8 \text{ and } 0 \le \ell \le L),$$

where the b_i^ℓ are defined as in Theorem 1 for each spline ℓ. The objective function and the constraints all have explicit analytical expressions, and so does their

gradients. We solve the problem with standard numerical methods for nonlinear constrained optimization.[3,4]

3.3. *Extrapolation in the Last Bracket*

The interpolation procedure only applies to fractiles between p_1 and p_K, but we generally also want an estimate of the distribution outside of this range, especially for $p > p_K$.[5] Because there is no direct estimate of the asymptotic Pareto coefficient $lim_{p\to 1} b(p)$, it is not possible to interpolate as we did for the rest of the distribution: we need to extrapolate it.

The extrapolation in the last bracket should satisfy the constraints imposed by the tabulation (on the quantile and the mean). In accordance with the principle of a regular Pareto curve, it should also ensure derivability of the quantile function at the boundary. To do so, we use the information contained in the four values (x_K, y_K, s_K, a_K) of the interpolation function at the last point. Therefore, we need an appropriate functional form for the last bracket with enough degrees of freedom to satisfy all the constraints. To that end, we turn to the generalized Pareto distribution.

Definition 1 (Generalized Pareto distribution)Let $\mu \in \mathbb{R}$, $\sigma \in]0,+\infty[$, and $\xi \in \mathbb{R}$. X follows a generalized Pareto distribution if for all $x \geq \mu$ ($\xi \geq 0$) or $\mu \leq x \leq \mu - \sigma/\xi$ ($\xi < 0$):

$$\mathbb{P}\{X \leq x\} = \mathrm{GPD}_{\mu,\sigma,\xi}(x) = \begin{cases} 1 - \left(1 + \xi \dfrac{x-\mu}{\sigma}\right)^{-1/\xi} & \text{for } \xi \neq 0 \\ 1 - \mathrm{e}^{-(x-\mu)/\sigma} & \text{for } \xi = 0. \end{cases}$$

μ is called the location parameter, σ the scale parameter, and ξ the shape parameter.

The generalized Pareto distribution is a fairly general family that includes as special cases the strict Pareto distribution ($\xi > 0$ and $\mu = \sigma/\xi$), the (shifted) exponential distribution ($\xi = 0$), and the uniform distribution ($\xi = -1$). It was popularized as a model of the tail of other distributions in extreme value theory by Pickands (1975) and Balkema and de Haan (1974), who showed that for a large class of distributions, the tail converges toward a generalized Pareto distribution.

If $X \sim \mathrm{GPD}(\mu,\sigma,\xi)$, the generalized Pareto curve of X is:

$$b(p) = 1 + \frac{\xi\sigma}{(1-\xi)[\sigma + (1-p)^{\xi}(\mu\xi - \sigma)]}.$$

We will focus on cases where $0 < \xi < 1$, so that the distribution is a power law at the limit ($\xi > 0$), but its mean remains finite ($\xi < 1$). When $\xi\mu = \sigma$, the generalized

[3]For example, standard sequential quadratic programming (Kraft, 1994) or augmented Lagrangian methods (Conn *et al*., 1991; Birgin and Martinez, 2008). See NLopt for details and open source implementations of such algorithms: http://ab-initio.mit.edu/wiki/index.php/NLopt_Algorithms.

[4]Adding one point at the middle of the interval is usually enough to enforce the constraint, but more points may be added if convergence fails.

[5]It is always possible to set $p_1 = 0$ if the distribution has a finite lower bound.

Pareto curve is constant, and the distribution is a strict power law with Pareto coefficient $b = 1/(1-\xi)$. That value also corresponds in all cases to the asymptotic coefficient $lim_{p\to 1} b(p) = 1/(1-\xi)$. However, there are several ways for the distribution to converge toward a power law, depending on the sign of $\mu\xi-\sigma$. When $\mu\xi-\sigma > 0$, $b(p)$ converges from below, increasing as $p\to 1$, so that the distribution gets more unequal in higher brackets. Conversely, when $\mu\xi-\sigma < 0$, $b(p)$ converges from above, and decreases as $p\to 1$, so that the distribution is more equal in higher brackets.

The generalized Pareto distribution can match a wide diversity of profiles for the behavior of $b(p)$, while offering the right number of degrees of freedom for our purpose. It has been shown to provide a better fit to the top income distribution than the standard Pareto distribution (Jenkins, 2017; Charpentier and Flachaire, 2019). In the context of our method, however, the value of its parameters is not of direct interest. In particular, the setting does not allow for a particularly accurate estimation of the asymptotic Pareto coefficient, and we do not focus on providing such an estimate. However, we can use it to find a reasonable functional form that makes an efficient use of the information at our disposal on the mean, the quantile, and its derivative at the last threshold. The generalized Pareto distribution offers a way to extrapolate the coefficients $b(p)$ in a way that is consistent with all the input data and preserves the regularity of the Pareto curve.

We assume that, for $p > p_K$, the distribution follows a generalized Pareto distribution with parameters (μ,σ,ξ), which means that for $q > q_K$ the CDF is:

$$F(q) = p_K + (1-p_K)\mathrm{GPD}_{\mu,\sigma,\xi}(q).$$

For the CDF to remain continuous and differentiable, we need $\mu = q_K$ and $\sigma = (1-p_K)/F'(q_K)$, where $F'(q_K)$ comes from the interpolation method of Section 3.1. Finally, for the Pareto curve to remain continuous, we need $b(p_K)$ equal to $1+\sigma/(\mu(1-\xi))$, which gives the value of ξ. That is, if we set the parameters (μ,σ,ξ) equal to:

$$\begin{aligned}
\mu &= s_K \mathrm{e}^{x_K - y_K} \\
\sigma &= (1-p_K)(a_K + s_K(1-s_K))\mathrm{e}^{2x_K - y_K} \\
\xi &= 1 - \frac{(1-p_K)\sigma}{\mathrm{e}^{-y_K} - (1-p_K)\mu},
\end{aligned}$$

then the resulting distribution will have a continuously differentiable quantile function and will match the quantiles and the means in the tabulation.

4. Tests Using Income Data from the US and France, 1962–2014

We test the quality of our interpolation method using data for the US (1962, 1954, and 1966–2014) and France (1994–2012). They correspond to cases for which we have detailed tabulations of the distribution of yearly pretax income based on quasi-exhaustive individual tax data (Garbinti *et al*., 2018; Piketty *et al*., 2018), so that we can know quantiles or shares exactly.

We call "DINA income" the income concept that we use as our benchmark, as it was defined and calculated in the context of the DINA project (Alvaredo

et al., 2020). The income that we consider includes all labor and capital income received by individuals. It also includes pension and unemployment insurance benefits and removes the corresponding social contributions. On the contrary, it does not remove income taxes and does not include other benefits (classified as social assistance benefits, rather than social insurance). These estimates are primarily based on administrative tax data and also use surveys to account for non-filers and tax-exempt income. See Piketty *et al*. (2018) and Garbinti *et al*. (2018) for a detailed definition and methodology. The inclusion of tax-exempt income is the main difference with the concept of "fiscal income" that was originally used in the top income literature (Atkinson and Piketty, 2007). It avoids having an income concept that is overly dependent on the local legislation of countries, making estimates more comparable. We also report comparisons using fiscal income directly in Appendix D. The statistical unit in both cases in the individual adult (age 20 or more), and income is split equally between adult household members. We compare the size of the error in generalized Pareto interpolation with alternatives most commonly found in the literature.

4.1. *Overview of Other Common Interpolation Methods*

We compare our interpolation method with the three main interpolation methods used in the top income literature (Atkinson, 2007). We designed our method primarily to improve the quality of estimates for the top of the distribution obtained from tax data, which explains our focus on these methods, and on the top half of the distribution. However, we also report results for the middle and the bottom of the distributions, which show that our method also works relatively well there.

There is a wide range of alternative interpolation approaches that are suited to various contexts. Some, like Jargowsky and Wheeler (2018), focus on cases where only the bracket thresholds and population share are available—while we consider cases in which the mean income in each bracket is also known. Other approaches seek to directly estimate a parametric model for the whole distribution: e.g., Villaseñor and Arnold (1989) and Kakwani and Podder (1976) fit a parametric model for Lorenz curves, and Chotikapanich *et al*. (2012) use the tabulation as moment conditions to fit a Beta II distribution. Our approach is less parametric and seeks to reproduce the statistics provided in the tabulation in input perfectly.

In Appendix D, we extend our comparison to some of these methods: one additional method based on the Pareto distribution method suggested by Cowell (2000, p. 158) and two methods that are fully parametric (Kakwani and Podder, 1976; Villaseñor and Arnold, 1989). The method of Cowell (2000, p. 158) is not widely used, in part because it does not lead to closed-form analytical expressions. The methods of Villaseñor and Arnold (1989) and Kakwani and Podder (1976) have been notably used by the World Bank in its PovcalNet database, but are less directly comparable to ours because they do not focus on the top of the distribution, and indeed perform relatively poorly in that part of the distribution. Overall, the generalized Pareto interpolation also compares quite favorably to them, though its primary strength is for the top of the distribution.

Method 1: Constant Pareto coefficient

That method was used by Piketty (2001) and Piketty and Saez (2003), and relies on the property that, for a Pareto distribution, the inverted Pareto coefficient $b(p)$ remains constant. We set $b(p) = b = \mathbb{E}[X|X > q_k]/q_k$ for all $p \geq p_k$. The p-th quantile becomes $q = q_k \left(\frac{1-p}{1-p_k}\right)^{-1/\alpha}$ with $\alpha = b/(b-1)$. By definition, $\mathbb{E}[X|X > q] = bq$, which gives the p-th top average and top share.

Method 2: log-linear interpolation

The log-linear interpolation method was introduced by Pareto (1896), Kuznets (1953), and Feenberg and Poterba (1993). It uses solely threshold information and relies on the property of Pareto distributions that $\log(1-F(x)) = \log(c) - \alpha \log(x)$. We assume that this relation holds exactly within the bracket $[p_k, p_{k+1}]$, and set $\alpha_k = -\frac{\log((1-p_{k+1})/(1-p_k))}{\log(q_{k+1}/q_k)}$. The value of the p-th quantile is again $q = q_k \left(\frac{1-p}{1-p_k}\right)^{-1/\alpha_k}$, and the top averages and top shares can be obtained by integration of the quantile function. For $p > p_K$, we extrapolate using the value α_K of the Pareto coefficient in the last bracket.

Method 3: mean-split histogram

The mean-split histogram uses information on both the means and the thresholds, but uses a very simple functional form, so that the solution can be expressed analytically. Inside the bracket $[q_k, q_{k+1}]$, the density takes two values:

$$f(x) = \begin{cases} f_k^- & \text{if} \quad q_k \leq x < \mu_k \\ f_k^+ & \text{if} \quad \mu_k \leq x < q_{k+1}, \end{cases}$$

where μ_k is the mean inside the bracket. This method is a special case of the split-histogram (Cowell and Mehta, 1982), with the breakpoint parameter inside each bracket set equal to the mean, which is the most common choice in the literature.[6] To meet the requirement on the mean and the thresholds, we set:

$$f_k^- = \frac{(p_{k+1} - p_k)(q_{k+1} - \mu_k)}{(q_{k+1} - q_k)(\mu_k - q_k)} \qquad \text{and} \qquad f_k^+ = \frac{(p_{k+1} - p_k)(\mu_k - q_k)}{(q_{k+1} - q_k)(q_{k+1} - \mu_k)}.$$

The mean-split histogram does not apply beyond the last threshold of the tabulation.

Comparison

Methods 1 and 2 make a fairly inefficient use of the information included in the original tabulation: method 1 discards the data on quantiles and averages at the

[6]That is, as noted by (Cowell and Mehta, 1982), the breakpoint of the interval $[q_k, q_{k+1}]$ could be different from μ_k, but not all values between q_k and q_{k+1} will work if we want to make sure that $f_k^- > 0$ and $f_k^+ > 0$. The breakpoint q^* must be between q_k and $2\mu_k - q_k$ if $\mu_k < (q_k + q_{k+1})/2$, and between $2\mu_k - q_{k+1}$ and q_{k+1} otherwise. Choosing $q^* = \mu_k$ ensures that the condition is always satisfied.

higher end of the bracket, whereas method 2 discards the information on averages. As a consequence, none of these methods can guarantee that the output will be consistent with the input. Method 3 does offer such a guarantee, but with a fairly unrealistic functional form: the density of the resulting distribution is piecewise uniform, exhibiting discontinuities at arbitrary points, as emphasized by Cowell and Mehta (1982).

Our generalized Pareto interpolation method makes use of all the information in the tabulation, so that its output is guaranteed to be consistent with its input. Moreover, contrary to all other methods, it leads a continuous density, hence a smooth quantile and a smooth Pareto curve. None of the other methods can satisfy this requirement, and their output exhibits stark irregularities at the beginning and the end of the brackets in the tabulation in input.

Application to France and the US

Using the individual income tax data, we compute our own tabulations in each year. We include four percentiles in the tabulation: $p_1 = 0.1$, $p_2 = 0.5$, $p_3 = 0.9$, and $p_4 = 0.99$.

We interpolate each of those tabulations with the three methods above, labeled "M1," "M2," and "M3" in what follows. We also interpolate them with our new generalized Pareto interpolation approach (labeled "M0"). We compare the values that we get with each method for the top shares and the quantiles at percentiles 30 percent, 75 percent, and 95 percent with the value that we get directly from the individual data. (We divide all quantiles by the average to get rid of scaling effects because of inflation and average income growth.) We report the mean relative error in Table 1:

$$\text{MRE} = \frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\widehat{y}_t - y_t}{y_t} \right|,$$

where y is the quantity of interest (income threshold or top share), and $\widehat{y}$ is its estimate using one of the interpolation methods.

The two standard Pareto interpolation methods (M1 and M2) are the ones that perform worst. M1 is better at estimating shares, whereas M2 is somewhat better at estimating quantiles. That shows the importance not to dismiss any information included in the tabulation, as exhibited by the good performance of the mean-split histogram (M3), particularly at the bottom of the distribution.

Our generalized Pareto interpolation method vastly outperforms the standard Pareto interpolation methods (M1 and M2). It is also better than the mean-split histogram (M3), except in the bottom of the distribution where both methods work well (but standard Pareto methods M1 and M2 fail badly).

Figure 2 shows how the use of different interpolation methods affects the estimation of the top 25 percent share and associated income threshold. Although all methods roughly respect the overall trend, they can miss the level by a significant margin. The generalized Pareto interpolation estimates the threshold much better than M1, M2, or M3.

TABLE 1
Mean Relative Error for Different Interpolation Methods

		Mean Relative Gap Between Estimated and Observed Values			
		M0	M1	M2	M3
US (1962–2014)	Top 80% share	0.044% (ref.)	0.54% (×12)	7.2% (×164)	0.03% (×0.7)
	Top 70% share	0.059% (ref.)	2.3% (×38)	6.4% (×109)	0.054% (×0.92)
	Top 25% share	0.093% (ref.)	3% (×32)	3.8% (×41)	0.54% (×5.8)
	Top 5% share	0.059% (ref.)	0.84% (×14)	4.4% (×76)	0.83% (×14)
	P20/average	1.4% (ref.)	39% (×28)	25% (×18)	2.1% (×1.5)
	P30/average	0.43% (ref.)	55% (×126)	29% (×67)	1.4% (×3.3)
	P75/average	0.32% (ref.)	11% (×35)	9.9% (×31)	5.8% (×18)
	P95/average	0.3% (ref.)	4.4% (×15)	3.6% (×12)	1.3% (×4.5)
France (1994–2012)	Top 80% share	0.16% (ref.)	0.51% (×3.1)	7.3% (×45)	0.21% (×1.3)
	Top 70% share	0.24% (ref.)	2.4% (×10)	6.5% (×27)	0.21% (×0.88)
	Top 25% share	0.25% (ref.)	1.9% (×7.9)	5.8% (×24)	0.28% (×1.1)
	Top 5% share	0.29% (ref.)	0.68% (×2.3)	11% (×36)	0.28% (×0.95)
	P20/average	4.9% (ref.)	29% (×5.9)	19% (×4)	4.3% (×0.87)
	P30/average	2.4% (ref.)	44% (×19)	25% (×10)	2.4% (×1)
	P75/average	0.83% (ref.)	6.1% (×7.4)	4.6% (×5.6)	4.7% (×5.7)
	P95/average	0.89% (ref.)	4% (×4.5)	1.9% (×2.1)	2.2% (×2.5)

DINA income. *Sources*: author's calculation from Piketty *et al*. (2018) (US) and Garbinti *et al*. (2018) (France). The different interpolation methods are labeled as follows. M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram. We applied them to a tabulation that includes the percentiles $p = 10$ percent, $p = 50$ percent, $p = 90$ percent, and $p = 99$ percent. We included the relative increase in the error compared to generalized Pareto interpolation in parentheses. We report the mean relative error, namely:

$$\frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|,$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods. We calculated the results over the years 1962, 1964, and 1966–2014 in the US and years 1994–2012 in France.

For the estimation of the top 25 percent share, M3 performs fairly well, unlike M1 and M2. To get a more detailed view, we therefore focus on a more recent period (2000–2014) and display only M0 and M3, as in Figure 3. We can see that M3 has, in that case, a tendency to overestimate the top 25 percent by a small yet

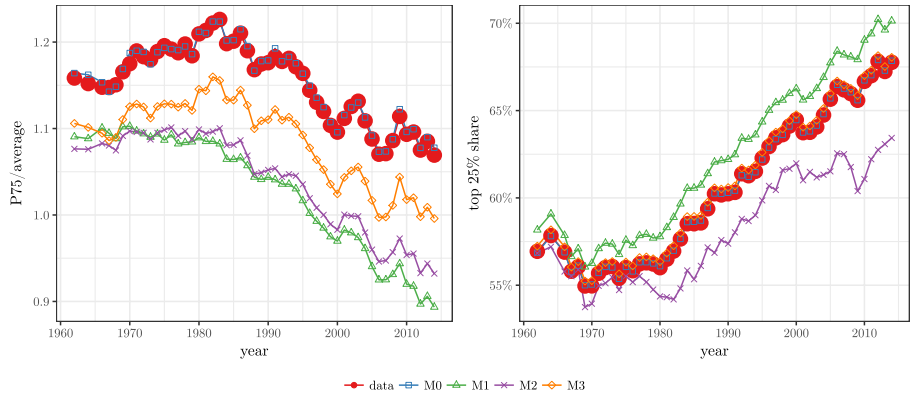

Figure 2. P75 Threshold and Top 25 Percent Share in the US (1962–2014),
Sources: author's computation from Piketty *et al*. (2018). M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram. [Colour figure can be viewed at wileyonlinelibrary.com]
Notes: Estimated Using All Interpolation Methods and a Tabulation with p = 10 percent, 50 percent, 90 percent, and 99 percent DINA Income

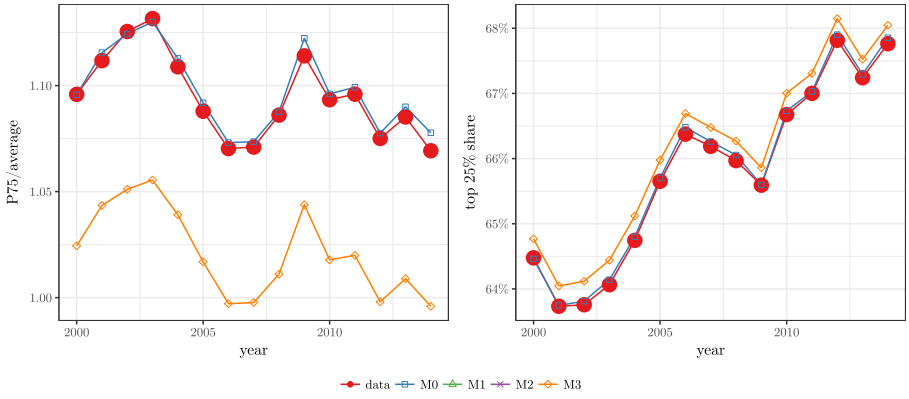

Figure 3. P75 Threshold and Top 25 Percent Share in the US (2000–2014),
Sources: author's computation from Piketty *et al*. (2018). M0: generalized Pareto interpolation. M3: mean-split histogram. [Colour figure can be viewed at wileyonlinelibrary.com]
Notes: Estimated Using Interpolation Methods M0 and M3, and a Tabulation with p = 10 Percent, 50 Percent, 90 Percent, and 99 Percent DINA Income

persistent amount. In comparison, M4 produces a curve almost identical to the real one.

We can also directly compare the generalized Pareto curves generated by each method, as in Figure 4. Our method, M0, reproduces the inverted Pareto coefficients $b(p)$ very faithfully, including above the last threshold (see Section 4.2). All the other methods give much worse results. Method M1 leads to discontinuous curve, which in fact may not even define a consistent probability distribution. The M2 method fails to account for the rise of $b(p)$ at the top. Finally, the M3 leads to an extremely irregular shape because of the use of a piecewise uniform distribution to approximate power law behavior.

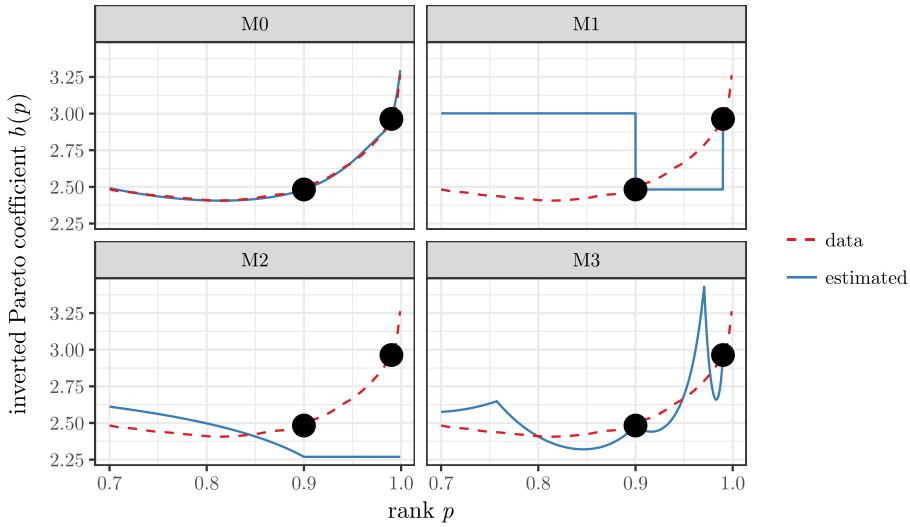

Figure 4. Generalized Pareto Curves Implied by the Different Interpolation Methods for the US Distribution of Income in 2010 DINA Income
Sources: author's computation from Piketty *et al.* (2018). M0: generalized Pareto interpolation. M1: constant Pareto coefficient. M2: log-linear interpolation. M3: mean-split histogram. [Colour figure can be viewed at wileyonlinelibrary.com]

Overall, the generalized Pareto interpolation method performs well. In most cases, it gives results that are several times better than methods commonly used in the literature, and it does so while ensuring a smoothness of the resulting estimate that no other method can provide. Moreover, it works well for the whole distribution, not just the top (like M1 and M2) or the bottom (like M3).

4.2. *Extrapolation methods*

Of the interpolation methods previously described, only M1 and M2 can be used to extrapolate the tabulation beyond the last threshold. Both assume a standard Pareto distribution. Method M1 estimates $b(p)$ at the last fractile p_K, and assumes a Pareto law with $\alpha = b(p_K)/(b(p_K) - 1)$ after that. Method M2 estimates a Pareto coefficient based on the last two thresholds, so in effect it assumes a standard Pareto distribution immediately after the second to last threshold.

The assumption that $b(p)$ becomes approximately constant for p close to 1, however, is not confirmed by the data. Figure 5 shows this for France and the US in 2010. The profile of $b(p)$ is not constant for $p \approx 1$. On the contrary, it increases faster than for the rest of the distribution.

In Section 3.3 we presented an extrapolation method based on the generalized Pareto distribution that had the advantage of preserving the smoothness of the Pareto curve, use all the information from the tabulation, and allow for a nonconstant profile of generalized Pareto coefficients near the top. As Figure 5 shows, this method leads to a more realistic shape of the Pareto curve.

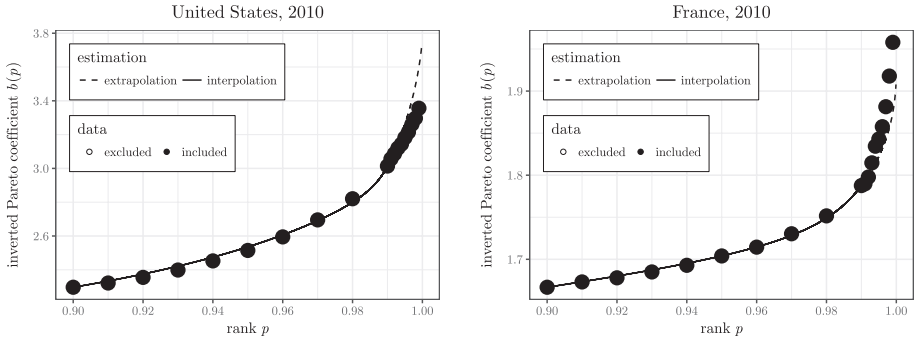

Figure 5. Extrapolation with Generalized Pareto Distribution Fiscal Income

Sources: author's computation from Piketty *et al*. (2018) (for the US) and Garbinti *et al*. (2018) (for France). Included points (hollow dots) come from the data but were not used in the estimation, while included points black dots were.

TABLE 2

Mean Relative Error on the Top 1 Percent for Different Extrapolation Methods, Knowing the Top 10 Percent and the Top 5 Percent

		Mean Relative Gap Between Estimated and Observed Values		
		M0	M1	M2
US (1962–2014)	Top 1% share	0.78% (ref.)	5.2% (×6.7)	40% (×52)
	P99/average	1.8% (ref.)	8.4% (×4.7)	13% (×7.2)
France (1994–2012)	Top 1% share	0.44% (ref.)	2% (×4.6)	11% (×25)
	P99/average	0.98% (ref.)	2.5% (×2.5)	2.4% (×2.4)

Fiscal income.

Sources: author's calculation from Piketty *et al*. (2018) (US) and Garbinti *et al*. (2018) (France). The different extrapolation methods are labeled as follows. M0: generalized Pareto distribution. M1: constant Pareto coefficient. M2: log-linear interpolation. We applied them to a tabulation that includes the percentiles p = 90 percent, and p = 95 percent. We included the relative increase in the error compared to generalized Pareto interpolation in parentheses. We report the mean relative error, namely:

$$\frac{1}{\text{number of years}} \sum_{t=\text{first year}}^{\text{last year}} \left| \frac{\hat{y}_t - y_t}{y_t} \right|,$$

where y is the quantity of interest (income threshold or top share), and $\hat{y}$ is its estimate using one of the interpolation methods. We calculated the results over the years 1962, 1964, and 1966–2014 in the US, and years 1994–2012 in France.

Table 2 compares the performance of the new method with the other ones, as we did in the previous section. Here, the tabulation in input includes p = 90 percent but stops at p = 95 percent, and we seek estimates for p = 99 percent.[7,8] Method M2

[7]Here, we use fiscal income instead of DINA income to avoid disturbances created at the top by the imputation of some sources of income in DINA income.

[8]We provide in appendix an alternative tabulation that stops at the top 1 percent and where we seek the top 0.1 percent. The performances of M0 and M1 are closer but M0 remains preferable.

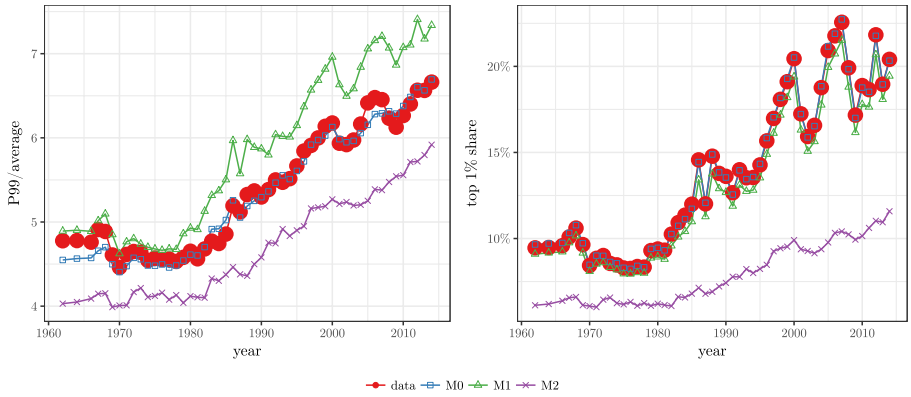

Figure 6. Comparison of Extrapolation Methods in the US for the Top 1 Percent, Knowing the Top 10 Percent and the Top 5 Percent Fiscal income.
Sources: author's computation from Piketty *et al*. (2018). [Colour figure can be viewed at wileyonlinelibrary.com]

is the most imprecise. Method M1 works quite well in comparison. However, our new method M0 gives even more precise results. This is because it can correctly capture the tendency of $b(p)$ to keep on rising at the top of the distribution.

Figure 6 compares the extrapolation methods over time in the US. We can see M1 overestimates the threshold by about as much as M2 underestimates it, whereas M0 is much closer to reality and makes no systematic error. For the top share, M1 is much better than M2. However, it slightly underestimates the top share because it fails to account for the rising profile of inverted Pareto coefficients at the top, which is why our method M0 works even better.

5. Precision

We now discuss a few extensions of the framework presented in this article, which allow us to analyze in greater detail the level of precision one can expect from the different ways of estimating the distribution of top incomes.

5.1. *Estimation of the Error*

When attempting to assess the error term associated with an interpolation method, the main difficulty is that most of the errors are not because of mere sampling variability (although part of it is), which we can assess using standard methods. It comes mostly from the discrepancy between the functional forms used in the interpolation and the true form of the distribution. Put differently, it corresponds to a "model misspecification" error, which is harder to evaluate. However, the generalized Pareto interpolation method does offer some solutions to that problem. We can isolate the features of the distribution that determine the error, and based on that provide approximations of it.

In this section, we remain concerned with the same definition of the error as in the previous one. Namely, we consider the difference between the estimate of a quantity by interpolation (e.g., shares or thresholds) and the same quantity defined

over the true population of interest. This is in contrast with a different notion of error common in statistics: the difference between an empirical estimate and the value of an underlying statistical model. If sample size was infinite—so that sampling variability would vanish—both errors would be identical. However, despite the large samples that characterize tax data, sampling issues cannot be entirely discarded. Indeed, because income and wealth distributions are fat-tailed, the law of large numbers may operate very slowly, so that both types of errors remain different even with millions of observations (Taleb and Douady, 2015).

We consider our notion of the error to be more appropriate in the context of the methods we are studying. Indeed, concerns for the distribution of income and wealth only arise to the extent that it affects the actual population, not a model of it. Moreover, this allows us to remain agnostic as to the "true" model for the distribution of income.

To get tractable analytical results, we also focus on the unconstrained interpolation procedure of Section 3.1, and thus leave aside the monotonicity constraint of the quantile. That has very little impact on the results in practice since the constraint is rarely binding, and when it is, the adjustments are small. For example, the monotonicity constraint is not binding in any of the tabulations interpolated in the previous section.

Let n be the size of the population (from which the tabulated data come). Recall that $x = -\log(1-p)$. Let $e_n(x)$ be the estimation error on $\varphi_n(x)$, and similarly $e'_n(x)$ the estimation error on $\varphi'_n(x)$. If we know both those errors, then we can retrieve the error on any quantity of interest (quantiles, top shares, Pareto coefficients, etc.) by applying the appropriate transforms. Our first result decomposes the error between two components. Like all the theorems of this section, we give only the main results. Details and proofs are in Appendix E.

Theorem 3 We can write $e_n(x) = u(x) + v_n(x)$ and $e'_n(x) = u'(x) + v'_n(x)$ where $u(x), u'(x)$ are deterministic, and $v_n(x), v'_n(x)$ are random variables that converge almost surely to zero when $n \to +\infty$.

We call the first terms $u(x)$ and $u'(x)$ the "misspecification" error. They correspond to the difference between the functional forms that we use in the interpolation, and the true functional forms of the underlying distribution. Even if the population size was infinite, so that sampling variability was absent, they would still remain nonzero. We can give the following representation for that error.

Theorem 4 $u(x)$ and $u'(x)$ can be written as a scalar product between two functions ε and φ''':

$$u(x) = \int_{x_1}^{x_K} \varepsilon(x,t)\varphi'''(t)\,\mathrm{d}t \qquad \text{and} \qquad u'(x) = \int_{x_1}^{x_K} \frac{\partial \varepsilon}{\partial x}(x,t)\varphi'''(t)\,\mathrm{d}t,$$

where $\varepsilon(x,t)$ is entirely determined by $x_1, \ldots, x_K$.

The function $\varepsilon(x,t)$ is entirely determined by the known values $x_1, \ldots, x_K$, so we can calculate it directly. Its precise definition is given in appendix. The other function, φ''', depends on the quantity we are trying to estimate, so we do not know it exactly. The issue is common in nonparametric statistics and complicates the

application of the formula.[9] However, if we look at the value of φ''' in situations where we have enough data to estimate it directly, we can still derive good approximations and rules of thumb that apply more generally.

We call $v_n(x)$ and $v_n'(x)$ the "sampling error." Even if the true underlying distribution matched the functional used for the interpolation, so that there would be no misspecification error, they would remain nonzero. We can give asymptotic approximation of their distribution for large n. We do not only cover the finite variance case ($\mathbb{E}[X^2] < +\infty$), but also cover the infinite variance case ($\mathbb{E}[X^2] = +\infty$), which leads to results that are less standard. Infinite variance is very common when dealing with distributions of income and wealth.

Theorem 5 $v_n(x)$ and $v_n'(x)$ converge jointly in distribution at speed $1/r_n$:

$$r_n \begin{bmatrix} v_n(x) \\ v_n'(x) \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{J}.$$

If $\mathbb{E}[X^2] < +\infty$, then $r_n = \sqrt{n}$ and $\mathcal{J}$ is a bivariate normal distribution. If $\mathbb{E}[X^2] = +\infty$ and $1 - F(x) \sim Cx^{-2}$, then $r_n = (n/\log n)^{1/2}$ and $\mathcal{J}$ is a bivariate normal distribution. If $\mathbb{E}[X^2] = +\infty$ and $1 - F(x) \sim Cx^{-\alpha}$ $(1 < \alpha < 2)$, then $r_n = n^{1-1/\alpha}$ and $\mathcal{J} \overset{\mathcal{D}}{=} (\gamma_1 Y, \gamma_2 Y)$, where Y follows a maximally skewed stable distribution with stability parameter α.

Again, we provide more detailed expressions of the asymptotic distributions in Appendix E alongside the proof of the result. More importantly, we also show that in practice, we always have $v_n(x) \ll u(x)$ and $v_n'(x) \ll u'(x)$, regardless of the precise characteristics of the underlying distribution. This means that sampling variability is negligible compared to the misspecification error. Therefore, we will from now on assume that $e_n(x) \approx u(x)$ and $e_n'(x) \approx u'(x)$.

5.2. *Optimal Choice of Brackets*

How many brackets do we need to achieve a given precision level, and how should they be placed? Based on Theorem 4, we can answer that question for any given φ''' by solving an optimization program. Therefore, if we pick a functional form for φ''' which is typical of what we observe, we get the solution of the problem for the typical income distribution.

We assume that we want our tabulation to span from the 10 percent to the 99.9 percent percentiles, so we set $p_1 = 0.1$ and $p_K = 0.999$. We pick the median profile of φ''' estimated over all available years for France and the US (see Figure 7 in appendix). For a given number K of thresholds, and using the derivative-free Nelder-Mead algorithm, we solve the optimization problem:

$$\min_{p_2,\dots,p_{K-1}} \left\{ \max_{t \in [x_1, x_K]} \int_{x_1}^{x_K} \varepsilon(x,t)\varphi'''(t)\,\mathrm{d}t \right\} \quad \text{st.} \quad p_1 < p_2 < \cdots < p_{K-1} < p_K,$$

[9]For example, the asymptotic mean integrated squared error of a kernel estimator depends on the second derivative of the density (Scott, 1992, p. 131).

TABLE 3
Optimal Bracket Choice for a Typical Distribution of Income

	3 Brackets	4 Brackets	5 Brackets	6 Brackets	7 Brackets
Optimal placement of thresholds	10.0%	10.0%	10.0%	10.0%	10.0%
	68.7%	53.4%	43.0%	36.8%	32.6%
	95.2%	83.4%	70.4%	60.7%	53.3%
	99.9%	97.1%	89.3%	80.2%	71.8%
		99.9%	98.0%	93.1%	86.2%
			99.9%	98.6%	95.4%
				99.9%	98.9%
					99.9%
Maximum relative error on top shares	0.91%	0.32%	0.14%	0.08%	0.05%

where as usual $x_k = -\log(1 - p_k)$ for $1 \leq k \leq K$.

Table 3 shows that an important concentration of brackets near the top is desirable, but that we also need quite a few to cover the bottom. Half of the brackets should cover the top 20 percent, most of which should be within just the top 10 percent. The rest should be used to cover the bottom 80 percent of the distribution. We can also see that a relatively small number of well-placed brackets can achieve remarkable precision: only six are necessary to achieve a maximal relative error of less than 0.1 percent.

Davies and Shorrocks (1989) studied a similar question and we can compare our results to theirs. Unlike this article, they focus on the estimation of a specific inequality indicator (the Gini coefficient) directly from grouped data, without an interpolation step. Our approach interpolates the grouped data and then seeks to minimize the maximum error on top over the whole distribution. Yet both sets of result provide similar recommendations: grouped data can achieve great accuracy in measuring inequality, and the optimal grouping somewhat concentrates groups at the top of the distribution.

5.3. *Comparison with Partial Subsamples*

We have seen that generalized Pareto interpolation can be quite precise, but how does it compare to the use of a subsample of individual data? The question may be of practical interest when researchers have access to both exhaustive data in tabulated form and a partial sample of individual data. Such a sample could either be a survey or a subsample of administrative data.

We may address that question using an example and Monte-Carlo simulations. Take the 2010 distribution of DINA income in the US. We can estimate that distribution and use it to simulate a sample of size $N = 10^8$ (the same order of magnitude as the population of the US).

Then, we create subsamples of size $n \leq N$ by drawing without replacement from the large population previously generated.[10] In the case of surveys, we ignore nonresponse and no misreporting, a simplification that favors the survey in the comparison. For each of those subsamples, we estimate the quantiles and top shares at different points of the distribution, and compare it to the same values in the original sample of size N. Table 4 shows the results for different values of n. We see that even for large

[10]This survey design is called simple random sampling.

TABLE 4
MEAN RELATIVE ERROR USING SUBSAMPLES OF THE FULL POPULATION

	Mean Percentage Gap Between Estimated and Observed Values for a Survey with Simple Random Sampling and Sample Size n					
	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$	$n = 10^7$	$n = 10^8$
Top 70% share	0.42	0.20	0.10	0.04	0.01	0.00
Top 50% share	1.26	0.63	0.32	0.13	0.04	0.00
Top 25% share	4.00	2.04	1.05	0.44	0.15	0.00
Top 10% share	9.29	4.80	2.50	1.05	0.35	0.00
Top 5% share	14.32	7.48	3.94	1.65	0.55	0.00
Top 1% share	29.13	16.01	8.57	3.61	1.21	0.00
Top 0.1% share	52.94	35.23	19.91	8.57	2.89	0.00
P30 threshold	4.67	1.44	0.45	0.15	0.04	0.00
P50 threshold	3.29	1.03	0.33	0.10	0.03	0.00
P75 threshold	2.92	0.91	0.31	0.10	0.03	0.00
P90 threshold	3.91	1.21	0.39	0.12	0.04	0.00
P95 threshold	5.86	1.76	0.59	0.18	0.06	0.00
P99 threshold	14.39	4.79	1.42	0.46	0.14	0.00
P99.9 threshold	44.31	16.29	5.47	1.70	0.49	0.00

Original sample of size $N = 10^8$ simulated using the distribution of 2010 DINA income in the US. Source: author's computations from Piketty *et al*. (2018).

samples ($n = 10^5$, $n = 10^6$, $n = 10^7$), the case for using tabulations of exhaustive data rather than subsamples to estimates quantities such as the top 1 percent or 0.1 percent share remains strong. Indeed, even with $n = 10^6$ observations, the typical error on the top 1 percent share is larger than what we get in Table 3, even with few thresholds. In practice, the thresholds may not be positioned in an optimal way as in Table 3, so may also want to compare the results with Table 1. The differences in the orders of magnitude are large enough, so that the implications of that comparison hold.

6. CONCLUDING COMMENTS

In this article, we introduce the concept of generalized Pareto curve to characterize, visualize, and estimate distributions of income or wealth. Based on quasi-exhaustive individual tax data, we reveal some stylized facts about the distribution of income that lets us move beyond the standard Pareto assumption. We find that although generalized Pareto curves can vary a lot over time and between countries, they tend to stay U-shaped.

Then we develop a method to interpolate tabulated data on income—as is typically available from tax authorities and statistical institutes—that can correctly reproduce the subtleties of generalized Pareto curves. In particular, the method guarantees the smoothness of the estimated distribution. It works especially well for the top half of the distribution. We show that method to be several times more precise than the alternatives most commonly used in the literature. In fact, it can often be more precise than using non-exhaustive individual data. Moreover, we can derive formulas for the error term that let us approximately bound the error of our estimates, and determine the number of optimally placed brackets that is necessary to achieve a given precision. The method could also be applied to wealth in cases where data under a similar form are available, which is much rarer.

We believe that more empirical work—especially a careful use of administrative data sources—is necessary to study those dynamics in a fully satisfying way. We hope that the interpolation method presented in this article will allow future researchers make progress in that direction. To that end, we made the methods presented in this article available as a R package named `gpinter`, and also in the form of an online interface that can be used without any installation or knowledge of any programming language. Both are available at http://wid.world/gpinter.

References

Alvaredo, F., L. Assouad, and T. Piketty, "Measuring Inequality in the Middle East 1990–2016: The World's Most Unequal Region?," *Review of Income and Wealth*, 65(4), 685–711, 2019.

Alvaredo, F., A. B. Atkinson, et al., *Distributional National Accounts Guidelines Methods and Concepts Used in the World Inequality Database*. Available at https://wid.world/document/distributional-national-accounts-guidelines-2020-concepts-and-methods-used-in-the-world-inequality-database/. 2020.

Atkinson, A. B., "Measuring Top Incomes: Methodological Issues," *Top Incomes over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries*, Oxford University Press, Oxford, 2007.

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

Atkinson, A. B., and A. J. Harrison, *Distribution of Personal Wealth in Britain*, Cambridge University Press, Cambridge, 1978.

Atkinson, A. B., and T. Piketty, *Top Incomes Over the Twentieth Century: a Contrast Between Continental European and English-Speaking Countries*, Oxford University Press, 2007.

Balkema, A. A., and L. de Haan, "Residual Life Time at Great Age," *Annals of Probability*, 2(5), 792–804, 1974.

Benhabib, J., and A. Bisin, "Skewed Wealth Distributions: Theory and Empirics," *NBER Working Paper Series*, 21924, 37, 2016.

Benhabib, J., A. Bisin, and S. Zhu, "The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents," *Econometrica*, 79(1), 123–157, 2011.

Bierbrauer, F. J., P. C. Boyer, and A. Peichl, "Politically Feasible Reforms of Nonlinear Tax Systems," *American Economic Review*, 111(1), 153–191, 2021.

Birgin, E. G., and J. M. Martìnez, "Improving Ultimate Convergence of an Augmented Lagrangian Method," *Optimization Methods Software*, 23(2), 177–195, 2008.

Bukowski, P., and F. Novokmet, "Inequality in Poland: Estimating the Whole Distribution by g-Percentile, 1983–2015," *WID.world Working Paper Series 2017/21*. Available at http://wid.world/wp-content/uploads/2017/11/Bukowski_Novokmet_WP_WIDworld_2017_21.pdf, 2017.

Cargo, G. T., and O. Shisha, "The Bernstein Form of a Polynomial," *Journal of Research of the National Bureau of Standards*, 60(B.1), 79–81, 1966.

Champernowne, D. G., "A Model of Income Distribution," *The Economic Journal*, 63(250), 318–351, 1953.

Chancel, L., and T. Piketty, "Indian Income Inequality, 1922–2015: From British Raj to Billionaire Raj?," *Review of Income and Wealth*, 65(S1), S33–S62, 2019.

Chang, W. et al., Shiny: Web Application Framework for R. R Package Version 1.0.3, 2017.

Charpentier, A., and E. Flachaire, Pareto Models for Top Incomes. Available at https://sites.google.com/site/emmanuelflachaire/publications, 2019.

Chotikapanich, D. et al., "Global Income Distributions and Inequality, 1993 and 2000: Incorporating Country-level Inequality Modeled with Beta Distributions," *Review of Economics and Statistics*, 94(1), 52–73, 2012.

Conn, A. R., N. I. M. Gould, and P. L. Toint, "A Globally Convergent Augmented Langrangian Algorithm for Optimization with General Constraints and Simple Bounds," *SIAM Journal on Numerical Analysis*, 28(2), 545–572, 1991.

Cowell, F. A., *Measuring Inequality. LSE Economic Series*, Oxford University Press, Oxford, 2000.

Cowell, F. A., and F. Mehta, "The Estimation and Interpolation of Inequality Measures," *Review of Economic Studies*, 49(2), 273–290, 1982.

Czajka, L., "Income Inequality in Côte d'Ivoire: 1985–2014." WID.world Working Paper 2017/8. Available at https://wid.world/document/income-inequality-cote-divoire-1985-2014-wid-world-working-paper-201708/, 2017.

Davies, J. B., and A. F. Shorrocks, "Optimal Grouping of Income and Wealth Data," *Journal of Econometrics*, 42(1), 97–108, 1989.
Feenberg, D. R., J. M. Poterba, "Income inequality and the incomes of very high-income taxpayers: Evidence from tax returns," *Tax Policy and the Economy*, 7, 145–177, 1993.
Fournier, J., "Generalized Pareto Curves: Theory and Application Using Income and Inheritance Tabulations for France 1901–2012." MA Thesis. Paris School of Economics, 2015.
Gabaix, X. et al., "The Dynamics of Inequality," *Econometrica*, 84(6), 2071–2111, 2016.
Garbinti, B., J. Goupille-Lebret, and T. Piketty, "Income Inequality in France, 1900–2014: Evidence from Distributional National Accounts (DINA)," *Journal of Public Economics*, 162(June), 63–77, 2018.
Jargowsky, P. A. and C. A. Wheeler, "Estimating Income Statistics from Grouped Data: Mean-Constrained Integration over Brackets," *Sociological Methodology*, 48(1), 337–374, 2018.
Jenkins, S. P., "Pareto Models, Top Incomes and Recent Trends in UK Income Inequality," *Economica*, 84(334), 261–289, 2017.
Jones, C. I., "Pareto and Piketty: The Macroeconomics of Top Income and Wealth Inequality," *Journal of Economic Perspectives*, 29(1), 29–46, 2015.
Jones, C. I., and J. Kim, "A Schumpeterian Model of Top Income Inequality," *Journal of Political Economy*, 126(5), 1785–1826, 2018.
Kakwani, N. C., and N. Podder, "Efficient Estimation of the Lorenz Curve and Associated Inequality Measures from Grouped Observations," *Econometrica*, 44(3), 630, 1976.
Kraft, D., "Algorithm 733: TOMP–Fortran Modules for Optimal Control Calculations," *ACM Transactions on Mathematical Software*, 20(3), 262–281, 1994.
Kuznets, S., Shares of Upper Income Groups in Income and Savings. National Bureau of Economic Research, Cambridge MA, 1953.
Lyche, T., and K. M⊘rken, "Spline Methods," Available at https://www.uio.no/studier/emner/matnat/math/MAT4170/v18/pensumliste/splinebook-2018.pdf, 2002.
Morgan, M., "Extreme and Persistent Inequality: New Evidence for Brazil Combining National Accounts, Surveys and Fiscal Data, 2001–2015," WID.world Working Paper Series 2017/12. Available at https://wid.world/document/extreme-persistent-inequality-new-evidence-brazil-combining-national-accounts-surveys-fiscal-data-2001-2015-wid-world-working-paper-201712/, 2017.
Nirei, M., "Pareto Distributions in Economic Growth Models IIR Working Paper 09-05. Available at http://hermes-ir.lib.hit-u.ac.jp/hermes/ir/re/17503/070iirWP09_05.pdf, 2009.
Novokmet, F., T. Piketty, and G. Zucman, "From Soviets to Oligarchs: Inequality and Property in Russia 1905–2016," *Journal of Economic Inequality*, 16(2), 189–223, 2018.
Pareto, V., Cours d'économie Politique. Available at https://www.cairn.info/cours-d-economie-politique-tomes-1-et-2--9782600040143.htm, 1896.
Pickands, J., "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics*, 3(1), 119–131, 1975.
Piketty, T., "Income Inequality in France, 1901–1998," *Journal of Political Economy*, 111(5), 1004–1042, 2003.
_____, *Les hauts revenus en France au XXème siècle*, Grasset, Paris, 2001.
Piketty, T., and E. Saez, "Income Inequality in the United States, 1913–1998," *Quarterly Journal of Economics*, 118(1), 1–39, 2003.
Piketty, T., E. Saez, and G. Zucman, "Distributional National Accounts: Methods and Estimates for the United States," *Quarterly Journal of Economics*, 133(5), 553–609, 2018.
Piketty, T., L. Yang, and G. Zucman, "Capital Accumulation, Private Property, and Rising Inequality in China, 1978–2015," *American Economic Review*, 109(7), 2469–2496, 2019.
Piketty, T., and G. Zucman, "Wealth and Inheritance in the Long Run," *Handbook of Income Distribution*, Vol. 2. Handbook of Income Distribution, Elsevier, Amsterdam, 1303–1368, 2015.
R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria, 2016.
Saez, E., "Using Elasticities to Derive Optimal Income Tax Rates," *Review of Economic Studies*, 68(1), 205–229, 2001.
Scott, D. W., *Multivariate Density Estimation*, John Wiley & Sons, Inc., New York City, 1992.
Simon, H., "On a Class of Skew Distribution Functions," *Biometrika*, 42(3–4), 425–440, 1955.
Taleb, N. N., and R. Douady, "On the Super-additivity and Estimation Biases of Quantile Contributions," *Physica A Statistical Mechanics and its Applications*, 429, 252–260, 2015.
van der Wijk, J., *Inkomens- En Vermogensverdeling*. Nederlands economisch instituut, De Erven F. Bohn, Haarlem, 1939.
Villaseñor, J. A., and B. C. Arnold, "Elliptical Lorenz Curves," *Journal of Econometrics*, 40(2), 327–338, 1989.
Wold, H. O. A., and P. Whittle, "A Model Explaining the Pareto Distribution of Wealth," *Econometrica*, 25(4), 591–595, 1957.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site: