Estimating Income Distributions From Grouped Data: A Minimum Quantile Distance Approach

Spasova, Tsvetana

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Spasova, T. (2023). Estimating Income Distributions From Grouped Data: A Minimum Quantile Distance Approach. *Computational Economics*, Early View, 1-18. https://doi.org/10.1007/s10614-023-10505-0

Computational Economics
https://doi.org/10.1007/s10614-023-10505-0

Estimating Income Distributions From Grouped Data: A Minimum Quantile Distance Approach

Tsvetana Spasova[1]

Accepted: 17 October 2023

Abstract

This paper focuses on the estimation of income distribution from grouped data in the form of quantiles. We propose a novel application of the minimum quantile distance (MQD) approach and compare its performance with the maximum likelihood (ML) technique. The estimation methods are applied using three parametric distributions: the generalized beta distribution of the second kind (GB2), the Dagum distribution, and the Singh–Maddala distribution. We provide the density-quantile functions for these distributions, along with reproducible `R` code. A simulation study is conducted to evaluate the performance of the MQD and ML methods. The proposed methods are then applied to data from 30 European countries, utilizing the aforementioned parametric distributions. To validate the accuracy of the estimates, we compare them with estimates obtained from more detailed and informative microdata sets. The findings confirm the excellent performance of the considered parametric distributions in estimating income distribution. Additionally, the MQD approach is identified as a straightforward and reliable method for this purpose. Notably, the MQD method displays superior robustness in comparison to the ML technique when it comes to selecting suitable starting values for the underlying computation algorithm, specifically when dealing with the GB2 distribution.

Keywords Minimum quantile distance · Maximum likelihood technique · Income distributions · Grouped data · GB2 distribution

JEL Classification C13 · C46 · D31

✉ Tsvetana Spasova
tsvetana.spasova@fhnw.ch

[1] School of Business, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Riggenbachstrasse 16, 4600 Olten, Switzerland

Published online: 15 November 2023

1 Introduction

Estimating income distribution in an accurate way is very important for the measurement of inequality and poverty, and more generally comparing welfare across space and time. An overview of the literature on modeling income distributions, various estimation methods and distribution specifications is available in the book by Kleiber and Kotz (2003), and the papers in Chotikapanich (2008).

When individual income data are available the estimation of income would be quite straightforward. However, very often the available income data is scarce, especially for many developing countries, which encumbers deriving representative income distribution models and inequality statistics. Frequently the income data are only available in grouped form, for example income deciles or income shares, mean incomes and Gini coefficients. The World Income Inequality Database (WIID), the World Inequality Database (WID) and the World Bank are among the largest databases providing grouped income data. However, when looking into smaller areas, the data provided can be only in the form of income quantiles due to privacy of personal data and the proximity of the considered areas as, for example, household income data at local levels provided by the French National Institute of Statistics and Economic Studies (INSEE). This paper focuses on estimating income distribution using only quantile income data and aims at determining a method suitable for such data.

In terms of modeling grouped income data, various approaches have been used depending on the data available. Two main strategies have been developed, either nonparametric techniques like for instance employing a nonparametric kernel density function (Sala-i-Martin, 2006), or parametric techniques assuming that the income distribution follows a parametric model. Parametric models are shown to perform very well when estimating income distributions and inequality measures (Chotikapanich et al., 2007) and even outperform the nonparametric techniques (Minoiu & Reddy, 2014; Jordá et al., 2021).

For the parametric modeling, it is crucial to choose a reliable estimation technique and a suitable parametric distribution model. Besides, the estimation techniques have to be adjusted to the grouped data types, usually grouped data with fixed bounds and random cell size or grouped data with fixed cell size and random bounds. Among the most common estimation techniques is the maximum likelihood based on sample proportions using a multinomial likelihood function [see, for example, (McDonald, 1984; Jöhnk & Niermann, 2002; Bandourian et al., 2003; Chotikapanich et al., 2018)]. Eckernkemper and Gribisch (2021) propose a general framework for ML and Bayesian estimation based on grouped data information accounting for known and unknown group boundaries. Another widely used technique is the method-of-moments approach where population and income shares are matched to their theoretical counterparts. Chotikapanich et al. (2007, 2012) apply it for the beta-2 distribution using population shares and class mean income data. Hajargasht et al. (2012) extended the work of Chotikapanich et al. (2007, 2012) to a generalized method-of-moments (GMM) approach and provided inference for the estimated distributions. Further, minimizing the distance

between a set of income indicators and their parametric representations is suggested by Graf and Nedyalkova (2014) and Hajargasht and Griffiths (2020) suggest minimum distance estimation of parametric Lorenz curves based on grouped data information.

In this work, we suggest the minumum quantile distance (MQD) method which is designed especially for quantile data (grouped data with fixed bounds) which as mentioned above could be the only grouped data available (for example, data from INSEE). Assuming that the income distribution of a country can be modeled with a specific parametric distribution, in this work we estimate the income distribution of each observed country by minimizing the distance between the empirical estimates of the respective country's income quantiles and their parametric representations. We compare our estimates with the estimates obtained with a ML method. At the end, we verify the results by comparing them with representative microdata.

Some of the earliest research work introducing the minimum quantile distance approach was done by Aitchison and Brown (1957) who applied the method to the log-normal distribution. After Parzen (1979) introduced the density-quantile function, LaRiccia and Wehrly (1985) showed the asymptotic properties of a family of minimum quantile distance estimators and applied it to the three-parameter log-normal distribution. Carmody et al. (1984) applied it to the three-parameter Weibull distribution. Jöhnk and Niermann (2002) compare it with other methods employing the Weibull distribution.

In the present study, we contribute to the literature by examining the performance of the MQD method applied to the generalized beta distribution of the second type (GB2), which is the mostly used distribution in recent studies on income distribution (Chotikapanich et al., 2018), the Dagum (1977) and the Singh-Maddala distributions. We provide the density-quantile functions for the considered distributions and reproducible `R` code (R Core Team, 2022). Further, we compare the MQD method with the ML. We estimate the income distribution of 30 European countries using data on their income deciles and quintiles. We use data from Eurostat, namely the European Union Statistics on Income and Living Conditions (EU-SILC 2011) data. Due to the fact that we have microdata for all of the observed countries, we have the opportunity to compare the accuracy of our estimates from the grouped data with the more representative microdata estimates. The findings of our study reveal that the MQD method performs as good as the ML method for both decile and quintile data. However, the MQD method exhibits greater robustness and lower sensitivity to starting values, as supported by a simulation study we conducted. The Dagum and the Singh–Maddala distributions are outperformed by the GB2 in terms of absolute differences between the estimated parametric quantiles and their observed nonparametric counterparts. We note that the GB2 outperformance is sometimes at the cost of introducing significant empirical and analytical complexity [see also (Bandourian et al., 2003)]. The Gini coefficient and the mean are best approximated with the Dagum distribution irrelevant of the estimation technique, when evaluating the estimates based on absolute error (difference between parametric estimates and estimates from the microdata).

This work is structured as follows. In Sect. 2.1, the MQD method is described. Section 2.2 outlines briefly the ML technique. In Sect. 2.3, the GB2, the Dagum and the

Singh–Maddala distributions are defined. Simulation results are shown in Sect. 3. The data being used and the empirical results are discussed in Sect. 4. Finally, we summarize and make some concluding remarks in Sect. 5.

2 Methodology

Let N be the number of income quantiles available for a given country and let $\mathbf{q} = (q(u_1), \cdots, q(u_N))^\top$ be a N-vector of sample quantiles with $q(u)$ denoting the uth quantile and $0 < u_1 < \cdots < u_N < 1$.

2.1 The Minimum Quantile Distance Method

Assuming that given data comes from a specific parametric distribution, one can represent the observed income quantiles parametrically with the quantile function of the assumed distribution. Then the representative parametric distribution can be estimated by minimizing the distance between the observed income quantiles and their parametric counterparts. This method was applied and proved to be consistent, asymptotically normal and robust against gross errors under the regularity conditions specified by LaRiccia and Wehrly (1985).

Let $\mathbf{Q}(\theta) = (Q(u_i;\theta))_{i=1}^N$ be a N-vector of theoretical quantiles of a given parametric distribution and θ the vector of the parameters of the considered distribution. Following LaRiccia and Wehrly (1985), the minimum quantile distance estimator is given by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \{\mathbf{q} - \mathbf{Q}(\theta)\}^\top \mathbf{H}(\theta) \{\mathbf{q} - \mathbf{Q}(\theta)\}, \tag{1}$$

where is $\mathbf{q}$ a N-vector of sample quantiles as defined above.

$\mathbf{H}(\theta)$ is the optimal weighting matrix defined as

$$\mathbf{H}(\theta) = \mathbf{D}(\theta)\mathbf{V}^{-1}\mathbf{D}(\theta), \tag{2}$$

which is the inverse of the asymptotic covariance matrix of $\sqrt{N}(\mathbf{q} - \mathbf{Q}(\theta))$ and $\mathbf{V}^{-1}$ is the inverse of the matrix $\mathbf{V}$ defined as

$$\mathbf{V} = \{\min(u_i, u_j) - u_i u_j\}_{N \times N} \tag{3}$$

and

$$\mathbf{D}(\theta) = \operatorname{diag}[fQ(u_1;\theta), \cdots, fQ(u_N;\theta)], \tag{4}$$

with $fQ(u;\theta) = f[Q(u;\theta);\theta]$ being the density-quantile function defined in LaRiccia and Wehrly (1985) and Parzen (1979).

2.2 Maximum Likelihood Estimation

Let the cumulative number of the observed income group observations be $s_i = \sum_{j=1}^{i} s_j$ with $i = 1, ..., N$ and $s = s_{N+1}$ be the total number of group observations.

Having the information on the income quantiles and the corresponding number of observations for each income group, we could use the maximum likelihood estimation technique. Following Eckernkemper and Gribisch (2021, Equations (4)–(6)) and Nishino and Kakamu (2011), we obtain the likelihood from a joint distribution of order statistics

$$\begin{aligned} L(\theta) =& s! \frac{(F(q(u_i);\theta))^{s_i-1}}{(s_i-1)!} f(q(u_1);\theta) \\ & \times \left\{ \prod_{i=2}^{N} \frac{(F(q(u_i);\theta) - F(q(u_{i-1});\theta))^{s_i - s_{i-1}-1}}{(s_i - s_{i-1} - 1)!} f(q(u_i);\theta) \right\} \\ & \times \frac{(1 - F(q(u_N);\theta))^{s-s_N}}{(s-s_N)!} \end{aligned} \tag{5}$$

Taking logarithms of Eq. 5, we obtain the log-likelihood

$$\begin{aligned} \log L(\theta) = C &+ \sum_{i=1}^{N} \log(f(q(u_i);\theta)) + (s_1 - 1)\log(F(q(u_1);\theta)) \\ &+ \sum_{i=2}^{N} (s_i - s_{i-1} - 1)\log[F(q(u_i);\theta) - F(q(u_{i-1});\theta)] \\ &+ (s - s_N)\log(1 - F(q(u_N);\theta)), \end{aligned} \tag{6}$$

where F is a cumulative distribution function of the considered parametric distribution, f the respective density function, θ the vector of the parameters of the considered distribution and $q(u_i)$ is the u_ith sample quantile as defined above.

2.3 The GB2 Distribution

The GB2 was introduced by McDonald (1984) and is acknowledged to perform in an excellent way when estimating income distributions [see (Kleiber & Kotz, 2003; Jenkins, 2009; Chotikapanich et al., 2018)]. It is a four-parameter distribution, and we will denote it as $GB2(\theta)$, where θ is the quadruple (a, b, p, q). Its density is

$$f(x;\theta) = \frac{a x^{ap-1}}{b^{ap} B(p,q)[1 + (x/b)^a]^{p+q}}, \quad x > 0, \tag{7}$$

where a, b, p and q are positive and $B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt$ is the beta function. When θ is obvious in the context, we write only $f(x)$.

The cumulative distribution function (cdf) is given by

$$F(x;\theta) = B\left(\frac{(x/b)^a}{1+(x/b)^a};p;q\right), \quad x > 0, \tag{8}$$

where $B(\nu;p,q) = \int_0^\nu t^{p-1}(1-t)^{q-1}dt/B(p,q)$ is the incomplete beta function ratio with $\nu = \frac{(x/b)^a}{1+(x/b)^a}$. $B(\nu;p,q)$ is commonly included as readily-computed function in statistical software.

The quantile function is given by Chotikapanich et al. (2018)

$$Q(u;\theta) = b\left(\frac{B^{-1}(u;p,q)}{1-B^{-1}(u;p,q)}\right)^{1/a}, \quad 0 < u < 1, \tag{9}$$

where $B^{-1}(u;p,q)$ is the quantile function of the standardized beta distribution evaluated at u.

The density-quantile function is a basic object in quantile-based methodology. It is obtained by substituting the density function (Eq. 7) into the quantile function (equation 9). For the GB2 distribution the density-quantile function is given by

$$fQ(u;\theta) = \frac{a\left(\frac{B^{-1}(u;p,q)}{1-B^{-1}(u;p,q)}\right)^{(ap-1)/a}}{bB(p,q)\left(1+\frac{B^{-1}(u;p,q)}{1-B^{-1}(u;p,q)}\right)^{p+q}} \tag{10}$$

where $B^{-1}(u;p,q)$ is the quantile function of the standardized beta distribution evaluated at u and $B(p,q)$ is the beta function.

The moment distribution function for the kth moment is given by

$$F_k(x;\theta) = B\left(\frac{(x/b)^a}{1+(x/b)^a};p+k/a,q-k/a\right), \tag{11}$$

where $B(\nu;p+k/a,q-k/a)$ is the incomplete beta function ratio defined as above with $\nu = \frac{(x/b)^a}{1+(x/b)^a}$.

The $k-$th moment is given by

$$\mu^{(k)} = \frac{b^k B(p+k/a)B(q-k/a)}{B(p,q)}.$$

The Gini coefficient was provided by McDonald (1984) and is given by

$$G = \frac{B(2q-1/a,2p+1/a)}{B(p,q)B(p+1/a,q-1/a)}\left(\frac{1}{p}J^{(1)} - \frac{1}{p+1/a}J^{(2)}\right) \tag{12}$$

where

Table 1 Singh-Maddala and Dagum distributions characteristics

Function	Dagum	Singh-Maddala
Density	$\frac{apx^{ap-1}}{b^{ap}[1+(x/b)^a]^{1+p}}$	$\frac{aqx^{a-1}}{b^a[1+(x/b)^a]^{1+q}}$
CDF	$\left[1+\left(\frac{x}{b}\right)^{-a}\right]^{-p}$	$1-\left[1+\left(\frac{x}{b}\right)^{a}\right]^{-q}$
Quantile	$b\left[u^{-1/p}-1\right]^{-1/a}$	$b\left[(1-u)^{-1/q}-1\right]^{1/a}$
Density-quantile	$\frac{ap(u^{-1/p}-1)^{(1-ap)/a}}{b(1+(1/(u^{-1/p}-1)))^{p+1}}$	$\frac{aq((1-u)^{-1/q}-1)^{(a-1)/a}}{b(1-u)^{-(q+1)/q}}$
(kth) Moment distribution	$B(\frac{(x/b)^a}{1+(x/b)^a};p+k/a,1-k/a)$	$B(\frac{(x/b)^a}{1+(x/b)^a};1+k/a,q-k/a)$
Moments (kth moment)	$\frac{b^k\Gamma(p+k/a)\Gamma(1-k/a)}{\Gamma(p)}$	$\frac{b^k\Gamma(1+k/a)\Gamma(q-k/a)}{\Gamma(q)}$
Gini	$\frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(2p)\Gamma(p+1/a)}-1$	$1-\frac{\Gamma(q)\Gamma(2q-1/a)}{\Gamma(q-1/a)\Gamma(2q)}$

Source: Kleiber and Kotz (2003)

Note: a, b, p, q are positive, $0 < u < 1$ and $x > 0$

$$J^{(1)} = {}_3F_2\left[1, p+q, 2p+\frac{1}{a}; p+1, 2(p+q); 1\right],$$
$$J^{(2)} = {}_3F_2\left[1, p+q, 2p+\frac{1}{a}; p+\frac{1}{a}+1, 2(p+q); 1\right],$$

where ${}_3F_2$ is the generalized hypergeometric function.

Amongst the special cases of the GB2 distribution are Dagum distribution($q = 1$) and the Singh–Maddala distribution ($p = 1$). These distributions are three-parameter distributions and the functions describing them are available in closed form. We provide the moments, Gini, density, cdf, quantile, density-quantile and moment distribution functions for the Dagum and the Singh–Maddala distributions in Table 1.

3 Simulation Results

In order to assess the effectiveness of the MQD method to the ML method as described in Sect. 2.2, we perform a simulation study. In this study, we assumed knowledge of the "true" distribution. The data was simulated from a GB2 distribution with parameter settings derived from our estimates obtained from income data of Austria ($a = 3.03, b = 21.71, p = 1.35, q = 1.61$), from a Dagum distribution with parameters ($a = 3.03, b = 21.71, p = 1.35$) and from a Singh–Maddala distribution with parameters ($a = 3.03, b = 21.71, q = 1.61$) as described in Sect. 2.1 for the MQD and Sect. 2.2 for the ML methods, respectively. For every distribution, we simulate $k = 5000$ and $k = 10{,}000$ observations in each trial and repeat this process for a total of $K = 500$ trials. Subsequently, we establish $N = 9$ and $N = 4$ group income boundaries based on the respective quantiles of the simulated data. For each of the K data sets, we estimate the parameters of the underlying GB2, Dagum and Singh-Maddala distributions using the two estimation methods.

Table 2 Mean squared errors for estimated distribution parameters (simulation results)

Groups	Parameters	GB2		Dagum		Singh-Maddala	
		MQD	ML	MQD	ML	MQD	ML
k = 5000							
	a	0.640	0.607	0.007	0.007	0.007	0.006
	b	10.761	24.614	0.868	0.939	1.228	1.228
	p	22.601	5.435	0.014	0.015	–	–
N = 9	q	7.924	16.250	–	–	0.303	0.030
	Mean	0.024	0.025	0.098	0.107	0.022	0.022
	Median	0.022	0.021	0.045	0.044	0.021	0.021
	Gini	8.9e−06	8.1e−06	2.7e−05	2.8e−05	9.8e−06	9.9e−06
	a	3.697	4.525	0.037	0.019	0.018	0.015
	p	81.313	392.104	3.066	3.127	4.318	3.655
	p	12.700	70.682	0.423	0.062	–	–
N = 4	q	39.314	173.489	–	–	0.122	0.105
	Mean	0.023	0.023	0.107	0.105	0.021	0.023
	Median	0.024	0.022	0.046	0.041	0.022	0.022
	Gini	1e-05	1e-05	2.6e−05	2.5e−05	1e-05	1.1e−05
k = 10,000							
	a	0.293	0.331	0.003	0.003	0.004	0.004
	b	1.059	1.955	0.436	0.495	0.688	0.604
	p	0.564	1.074	0.007	0.008	–	–
N = 9	q	0.965	2.065	–	–	0.017	0.014
	Mean	0.013	0.011	0.051	0.049	0.011	0.011
	Median	0.011	0.011	0.0422	0.023	0.010	0.010
	Gini	4.9e−06	4.7e−06	1.4e−05	1.4e−05	5.3e−06	5.7e−06
	a	1.996	2.290	0.008	0.008	0.009	0.009
	p	38.390	156.787	1.811	1.381	1.481	2.049
	p	7.262	21.837	0.047	0.026	–	–
N = 4	q	20.652	73.835	–	–	0.029	0.054
	Mean	0.011	0.012	0.055	0.048	0.011	0.011
	Median	0.012	0.012	0.022	0.022	0.010	0.010
	Gini	4.8e−06	4.6e−06	1.4e−05	1.4e−05	4.6e−06	5.4e−06

The reported results are based on 500 Monte Carlo replications with the following parameters setting: $a = 3.03, b = 21.71, p = 1.35, q = 1.61$ for the GB2 distribution (Austria) and $p = 1$ and $q = 1$ for the Dagum and the the Singh–Maddala distributions respectively

Table 2 presents the Mean Squared Error (MSE) results for the given parameter settings of the considered case, which were obtained from 500 independently simulated data sets.

The MSE are decreasing with increasing sample size, indicating the consistency of the estimates. The MSE of the distribution parameters estimates exhibit negligible differences and are consistently small for both the MQD and ML methods when

employing the Dagum and the Singh-Maddala distributions. However, the GB2 distribution estimated with the ML method has much larger MSE than the estimates computed with the MQD method which reflects that the MQD is more robust and less sensitive to starting values.

Notably, the disparities in the estimates of the mean, the median and the Gini coefficient between the grouped and raw data are minimal. This implies that the process of grouping data only results in modest reductions in estimation uncertainty when it comes to the income distribution and related metrics such as the Gini coefficient. This finding carries substantial implications given the prevalent use of grouped data in international income analysis. It challenges the common assumption that grouped data entails significant statistical limitations compared to raw data.

4 Applications to Income Data

We use income deciles data for 30 European countries for the year 2010. The income we use is equivalized disposable income in purchasing power parities and has been scaled by a thousand (the given income divided by 1000). Table 4 in Appendix A. Tables shows a complete list of the countries used with their country codes and names as given in EU-SILC. The average sample size is 7836. The used income deciles along with the mean incomes and Gini coefficients for each country are provided in Table 5 in Appendix A. Tables.

We have estimated the income deciles directly from the cross-sectional microdata set "EUSILC UDB 2011 version 2 of August 2013", Eurostat (2011). This cross-sectional data set is part of the EU-SILC data which provides representative data on income, poverty, social exclusion and living conditions for most of the European countries. The EU-SILC data for each country is provided to Eurostat by the relevant national statistical offices which collect the data according to the methodology suggested by Eurostat. We provide more computational details and the full `R` code for replicating the results in Appendix C. code.

Our estimates are based on nine income deciles $[q(u_1), q(u_2), ..., q(u_8), q(u_9)]$ and income quintiles $[q(u_2), q(u_4), q(u_6), q(u_8)]$. Table 3 provides the absolute differences between the empirical estimates for the Gini coefficients, the mean, the observed

Table 3 Absolute error of the estimates (difference between parametric estimates and estimates from the microdata)

Method	Distribution	Deciles			Quintiles		
		Quantiles	Mean	Gini	Quantiles	Mean	Gini
MQD	GB2	57.87	229.64	0.01253	91.99	331.25	0.01981
	Dagum	109.67	182.74	0.00883	123.32	230.70	0.01097
	Singh–Maddala	89.29	197.74	0.00989	109.32	286.14	0.01378
MLE	GB2	57.59	227.94	0.01234	91.24	316.73	0.01948
	Dagum	108.08	169.44	0.00880	121.12	227.22	0.01081
	Singh–Maddala	88.48	200.45	0.01010	108.32	284.72	0.01377

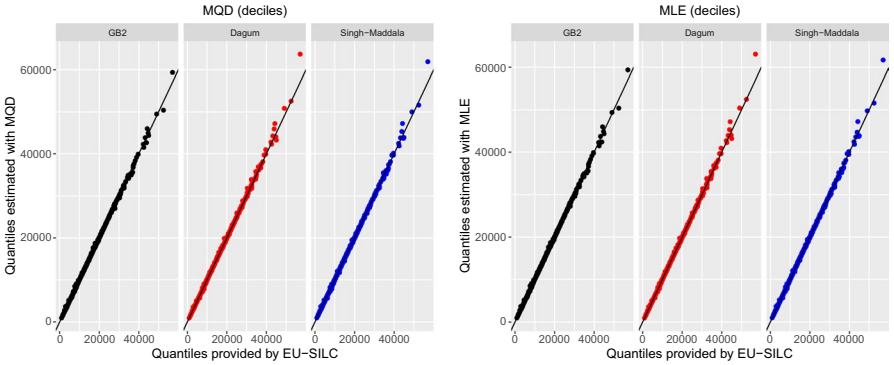

Fig. 1 Q-Q plots observed vs. estimated quantiles with MQD and MLE (all countries)

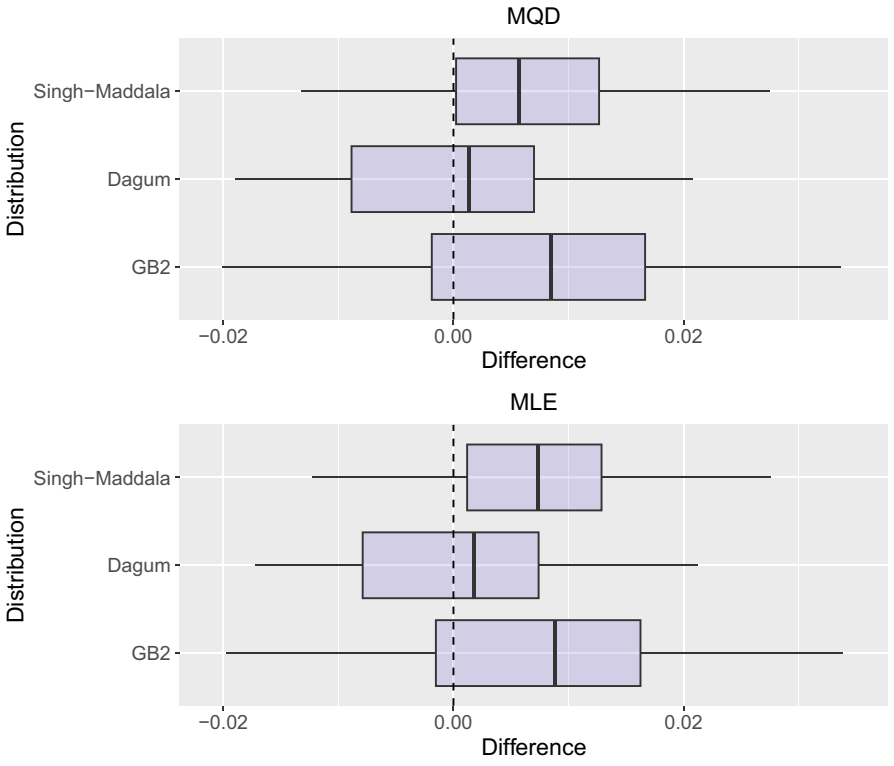

Fig. 2 Differences between estimated and observed Gini index (estimates from deciles ($N = 9$))

quantiles and their parametric counterparts approximated using the suggested parametric model. The columns “quantiles" provide the average absolute difference for all the estimated quantiles and countries. The GB2 estimates provide the smallest absolute differences for both methods MQD and ML. However, the GB2 is very sensitive to the starting values of the computation algorithm, especially for the ML method. For

some of the countries we had to adjust them in order to have convergence of the algorithm. In terms of mean and Gini coefficients, the Dagum distribution is the one which provides the smallest differences between observed and estimated values.

Figure 1 displays the empirical quantiles plotted against the theoretical quantile functions aggregated together for all the considered countries. The theoretical quantiles are estimated with the MQD and the ML methods using deciles ($N = 9$) grouped data. It is confirmed that the GB2 distribution provides the best estimates also for the distribution tails.

Figure 2 shows boxplots of the differences between the estimated and the observed Gini index for all the considered countries, methods and distributions using deciles ($N = 9$) grouped data. The GB2 estimates have the largest median (≈ 0.009) and deviation from the observed values and thus confirm the results in Table 3. The difference from the observed Gini coefficients are in the interval $[-0.02;0.035]$. The Dagum distribution provides the best estimate with the smallest median (≈ 0.006 estimated with MQD method).

5 Conclusion and Further Research

Considering the importance of the exact estimation of inequality and adding the fact that still only sparse income data is available for many countries, it is crucial to find a well-performing method for estimating income distributions. This work proposes a method for estimating the income distribution when only quantile data is available. We suggested the MQD method and applied it to the GB2, Dagum and Singh–Maddala distributions. We use decile ($N = 9$) and quintile ($N = 4$) grouped data as starting values for 30 European countries. We note that the absolute differences between the parametric estimates and their nonparametric counterparts estimated from the microdata are preserved when we use quintiles instead of deciles. These results are confirmed by a simulation study. Further, we note that MQD method is more robust than the ML technique in terms of starting values for the underlying computation algorithm, especially for the GB2 distribution. The Dagum and the Singh–Maddala distributions are outperformed by the GB2 in terms of absolute differences between the estimated parametric quantiles and their observed counterparts. The Gini coefficient and the mean are best approximated with the Dagum distribution irrelevant of the estimation technique.

One of the potential computational challenges associated with the methodology used in this article relates to the starting values for the underlying computation algorithm. During our simulation study, we faced difficulties when we started with a different value than the mean of the considered data, for the scale parameter b. Therefore, we always used as initial value the mean of the given data. An interesting extension of the study could investigate the impact on accuracy when the mean of the underlying data is not available as a starting point for estimation. This would involve exploring alternative measures or techniques that can serve as effective substitutes for the mean in the estimation process.

Appendix A: Tables

Table 4 shows a complete list of the countries used in this work with their country codes and names as given in EU-SILC. The average sample size is 7, 836. The population size of a country is computed as the sum of the product of the household size and the household weight.

Table 4 Country codes and names in EU-SILC

Country code	Country name	Sample size	Population size
AT	Austria	6187	8,315,881
BE	Belgium	5910	10,826,442
BG	Bulgaria	6554	7,518,649
CH	Switzerland	7502	7,619,680
CY	Cyprus	3917	839,751
CZ	Czech Republic	8866	10,434,558
DE	Germany	13,512	80,845,125
DK	Denmark	5331	5,512,919
EE	Estonia	4993	1,328,259
EL	Greece	6029	10,991,212
ES	Spain	13,109	45,900,276
FI	Finland	9351	5,294,659
FR	France	11,360	61,359,753
HR	Croatia	6403	4,225,193
HU	Hungary	11,685	9,850,181
IS	Iceland	3018	300,766
IT	Italy	19,399	60,683,909
LT	Lithuania	5201	3,234,482
LU	Luxembourg	5464	497,640
LV	Latvia	6599	2,049,851
MT	Malta	4076	412,580
NL	Netherlands	10,492	16,526,278
NO	Norway	4628	4,961,793
PL	Poland	12,871	37,473,013
PT	Portugal	5740	10,636,979
RO	Romania	7675	21,501,653
SE	Sweden	6717	9,531,043
SI	Slovenia	9247	2,003,382
SK	Slovakia	5200	5,392,446
UK	United Kingdom	8058	61,770,154

Table 5 provides the observed quantiles used in this work for estimating the corresponding distribution parameters with the MQD method. Table 5 displays also the observed mean incomes and Gini coefficients for each country. The empirical estimates called “observed" are computed from the microdata set “EUSILC UDB 2011 version 2 of August 2013" using the `"quantile"` option of the `wtd.quantile` function from the `R` package `Hmisc` (Harrell Jr et al., 2015) and the observed mean using the function `weighted.mean` (package `stats`). The income deciles and the mean values are given in thousands of purchasing power parities. The Gini coefficients are computed with the `gini` function from the `R` package `reldist` (Handcock, 2015) using the corresponding sample weights.

Table 5 Observed income deciles and mean

Country	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	Mean	Gini
AT	11.302	14.01	16.034	18.064	20.251	22.511	25.14	28.557	34.978	22.458	0.263
BE	9.526	11.778	13.696	15.753	17.994	20.014	22.371	25.339	30.374	19.462	0.262
BG	2.28	3.224	4.032	4.84	5.7	6.658	7.655	9.171	11.814	6.725	0.351
CH	12.09	15.25	17.894	20.41	23.069	26.151	29.534	34.248	42.292	26.552	0.296
CY	10.29	12.762	14.871	17.049	19.239	21.762	24.763	29.211	37.204	22.378	0.291
CZ	5.961	7.288	8.154	8.992	9.859	10.958	12.145	13.979	17.323	11.167	0.252
DE	9.207	12.016	14.138	16.155	18.242	20.661	23.426	27.163	33.326	20.672	0.288
DK	10.331	12.82	14.72	16.697	18.68	20.77	22.993	25.986	30.973	20.417	0.267
EE	3.488	4.624	5.495	6.328	7.337	8.545	9.926	11.892	15.127	8.614	0.319
EL	4.977	6.656	8.242	9.658	11.481	13.326	15.26	17.98	22.625	13.201	0.335
ES	5.133	7.399	9.265	10.93	12.906	15.021	17.616	20.911	26.608	14.736	0.337
FI	9.906	12.039	13.998	15.843	17.744	19.764	22.165	25.238	30.237	19.633	0.258
FR	9.871	12.354	14.26	16.111	18.062	20.211	23.011	26.867	34.805	21.57	0.308
HR	3.138	4.339	5.387	6.369	7.304	8.226	9.536	11.149	14.179	8.122	0.310
HU	3.778	4.759	5.561	6.28	7.017	7.837	8.909	10.354	12.718	7.903	0.268
IS	10.579	12.706	14.25	15.584	17.145	18.808	20.738	23.185	27.44	18.717	0.235
IT	6.967	9.367	11.54	13.488	15.514	17.652	20.153	23.531	29.361	17.541	0.319
LT	2.639	3.702	4.45	5.218	6.165	7.111	8.161	9.824	12.831	7.097	0.328
LU	14.981	17.741	20.672	23.526	26.668	30.354	34.606	39.687	49.036	30.091	0.271
LV	2.365	3.456	4.165	4.899	5.666	6.696	8.033	9.652	12.59	6.955	0.353
MT	7.377	9.201	10.7	12.258	14.034	15.857	18.09	20.854	25.047	15.688	0.274
NL	10.95	13.44	15.215	16.912	18.751	20.857	23.433	26.808	32.592	20.922	0.253
NO	14.253	17.874	20.185	22.233	24.197	26.456	29.013	32.547	38.191	25.95	0.229
PL	4.034	5.163	6.171	7.189	8.207	9.354	10.585	12.546	15.874	9.494	0.310
PT	4.601	6.112	7.211	8.42	9.584	11.012	12.923	15.524	21.034	11.86	0.342
RO	1.381	1.992	2.543	3.039	3.554	4.111	4.79	5.727	7.249	4.056	0.333
SE	10.039	12.529	14.675	16.61	18.474	20.322	22.49	25.318	29.864	19.608	0.243
SI	7.394	9.611	11.231	12.476	13.798	15.26	16.859	19.101	22.592	14.815	0.238
SK	4.892	6.315	7.227	8.086	8.856	9.873	11.084	12.56	15.348	9.802	0.257
UK	8.751	11.151	13.121	15.011	17.192	19.966	23.041	27.618	34.589	20.867	0.330

Appendix B: Code

In this appendix, the code used in this work for estimating the GB2, Dagum and Singh–Maddala distributions parameters a, b, p and q with the minimum quantile distance method is provided. To reduce precision loss (due to disproportionately large parameter values) in our computations, we scale the income deciles by 1, 000 (the observed ones divided by 1, 000). We perform the optimization of the minimum quantile distance estimator $\hat{\theta}$ (given in Eq. 1) with the statistical software `R` (R Core Team, 2022) using the function `optim` (from the `R` package `stats`). We employ the `L-BFGS-B` optimization method which is a modification of the quasi-Newton method. It is crucial to set the starting value of the parameter b equal or close to the (scaled by a thousand) mean income of each country. Otherwise, the algorithm may not converge. Further, we set the initial values of $a = 3$, $p = 1$, $q = 1$ for all the observed countries.

```
#############################################################
## 1. Common functions.
## GB2 quantile function as in equation (II.8)
qgb2 <- function(u, a, b, p, q) {
 b*(qbeta(u, p, q)/(1-qbeta(u, p, q)))^(1/a)
}
## Density-quantile function as in equation (II.9)
d_qgb2_all <- function(u, a, b, p, q){
  (a*(qbeta(u, p, q)/(1-qbeta(u, p, q)))^((a*p-1)/a))/(b*beta
     (p,q)*(1+(qbeta(u, p, q)/(1-qbeta(u, p, q))))^(p+q))
}
## Alternatively one can use the predefined GB2 functions
## from package GB2 (Graf and Nedyalkova 2011)
# library(GB2)
# d_qgb2_all <- function(u, a, b, p, q){
#  dgb2(qgb2(u, a, b, p, q), a, b, p, q)
# }
## List of quantiles to take from input vector of quantiles
quantile_list <- c(1,2,3,4,5,6,7,8,9)
quantile_values <- c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)

## alternatively one can take fewer quantiles
## (e.g. 5 from all the 9 in the input vector)
#  quantile_list <- c(1,3,5,7,9)
#  quantile_values <- c(0.1,0.3,0.5,0.7,0.9)

## Inverse of matrix V defined in equation (II.3)
v_mat <- c()
for (u in quantile_values) {
  v_mat <- append(v_mat,
                  sapply(quantile_values,
                         function(a, b) { min(a,b)-a*b },
                         u))
}
v_mat <- matrix(v_mat, nrow = length(quantile_values),
                ncol = length(quantile_values),
                byrow = TRUE)
```

```
v_qr <- qr(v_mat, tol = 1e-10)

## Minimum quantile distance function as in equation (II.1)
min_quant_dist <- function(theta, eqincome, weight) {
  d <- diag(d_qgb2_all(quantile_values, theta[1],
                       theta[2], theta[3], theta[4]))
  mat_inv <- d %*% qr.coef(v_qr, d)
  qmQ <- est_quant[quantile_list] -
    qgb2(quantile_values, theta[1],
         theta[2], theta[3], theta[4])
  t(qmQ) %*% mat_inv %*% qmQ
}
############################################################
# 2. Country processing code
## For each country read input data.
## Example values: Switzerland (CH)
country <- "CH"
est_quant <- c(12.090, 15.250, 17.894, 20.410, 23.069,
               26.151, 29.534, 34.248, 42.292)
eqincome <- c(......) # from microdata set
hweight <- c(......) # from microdata set
curr_mean <- 26.552
# curr_mean <- weighted.mean(eqincome, hweight)

## Give initial values for a, b and p.
initial_values <- c(3, curr_mean, 1, 1)

## Optimize the a, b and p distribution parameter values
optim(initial_values, function(theta) {
  theta <- c(theta[1], theta[2], theta[3], theta[4])
  # Set theta[3] = 1 for Singh-Maddala
  # Set theta[4] = 1 for Dagum
  min_quant_dist(theta, eqincome, hweight)
}, method = "L-BFGS-B",
   lower = c(0.01, 0.01, 0.01, 0.01),
   upper = c(Inf, Inf, Inf, Inf))
```

Acknowledgements The author thanks Eurostat for providing the data. Any views expressed in this article are those of the author, and do not necessarily reflect the official position of Eurostat, the European Commission or any of the national authorities whose data have been used. The author would also like to thank Christian Kleiber and Kurt Schmidheiny for their insightful comments and suggestions. Comments from two anonymous referees and the editor have led to substantial improvements in the paper. Financial support by WWZ-Forum (WWZ Förderverein) is gratefully acknowledged.

Funding Open access funding provided by FHNW University of Applied Sciences and Arts Northwestern Switzerland.

Data availability The quantile data generated and analysed with the MQD method in this work is available in Table 5 in the current article. This data is computed from the microdata set "EUSILC UDB 2011 version 2 of August 2013", Eurostat (2011) which can be accessed as described in "How to Apply for Microdata Access?", Eurostat (2022). Due to the confidential nature of the data, we cannot grant access to the microdata.However, we provide `R` code for replicating the main results of this work in Appendix C. code. The data necessary for the code is available in this article.

Declarations

Conflict of interest The author declares no conflict of interest.

References

Aitchison, J., & Brown, J. A. C. (1957). *The lognormal distribution*. Cambridge University Press.

Bandourian, R., McDonald, J. B., & Turley, R. S. (2003). A comparison of parametric models of income distribution across countries and over time. *Estadistica, 55*, 135–152.

Carmody, T. J., Eubank, R. L., & LaRiccia, V. N. (1984). A family of minimum quantile distance estimators for the three-parameter Weibull distribution. *Statistische Hefte, 25*, 69–82.

Chotikapanich, D. (Ed.). (2008). *Modeling income distributions and Lorenz curves*. Springer.

Chotikapanich, D., Griffiths, W. E., Hajargasht, G., Karunarathne, W., & PrasadaRao, D. S. (2018). Using the GB2 income distribution. *Econometrics, 6*(2), 21.

Chotikapanich, D., Griffiths, W. E., & Prasada Rao, D. S. (2007). Estimating and combining national income distributions using limited data. *Journal of Business and Economic Statistics, 25*(1), 97–109.

Chotikapanich, D., Griffiths, W. E., Prasada Rao, D. S., & Valencia, V. (2012). Global income distributions and inequality, 1993 and 2000: Incorporating country-level inequality modeled with Beta distributions. *The Review of Economics and Statistics, 94*(1), 52–73.

Dagum, C. (1977). A new model of personal income distribution: Specification and estimation. *Économie Appliquée, 30*, 413–437.

Eckernkemper, T., & Gribisch, B. (2021). Classical and Bayesian inference for income distributions using grouped data. *Oxford Bulletin of Economics and Statistics, 83*(1), 0305–9049.

Eurostat: "EUSILC UDB 2011 version 2 of August 2013". (2011). *European Union Statistics on Income and Living Conditions*.

Eurostat: "How to Apply for Microdata Access?". *European Union Statistics on Income and Living Conditions*. (2022). https://ec.europa.eu/eurostat/documents/203647/771732/How_to_apply_for_microdata_access.pdf

Graf, M., & Nedyalkova, D. (2014). Modeling of personal income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth, 60*(4), 821–842.

Hajargasht, G., & Griffiths, W. (2020). Minimum distance estimation of parametric Lorenz curves based on grouped data. *Econometric Reviews, 39*(4), 344–361.

Hajargasht, G., Griffiths, W., Brice, J., Prasada Rao, D. S., & Chotikapanich, D. (2012). Inference for income distributions using grouped data. *Journal of Business and Economic Statistics, 30*(4), 563–575.

Handcock, M. S.(2015) *Relative distribution methods*. Version 1.6–4. Project home page at http://www.stat.ucla.edu/~handcock/RelDist

Harrell Jr, F. E. (2015) with contributions from Charles Dupont, and many others: *Hmisc: Harrell Miscellaneous*. R package version 3.17–0 . http://CRAN.R-project.org/package=Hmisc

Jenkins, S. P. (2009). Distributionally-sensitive inequality indices and the GB2 income distribution. *Review of Income and Wealth, 55*, 392–398.

Jöhnk, M. D., & Niermann, S. (2002). Parameter estimation with grouped data according to the linearization method—A comparison with alternative approaches. *Statistical Papers, 43*(2), 237–255.

Jordá, V., Sarabia, J. M., & Jäntti, M. (2021). Inequality measurement with grouped data: Parametric and non-parametric methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(3), 964–984.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Wiley.

LaRiccia, V. N., & Wehrly, T. E. (1985). Asymptotic properties of a family of minimum quantile distance estimators. *Journal of the American Statistical Association, 80*(391), 742–747.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica, 52*(3), 647–665.

McDonald, J. B., & Ransom, M. R. (1979). Functional forms, estimation techniques and the distribution of income. *Econometrica, 47*(6), 1513–1525.

Minoiu, C., & Reddy, S. G. (2014). Kernel density estimation on grouped data: The case of poverty assessment. *The Journal of Economic Inequality, 12*, 163–189.

Nishino, H., & Kakamu, K. (2011). Grouped data estimation and testing of Gini coefficients using log-normal distributions. *Sankhya B, 73*, 193–210.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association, 74*(365), 105–121.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.Rproject.org/

Sala-i-Martin, X. (2006). The world distribution of income: Falling poverty and... convergence, period. *The Quarterly Journal of Economics, 121*(2), 351–97.