Survey Research Methods (2024)
Vol. 18, No. 3, pp. 251-261
https://doi.org/10.18148/srm/2024.v18i3.8187
European Survey Research Association

©2024 Author(s)
CC BY 4.0

How Many Brackets Should We Ask For to Derive Adequate Metric Information for Income and Wealth?

Maximilian Longmuir[1] · Markus M. Grabka[2]
[1]City University of New York, Stone Center on Socio-Economic Inequality, Graduate Center
[2]DIW Berlin, SOEP

This paper investigates how the number of brackets and the choice of upper cut-offs in grouped data affect the metric approximation of income and wealth. The literature currently lacks a definition of what should be considered *too few* brackets or *too-low* cut-offs. Using German survey data, we show that more than six (eight) brackets and an upper cut-off at the 95th (97th) percentile are sufficient to provide an adequate approximation of the income (wealth) distribution.

Keywords: grouped data; income; gross wealth; survey design

1 Introduction

Many data sets only include information on income or wealth in brackets.[1] Advantages of grouping information in aggregated brackets include reduced survey length and, more importantly, fewer missing values (Heeringa and Suzman 1995), for example, when a respondent cannot give a precise figure or is unwilling to provide detailed information. Even though many surveys and administrative data sets include brackets in their questionnaires, there is little evidence for how many brackets should be included to adequately approximate the income or wealth distribution. Applied researchers may prefer as many brackets as possible because a higher number reduces the unknown variance within brackets. For survey designers, however, a higher number of brackets comes at a cost: Not only do more brackets make questionnaires longer, but they may also result in increased item non-response if the bracket size is too small. This paper addresses this trade-off by determining an empirical minimum number of brackets that survey designers should include.

We show a stylized version of a question module asking for disposable household income information in Fig. 1. In this example, the questionnaire provides seven brackets, five of which are characterized by a minimum and maximum value. The last bracket is typically open-ended above a threshold, which we refer to as the *upper cut-off* of the grouped distribution. At the same time, several alternative estimation methods are available, including parametric and non-parametric approaches, to derive metric information from grouped data[2] and even also tools to calculate distributional statistics from grouped data (Ho and Reardon 2012; Jargowsky and Wheeler 2018; Jenkins 2012; Scott and Sheather 1985; Von Hippel et al. 2016). However, there is no empirical evidence on the measured quality of different numbers of brackets and very little evidence on the quality of the metric approximation (Carr 2022). We argue that two configurations may potentially bias the distributional parameter of interest. In the first configuration, too few brackets measure the distribution of income or wealth. In the second, the upper cut-off is too low. In the research to date, however, there is no definition of what constitutes *too few* brackets or *too-low* cut-offs.

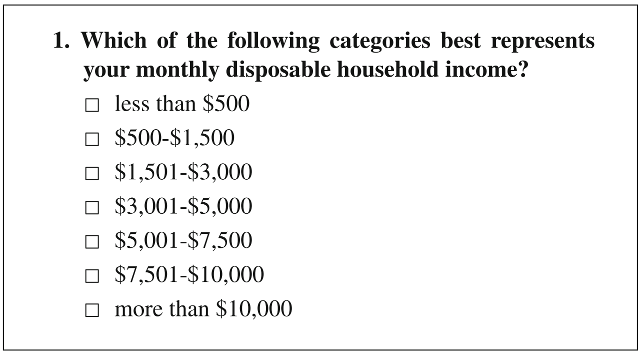
We assess the quality of grouped data information for varying numbers of brackets and cut-off points. For that cause, we artificially group metric income and wealth scales and estimate several moments of the distribution. Then, we measure the quality of those estimates by comparing them

[1] Also referred to as binned, censored, or grouped data.

[2] An overview of the different estimation methods is provided in Table 3 in the Appendix.

This article (https://doi.org/10.18148/srm/2024.v18i3.8187) contains supplementary material.

Corresponding author: Maximilian Longmuir, Stone Center on Socio-Economic Inequality, Graduate Center, City University of New York, 365 5th Ave, New York City, NY 10016, USA (Email: mlongmuir@gc.cuny.edu)



1. **Which of the following categories best represents your monthly disposable household income?**
   - □ less than $500
   - □ $500-$1,500
   - □ $1,501-$3,000
   - □ $3,001-$5,000
   - □ $5,001-$7,500
   - □ $7,501-$10,000
   - □ more than $10,000

Fig. 1

Questionnaire item asking for income: Example with seven brackets and an upper cut-off at $10,000

with those based on the original, ungrouped metric scale. In doing so, we extend the methodology proposed by Jargowsky and Wheeler (2018), who introduced a procedure to calculate distributional statistics from grouped data called mean constrained integration over brackets (MCIB). The analysis confirms that our extension of the MCIB method is an effective procedure for estimating complete metrical distributions from grouped income and wealth data– even if the moments of the original distribution are unknown. Our data bases are monthly net household income and gross household wealth data from the German Socio-Economic Panel (SOEP) in 2017.

Our analysis reveals three central findings. First, higher numbers of brackets combined with higher upper cut-offs lead to better approximations. The distribution of net income (gross wealth) can be approximated with a correlation over 0.95 if the number of brackets is higher than six (16) and the cut-off is above the 95th (97th) percentile. Second, fewer than six (eight) brackets are insufficient to produce a reliable metric approximation. Third, the quality of the approximated distribution of wealth is more sensitive to the choice of the number of brackets and upper cut-off, as it is typically more skewed. Therefore, a cut-off below the 95th percentile leads to an imprecise approximation of the wealth distribution, and more brackets do not improve it. For income, low cut-offs can be compensated for by including more brackets.

Our results provide helpful guidance for applied researchers, data methodologists, and survey designers alike. On the one hand, applied researchers can refine the approximation if their analysis relies on grouped data. On the other hand, researchers can improve their data infrastructure by choosing an appropriate number of brackets and a suitable upper cut-off when designing new questionnaires. We see the number of brackets as the primary dimension that can be controlled for, as the actual distribution of interest is typically unknown. Nevertheless, this article shows how a low upper cut-off can be compensated for with more brackets and when this increase reaches its limits.

2 Analytical Strategy and Data

2.1 Research design

Based on a long-term panel study with metric income and wealth information, we artificially group the available data with different numbers of brackets and upper cut-off limits. We chose net household income and gross household assets as they are typically fundamental elements of socioeconomic surveys. Once we set the upper cut-off, the remaining $n-1$ brackets were arranged into equally sized quantiles. For instance, if we created 12 brackets and set the 99th percentile as the upper cut-off, we defined a bracket size as $99/(12-1) = 9$ percentiles.[3] We then apply an extended version of the mean-constrained integration over brackets method (MCIB) to each setting, drawing 20,000 (15,000) values for income (wealth) from the approximated density functions. We compare the resulting values in different ways. First, we compare key moments (mean, standard deviation, various percentiles) and various inequality measures of the distribution of the two variables of interest based on the original and artificially grouped data. Second, we analyze correlation coefficients for different numbers of brackets and upper cut-off limits. Third, we show how different numbers of brackets and upper cut-offs affect the estimated percentiles of the two exemplary variable distributions.

We refrained from including zero values in our estimates for two practical reasons: first, our estimated distribution can be compared directly with the original survey data. Second, participants were only asked for their net income or gross wealth in the survey modules if it was above zero. If the lowest group included zero values, the estimation at the bottom tail might be more biased. However, this can be circumvented by offering an additional item with a zero value instead of a filter question.

[3] One could argue that the bracket size could be another variable to maximize the quality of the approximation. For comparability reasons, we defined the bracket size equally across percentiles. It may be that fewer brackets are needed if the bracket size differs across the distribution. We discuss this question with an example in Sect. 4.

2.2 MCIB

We use the mean-constrained integration over brackets method (MCIB) developed by Jargowsky and Wheeler (2018) to analyze how many brackets are needed to approximate a good fit for the underlying metric concept. We opt for the parametric MCIB method, as other approaches may ignore variances within brackets, as is the case with midpoint estimations, or need at least one moment of the actual distribution, as is the case with random empirical distribution. The MCIB approach thus can be seen as a proper procedure for our analysis as it fits the income and wealth distribution well. Carr (2022) supports this argument by comparing several estimation methods and showing that the MCIB is the best approach to estimating percentiles of the distribution.

The MCIB estimation assumes three different types of distributions within the brackets. The first bracket is specified as a uniform distribution. The following closed-ended brackets are described via linear density functions. The top bracket is characterized as an open-ended definition with a lower threshold. From there, the MCIB approach assumes a Pareto distribution, a common assumption used to approximate the top of the income or wealth distribution (Cowell 2011; Cowell and Van Kerm 2015; Jenkins 2017).

Following Jargowsky and Wheeler (2018), the MCIB estimation follows three steps: first, we estimate the density functions of the closed-end brackets. Each bracket b includes n_b households, adding to the total number N. The linear density functions for each closed-ended bracket can be defined as

$$f_b(y) = \frac{m_b y + c_b}{N}, \tag{1}$$

where m_b is the slope and c_b is the constant of the line that describes the relative frequency of households in the bracket. The slopes and intercepts for the brackets are calculated by taking the number of households in each bracket divided by the width of the bracket, which is the frequency per dollar of income for each bracket. Then, the slopes m_b are calculated as the average of the slopes from bracket $b-1$ to b and from b to $b+1$. The constants c_b are then calculated to force the line of slope m_b through the frequency point relative to the neighboring brackets, thus preserving the correct overall frequency for the bracket b (Liebenberg and Kaitz 1951). The density sums to $\frac{n_b}{N}$, i.e., the bracket's contribution to the overall income or wealth probability function.

Second, we estimate the mean and Pareto parameters for the open-ended top bracket B. Here, it is crucial to tackle the problem of unlimited possible values in the top bracket. Jargowsky and Wheeler (2018) work around this by defining the overall mean of income as the total income minus the aggregate of all income below the top bracket, divided by the number of households in the top bracket. In this formulation, the mean of the top bracket is constrained by the overall mean. Assuming linear trends in the household distribution, Jargowsky and Wheeler (2018) show that one can approximate the mean incomes in the brackets below the top bracket. Inserting this approximation in the overall mean allows us to calculate the top bracket mean μ_B, which, in turn, is required for estimating the Pareto alpha parameter

$$\alpha = \frac{\mu_B}{\mu_B - \beta}, \tag{2}$$

where β represents the upper cut-off. In our analysis, we show how different choices of upper cut-off β affect the quality of the income and wealth distribution approximation.

Jargowsky and Wheeler (2018) empirically validate their approach by grouping household incomes in 297 metropolitan areas in the United States and they compare their distributional estimations with the original distribution. They found that for several moments of the distribution, the MCIB approach provides better estimates than methods applied previously in the literature.[4]

We extended the MCIB approach by randomly generating numbers from the approximated density functions $f_b(y)$.[5] These numbers were randomly merged into the observations in the data set depending on their preassigned bracket. In this way, we imputed a full metrical distribution based on grouped information.

Our approach has limitations, as discussed in Jargowsky and Wheeler (2018). Generally, it is difficult to fit the tails of the distribution. Especially at the top, values may be exceptionally high. Jargowsky and Wheeler (2018) state that their estimates are generally reliable but lose accuracy below the 5th percentile or above the 95th percentile. As we empirically investigate different cut-offs of the top bracket directly, our contribution helps to qualify the imprecision at the tails. Nevertheless, we winsorize all our distributions at the 0.5 and the 99.5 percentile, so we cannot include the very bottom or top of the distribution.

[4] For instance, compared to approaches using the midpoint or the mean of the individual bracket; see Jargowsky and Wheeler (2018) for a detailed discussion.

[5] This step required an adjustment of the Stata command *mcib*, which only allows estimation of moments of the distribution, percentiles, and inequality measures. The authors provide the code upon request.

2.3 Data

We applied the extended MCIB to the net monthly household income and gross household wealth of German households using 2017 SOEP data. The SOEP is a panel survey of individuals in private households in Germany that started in 1984 and has been repeated annually up to the present day (Goebel et al. 2019). The SOEP currently consists of around twenty subsamples, ranging from pure random samples to oversamples of certain subgroups such as high-income households, migrants, refugees, or families with many children. All of them are randomly drawn in a multi-stage sampling procedure (Siegers et al. 2022). The first stage of the stratification is usually at the regional level (Nuts1, Nuts2, Nuts3, or municipality size) and clustered for primary sampling units (PSU). The second stage encompasses a random walk in each PSU. Information for migration or refugee samples are either coming from administrative data from the German Federal Employment Agency or the German Central Register of Foreigners.

The question about net monthly household income is asked every year at the end of the household questionnaire after various types of income such as receipt of child benefit, housing benefit, basic security in old age, social assistance, or capital income have been asked in order to obtain a better assessment of a household's financial situation.[6] The wealth module is included in the questionnaire every five years and consists of twelve asset and debt components. On the assets side, questions are asked about the market value of real estate assets, investments, private insurance, building savings contracts, and tangible assets. As the wealth module was last surveyed in 2017, we chose this year as the basis of our analysis. Our data set contained metrical distributions with 19,700 households with positive net income and 12,698 households with positive gross wealth.[7]

3 Results

This section describes our three major results. First, we show that the extended MCIB approach can be used to approximate income and wealth distributions. Second, we provide the correlation between the original and approximated income and wealth distributions for various numbers of brackets and cut-off limits. Third, we demonstrate how the different settings fit the percentiles of the distributions.

We start with an application of the extended MCIB approach. Table 1 provides key moments from the net income and gross wealth distribution based on the original and artificially grouped data using the extended MCIB. The second and fifth columns list several statistics and inequality measures for the original distribution, and the third and sixth columns show the approximation based on MCIB. Additionally, the fourth and seventh columns show the differences in percent, and the differences in percentiles are normalized by the mean. In this example, we include ten brackets and a cut-off at the 98th percentile because, for instance, the European Social Survey uses ten brackets for grouped income data.

The table shows that the estimates from the extended MCIB adequately fit the income and wealth distribution. The mean, median, and standard deviation are close to the original data in the SOEP. The distribution between the 5th and the 99th percentile is well approximated in this setting. While the mean-adjusted difference is small at the bottom tail but relatively large at the 99th percentile, with 14% for income and 61% for wealth. Generally, the mean-adjusted difference at the top is greater for wealth. Notably, the number of households with positive net income, at 19,700, is higher than the number of households with positive gross wealth, which aligns with previous findings for Germany (Grabka and Westermeier 2014). Finally, the inequality measures from the extended MCIB are also close to those in the original data. Here, the difference is provided in percent. The difference is relatively large for the mean log deviation and the Theil index for income, but the absolute difference is small. The approximated inequality measures for wealth are close to the original estimates.[8] At 0.96 and 0.91, the correlation between net income and gross wealth is high.

Table 1 shows that the extended MCIB fits several moments of the distribution well for both concepts. Therefore, fitting a metric distribution based on MCIB is feasible for net income and gross wealth. However, this finding holds for only one specific set of grouped data. We are interested in how well the distribution of income and wealth can be approximated with various brackets and upper cut-off limits.

We now focus on the correlation between the original metric distribution and the MCIB approximation from artificially grouped data. The correlation allows us to specify the quality of the imputed distribution by a simple number

[6] The original question is as follows: If you look at the total income of all of the members of your household: what is your monthly household income today? Please state the net monthly income after deductions for taxes and social security. Please include regular income such as pensions, housing allowances, child benefits, grants for higher education, maintenance payments, etc. If you do not know the exact amount, please estimate the amount per month ...euros per month.

[7] Note that the SOEP imputes missing values for both variables.

[8] Note that this holds for the distribution between the 0.5th and the 99.5th percentile. We obtain similar findings when we run this estimation with single wealth components instead of gross wealth for different survey years. Estimates are available upon request.

Table 1

Original data compared to MCIB estimations with ten brackets with an upper cut-off limit to the 98th percentile

	net household income			gross household wealth		
	orig.	MCIB	Δ%	orig.	MCIB	Δ%
mean	2654	2665	0.41	271,144	274,028	1.06
	(7)	(7)	(0.38)	(2167)	(2143)	(1.11)
sd	1748	1828	4.59	416,787	401,723	–3.61
	(7)	(9)	(0.65)	(5360)	(4465)	(1.65)
median	2250	2261	0.51	156,000	155,722	–0.18
p1	320	76	–9.18	400	533	0.05
p5	518	383	–5.10	2300	2728	0.16
p10	820	771	–1.86	5000	5420	0.15
p25	1400	1397	–0.11	23,000	23,042	0.02
p75	3500	3542	1.57	350,000	346,950	–1.12
p90	5000	4955	–1.68	610,987	672,557	22.71
p95	6000	6068	2.55	900,000	997,407	35.92
p99	9000	9381	14.35	2,300,000	2,133,898	–61.3
N	19,700	19,700		12,698	12,698	
Inequality Measures						
Mean Log Dev	0.22	0.28	25.11	1.14	1.14	–0.67
	(0.00)	(0.00)	(1.11)	(0.01)	(0.01)	(0.89)
Theil	0.20	0.22	10.03	0.73	0.72	–1.41
	(0.00)	(0.00)	(0.86)	(0.01)	(0.01)	(1.11)
Gini	0.35	0.36	3.48	0.62	0.62	0.38
	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.46)
COV	0.66	0.69	4.16	1.54	1.47	–4.63
	(0.00)	(0.00)	(0.51)	(0.01)	(0.01)	(1.00)
Rel. Mean Dev	0.25	0.26	3.07	0.45	0.46	0.73
	(0.00)	(0.00)	(0.41)	(0.00)	(0.00)	(0.52)
Correlation		0.96			0.91	
		(0.00)			(0.00)	

Compiled by authors based on SOEP v37. The table provides several moments of the net income and gross wealth distribution at the household level in 2017, with non-negative and non-zero values. The original distribution (orig.) is taken directly from the SOEP data. For the MCIB distribution, the data are artificially grouped into 10 brackets with an upper cut-off at the 98th percentile. Variables are winsorized at the 0.5th and the 99.5th percentiles. The difference in percent is mean-adjusted for the percentile estimates. Bootstrapped standard errors are in parentheses based on 500 bootstrap weights.

between –1 and 1. In our application, a correlation equal to one would describe a perfect estimation using the MCIB method. The quality of an approximation described by correlation is ultimately a normative decision: We argue that an estimation above 0.95 is an excellent approximation and one above 0.90 is a good approximation. Additionally, the estimates should fit the percentiles of the respective distribution well.

Figure 2 provides the correlation for numbers of brackets (x-axis) and several upper cut-off percentiles (y-axis) for net household income and gross household wealth, respectively. The left panel shows that the correlation is relatively high, with a value larger than 0.91, as soon as more than six brackets are included. Setting the cut-off high enough improves the correlation to a value above 0.99 with more than 16 brackets and a cut-off above the 97th percentile. Moreover, high correlations can be achieved by a relatively high number of brackets and a low upper cut-off or vice versa.

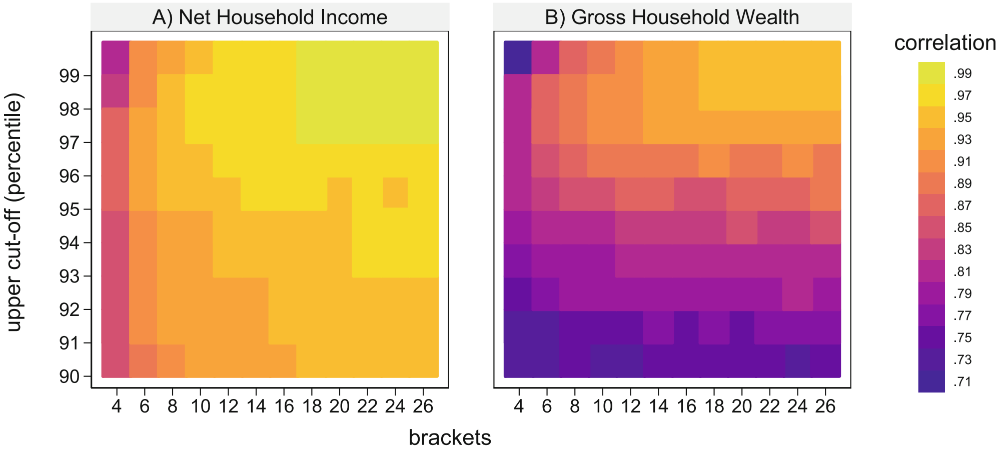


Fig. 2

Correlation between original and grouped data by number of brackets and upper cut-off percentiles. Compiled by authors based on SOEP v37. The figure shows the correlation between the original distribution and the MCIB approximation for net income (left panel) and gross wealth (right panel) at the household level in 2017. The y-axis depicts different upper cut-off limits; the x-axis shows different numbers of brackets. Variables are winsorized at the 0.5th and the 99.5th percentiles, respectively

Generating high correlations is considerably more difficult for wealth than for income. The right panel in Fig. 2 provides the same graph for gross household wealth. The correlation rises above 0.91 if more than ten brackets are included, and the upper cut-off is above the 97th percentile. Hence, the correlation is more sensitive to the cut-off setting than income. It shows that more brackets cannot offset a too-low upper cut-off. Nevertheless, if the cut-off is above the 97th percentile, the correlation can reach levels above 0.95 for more than 16 brackets.

We learn from Fig. 2 that a high correlation between the original and grouped data can be achieved in different settings of grouped data for income and wealth.[9] The higher sensitivity of gross wealth is not surprising, as the distribution of gross wealth is generally more skewed than the distribution of net household income. For income, a low upper cut-off can be compensated for with more brackets. However, this does not hold for wealth: Setting an upper cut-off below the 95th percentile generally provides approximations with a correlation of 0.85 or lower.

As a final step, we want to see how the different numbers of brackets and upper cut-offs affect the estimated percentiles of the net household income and the gross wealth distribution. This is necessary because the extended MCIB can still achieve a high correlation while structurally over- or underestimating the values of the distribution. Accordingly, Figures 3 to 6 are arranged similarly for both concepts. The figures show the mean-adjusted percentile difference in percent between the extended MCIB estimations and the original distribution on the y-axis. The percentiles of the income or wealth distribution are on the x-axis. The upper two panels provide the mean-adjusted difference for 6, 10, 14, 18, 22, and 26 brackets for the upper cut-off at the 95th percentile (left) and the 98th percentile (right). The lower two panels depict the mean-adjusted difference for the upper cut-offs at the 90th, 92nd, 94th, 96th, and 98th percentile for 10 and 26 brackets, respectively.[10]

Figures 3 and 4 display the results for net household income. In Figure 3, including more brackets gradually reduces the mean-adjusted difference along the percentiles for both cut-offs. Setting six or ten brackets leads to underestimating lower percentiles and overestimating the top percentiles to a lesser extent. The more brackets we include, the smaller the mean-adjusted difference along percentiles. However, the gain from more than ten brackets is small. We vary the upper cut-offs for ten or 26 brackets in the

[9] This is further supported by Tables 4 and 5 in the Appendix. They show descriptive statistics for the 90th, 95th, and 99th cut-offs for several bracket configurations for income and wealth, respectively. High cut-off thresholds and more brackets lead to excellent approximations of income and wealth.

[10] We include only some settings for the sake of clarity. Additionally, in Figures 7 to 10 in the Appendix, we provide bracket variation for a cut-off at the 99th percentile and cut-off variation for four brackets. The percentile differences of all other bracket and cut-off combinations are available upon request.

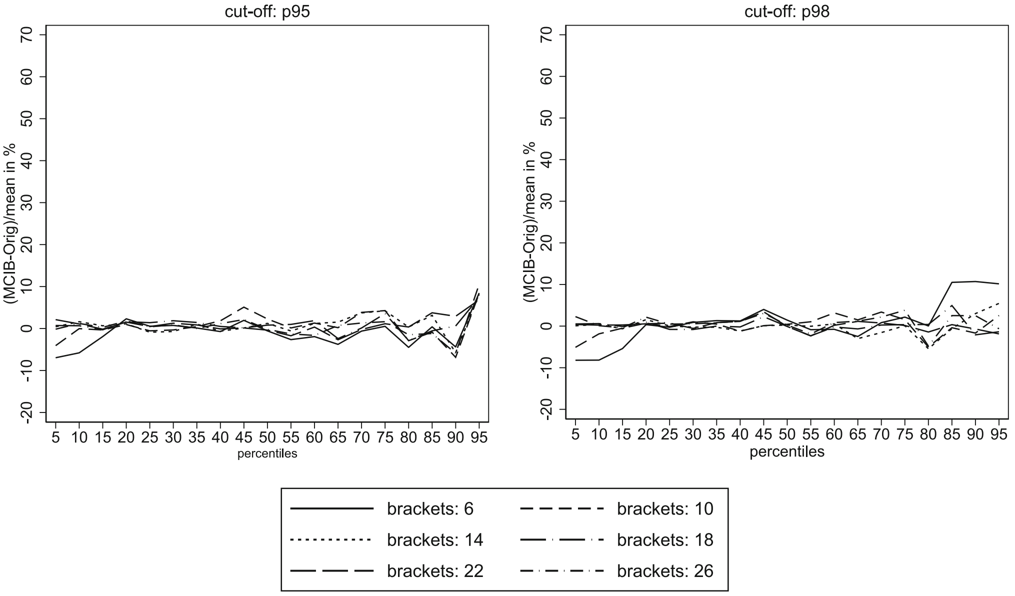


Fig. 3

Variation of number of brackets: net income. Compiled by authors based on SOEP v37. The panels show the mean-adjusted difference in percent between the original distribution and the MCIB approximation (y-axis) along percentiles of households' net income (x-axis) in 2017. The panels depict differences for various numbers of brackets. Variables are winsorized at the 0.5th and the 99.5th percentile, respectively

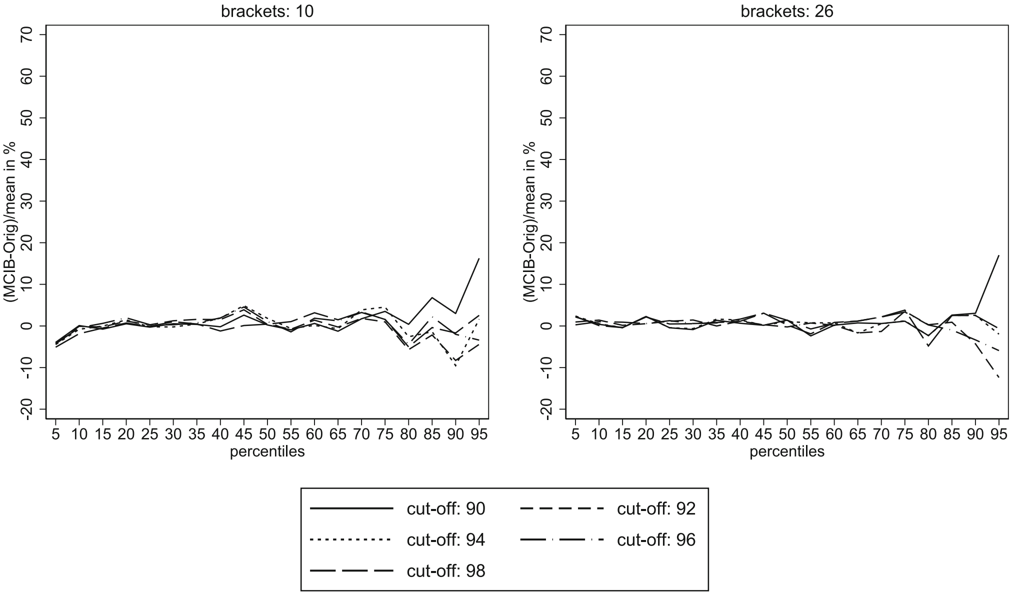


Fig. 4

Variation of cut-off limits: net income. Compiled by authors based on SOEP v37. The panels show the mean-adjusted difference in percent between the original distribution and the MCIB approximation (y-axis) along percentiles of households' net income (x-axis) in 2017. The panels depict differences for various cut-off limits. Variables are winsorized at the 0.5th and the 99.5th percentile, respectively

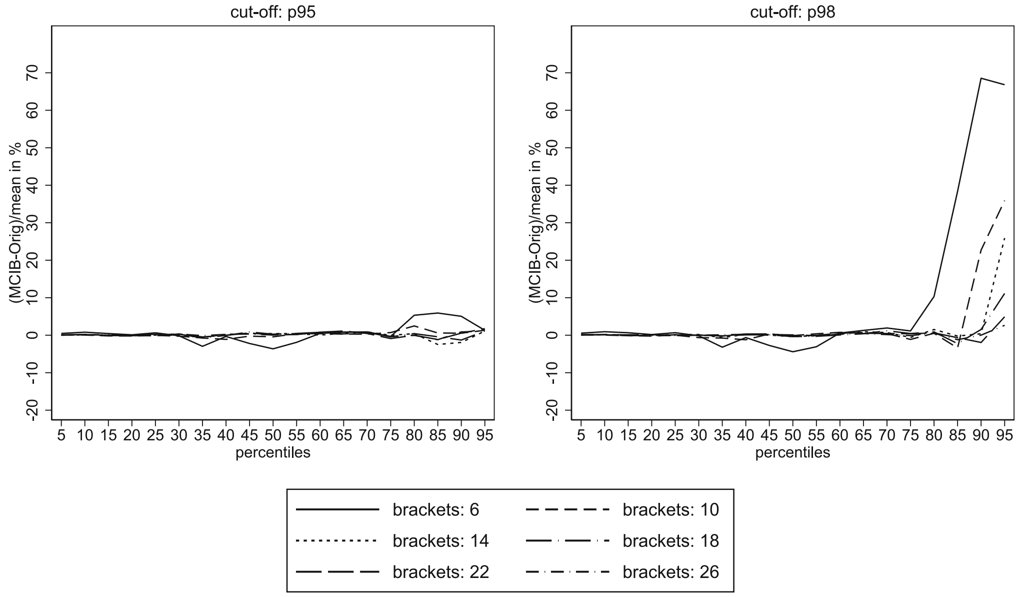


Fig. 5

Variation of number of brackets: gross wealth. Compiled by authors based on SOEP v37. The panels show the mean-adjusted difference in percent between the original distribution and the MCIB approximation (y-axis) along percentiles of gross wealth at the household level (x-axis) in 2017. The panels depict differences for various numbers of brackets. Variables are winsorized at the 0.5th and the 99.5th percentile, respectively

lower two panels. The mean-adjusted difference is small at the lower end but increases at the top. For ten brackets, a cut-off at the 90th percentile induces large mean-adjusted differences at the 85th percentile. A higher cut-off point lessens the bias up to the 90th percentile. With 26 brackets, the bias remains small up to the 95th percentile.

The figures reveal that the net income distribution is well approximated, at least up to the 85th percentile if we include more than 6 brackets. Large parts of the distribution can be approximated reasonably well even with six brackets. The distortions for lower percentiles are low, as the absolute difference is still relatively small, and the mean-adjusted differences account for this. The figures support the findings of the correlations above that the precision of the approximation reduces with fewer brackets, but it is still relatively high at later cut-offs. Therefore, we argue that more than six brackets and an upper cut-off limit above the 95th percentile limit are sufficient for a high correlation and a good fit of the percentiles for household net income.

Turning to Figures 5 and 6, the results for gross wealth are different. For all bracket settings with a cut-off at the 95th percentile, the two panels in Fig. 5 show mean-adjusted differences below 10 percent across all percentiles. At an upper cut-off at the 98th percentile, the mean-adjusted difference is up to >60 percent for 6 brackets at the top percentiles. Therefore, a setting with 6 brackets is not as close to the original distribution with a cut-off at the 98th percentile as it is with a cut-off at the 95th percentile. However, increasing the number of brackets reduces the differences in the distribution's tails. The two panels in Fig. 6 show that the gross wealth distribution's percentiles fit relatively well with several upper cut-off limits. Again, the cut-off at the 98th percentile is biased at the top, especially for ten brackets. The lower right panel shows that the percentiles seem to fit well if the number of brackets is relatively large.

Figures 5 and 6 show that gross wealth can be approximated well if more than 6 brackets are provided. In accordance with the findings above, it is more sensitive to the different cut-offs, which can distort the top 20 percent even if we use up to 14 brackets. Increasing the number of brackets reduces the bias at the top substantially. A later cut-off can additionally lead to biased top percentiles, but as we learned above, the correlation is generally higher for later levels of cut-offs. For example, with a cut-off at the 95th percentile, the gross wealth distribution's percentiles are well approximated, but the correlation is below 0.90.

Evaluating our findings for gross wealth as strictly as for net income, we advise more than 16 brackets and an upper cut-off beyond the 97th percentile for an *excellent* approximation. However, a setting with more than six brackets with a cut-off beyond the 96th percentile still achieves a *good* approximation, but in this case, one should be aware that

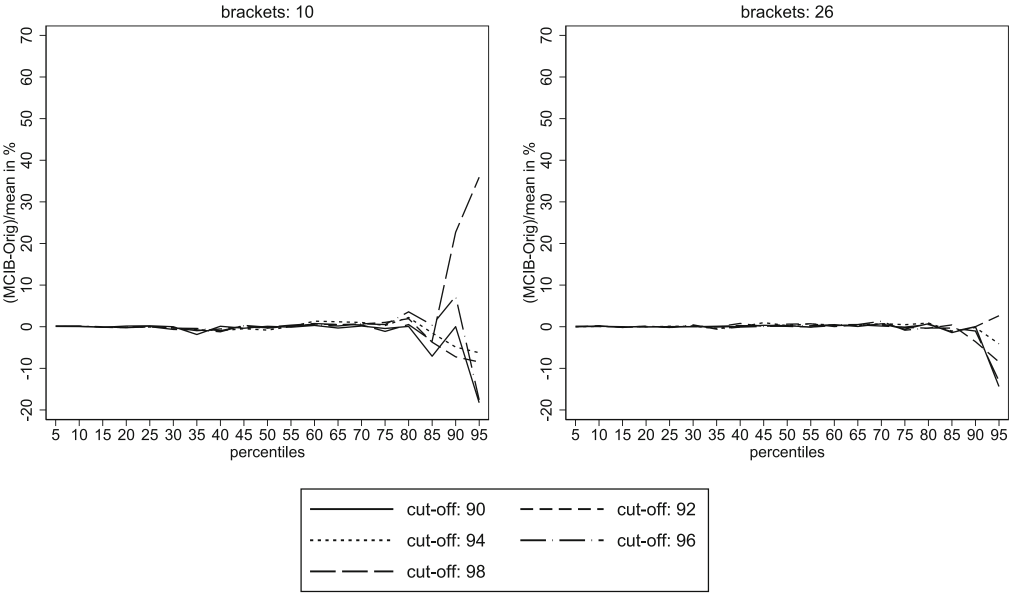


Fig. 6

Variation of cut-off limits: gross wealth. Compiled by authors based on SOEP v37. The panels show the mean-adjusted difference in percent between the original distribution and the MCIB approximation (y-axis) along percentiles of gross wealth at the household level (x-axis) in 2017. The panels depict differences for various upper cut-off limits. Variables are winsorized at the 0.5th and the 99.5th percentile, respectively

the correlation is lower. We argue that the appropriate setting depends on the application at hand: If researchers want to use individual metric values, more brackets improve the precision of the estimates. If a good approximation of the percentiles is required, fewer brackets and lower cut-offs are sufficient.

4 Discussion and Conclusion

Our analysis provides researchers with guidance on the quality of grouped data, such as information on income and wealth, with varying numbers of brackets and cut-off points. Generally, our findings encourage the use of grouped data in empirical analyses. More than 6 brackets provide close approximations of net household income, while gross wealth requires more than 16 brackets for similar quality. Nevertheless, 8 brackets are still sufficient for good approximations if the upper cut-off is at the 97th percentile or higher.

The upper cut-offs can lead to problems for researchers. Suppose no background information is available for a given country and wealth concept. In such cases, it is impossible to find a suitable cut-off ex-ante, that is, before a survey is conducted or administrative data becomes available. A solution could be to ask for metric values and brackets if the value is unknown.[11] In future years, the metric values can help to set the upper cut-off limits.

The findings proved valid after several robustness checks. One aspect that affects the quality of the MCIB procedure approximation is the underlying number of observations in the database. Our robustness analyses show that randomly reducing the number of observations does not affect the estimates substantially. Dropping 50 percent, or even 75 percent, of the SOEP sample (Table 6 in the Appendix) shows that more than 6000 (more than 2000) observations provide good approximations of the net income (gross wealth) distribution. We see this as suggestive evidence that our findings may also hold for smaller surveys.

For the sake of comparability, our analyses are based on brackets with the same percentile size. One could argue that the size variation in the brackets may affect the approximation's quality. We address this in two ways: first, we spread the brackets along the household gross wealth distribution with ten brackets and a cut-off at the 98th percentile. We show the old and the new settings in kernel density plots in Fig. 11 in the Appendix. Additionally, Table 7 shows

[11] Surveys such as the European Quality of Life Survey and the Panel Study of Income Dynamics in the United States apply this methodology.

descriptive statistics for the original and new approximation. The results remain relatively stable, with small gains at the 90th and 95th percentile and an increase in the correlation. Second, we estimate the correlations between the original and grouped income data using equally sized income brackets. Figure 12 in the Appendix shows the results. The estimates show an overall slightly smaller correlation, i.e., around 0.95 in their maximum. This is due to less precise estimates, especially at the lower end of the distribution. However, the recommended number of brackets holds. These robustness checks show that the bracket size configuration matters for the approximation quality. Some settings are superior to the quantile approach from our main analysis, while equal widths produce slightly less precise estimates. However, the use of brackets of the same size rarely occurs in practice (for example, the grouped query of income in the German microcensus), since the density of income is usually higher at the bottom of the distribution and it therefore makes sense to use a different bracketing than for the upper half of an income distribution. Our analysis suggests that including smaller brackets for lower levels of income and wealth might be favorable for capturing the true distribution.

The baseline guidelines derived from our main analysis pertain specifically to Germany in 2017, prompting the question if the results can be generalized to other national contexts. To explore this, we conducted simulations across various income distributions characterized by both elevated and diminished variances. Our findings affirm that the established baseline rule for income remains valid for distributions exhibiting a Gini coefficient typically ranging from 0.2 to 0.4. When simulating a larger variance, exemplified by a Gini coefficient of 0.6, the relevance of cut-offs increases, rendering the baseline rule of wealth more pertinent.[12] We see these findings as suggestive evidence for the potential universality of our baseline guidelines, as the simulated distributions could mirror income distributions from other nations. However, further exploration is warranted, particularly to pinpoint the precise transitional threshold between a Gini coefficient of 0.4 and 0.6, at which the baseline rule for wealth supersedes that for income. Further investigating this threshold in an international context would be a valuable extension of our study.

We conclude that the findings in this paper provide researchers with useful information about the quality of existing grouped income and wealth data and how to organize new survey items in the future.

Acknowledgements For most helpful discussions and feedback, the authors thank Charlotte Bartels, Yannick Harksen, Philipp M. Lersch, Daniel Schnitzlein, Carsten Schröder, Nhat An Trinh, and the participants of the 1st Postdoc and Assistant Professor workshop of the Berlin School of Economics. We also like to thank the editor for valuable suggestions and criticism. This research has been supported by the German Research Foundation (DFG), grant numbers GR 3239/6-1, LE 3612/5-1, and SCHN 1501/3-1.

[12] A detailed description and discussion of the simulations is provided in Appendix B.

References

Carr, A. (2022). Lorenz interpolation: A method for estimating income inequality from grouped income data. *Sociol Methodol*. https://doi.org/10.1177/00811750221085586

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

Cowell, F. A., & Van Kerm, P. (2015). *Wealth inequality: A survey*. London.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrb Natl Okon Stat*, *239*(2), 345–360. https://doi.org/10.1515/jbnst-2018-0022

Grabka, M. M., & Westermeier, C. (2014). Persistently high wealth inequality in germany. *DIW Econ Bull*, *4*(6), 3–15.

Heeringa, S. G., & Suzman, R. (1995). *Unfolding brackets for reducing item nonresponse in economic surveys*. National Institute on Aging, National Institutes of Health.

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *J Educ Behav Stat*, *37*(4), 489–517.

Hout, M. (2004). *Getting the Most Out of the GSS Income Measures*

Jargowsky, P. A., & Wheeler, C. A. (2018). Estimating income statistics from grouped data: Mean-constrained integration over brackets. *Sociol Methodol*, *48*(1), 337–374.

Jenkins S (2012) GB2FIT: stata module to fit generalized beta of the second kind distribution by maximum likelihood. Boston College Department of Economics

Jenkins, S. (2009). Distributionally-sensitive inequality indices and the GB2 income distribution. *Rev Income Wealth*, *55*(2), 392–398.

Jenkins, S. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica*, *84*(334), 261–289.

King, M. M. (2022). REDI for Binned Data: A Random Empirical Distribution Imputation Method for Estimating Continuous Incomes. *Sociol Methodol*. https://doi.org/10.1177/00811750221108086

Liebenberg, M., & Kaitz, H. (1951). An income size distribution from income tax and survey data, 1944. In: *Studies in Income and Wealth*. NBER, In, pp 378–462.

McDonald, J. B., & Ransom, M. R. (1979). Alternative parameter estimators based upon grouped data. *Commun Stat - Theory Methods*, *8*(9), 899–917. https://doi.org/10.1080/03610927908827806

McDonald, J. B., & Xu, Y. J. (1995). A generalization of the beta distribution with applications. *J Econom*, *66*(1-2), 133–152. https://doi.org/10.1016/0304-4076(94)01612-4

Scott, D. W., &Sheather, S. J. (1985). Kernel density estimation with binned data. *Commun Stat Methods*, *14*(6), 1353–1359.

Siegers, R., Steinhauer, H. W., & Schütt, J. (2022). *Soep-core v37-documentation of sample sizes and panel attrition in the german socio-economic panel* (soep)(1984 until 2020). Tech. rep., SOEP Survey Papers.

Von Hippel, P., Hunter, D., & Drown, M. (2017). Better Estimates from Binned Income Data: Interpolated CDFs and Mean-Matching. *SocScience*, *4*, 641–655. https://doi.org/10.15195/v4.a26

Von Hippel, P. T., Scarpino, S. V., & Holas, I. (2016). Robust estimation of inequality from binned incomes. *Sociol Methodol*, *46*(1), 212–251.