

Emmenegger, Jana; Münnich, Ralf T.

Article

Localising the upper tail: How top income corrections affect measures of regional inequality

Journal of Economics and Statistics

Provided in Cooperation with:

De Gruyter Brill

Suggested Citation: Emmenegger, Jana; Münnich, Ralf T. (2023) : Localising the upper tail: How top income corrections affect measures of regional inequality, Journal of Economics and Statistics, ISSN 2366-049X, De Gruyter Oldenbourg, Berlin, Vol. 243, Iss. 3/4, pp. 285-317, <https://doi.org/10.1515/jbnst-2022-0015>

This Version is available at:

<https://hdl.handle.net/10419/333250>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>

Jana Emmenegger* and Ralf Münnich

Localising the Upper Tail: How Top Income Corrections Affect Measures of Regional Inequality

<https://doi.org/10.1515/jbnst-2022-0015>

Received March 10, 2022; accepted October 18, 2022

Abstract: Poor coverage of top incomes in surveys, also referred to as the “missing rich” problem, leads to severe underestimation of income inequality. At the regional level this shortcoming is even more eminent due to small regional sample sizes. Tax records contain more accurate income information at the top and cover all regions equally well. Top-income correction approaches tackle the missing rich problem by imputing top incomes from tax to survey data. While existing methods focus on adjustments at the national level, our paper provides corrections of the *regional* income distributions in survey data by exploiting the tax data’s regional variability. We impute top incomes in the survey data from the German Micro-census based on region-specific Pareto and generalized Pareto distributions estimated from tax records. The combined survey and tax data provide new estimates of regional income inequality in Germany. Our findings indicate that inequality between and within the regions is much larger than previously understood with the magnitude of the adjustment depending on the federal states’ level of inequality in the tail.

Keywords: Gini coefficient, Generalized Pareto distribution, spatial income inequality, tax record data

JEL Classification: D31, C46, C83

1 Introduction

Economic inequality has been at the forefront of public debate for decades and is still one of the most controversial issues of our society today (Piketty 2015). It is widely recognized that inequality has negative consequences, such as bad health and crime, for society as a whole (Frank 2013). While inequality measures have for a

***Corresponding author: Jana Emmenegger**, Federal Statistical Office of Germany (DESTATIS), Wiesbaden, Germany, E-mail: jana.emmenegger@destatis.de

Ralf Münnich, Trier University, Trier, Germany

long time focused on national trends (Cowell 2000), recent research increasingly evolves into small-scale analyses (Florida and Mellander 2016; Lee et al. 2016; Moser and Schnetzer 2017). In Germany, equality of living conditions has constitutional status. Reliable statistical estimates of regional inequality are an important precondition for an informed debate and well-targeted policy. However, regional inequality estimates face major data limitations. Survey-based regional inequality estimates are constrained by small regional samples and usually biased by survey undercoverage of top incomes. This conceals the extent of regional income inequality. We address this problem by correcting survey top income distributions at the level of federal states based on regional top shares derived from tax data. We then explore the consequences of the conducted top-corrections on regional income inequality.

The literature on income inequality is bipartite: On the one hand, most official statistics and academic analyses of inequality utilise household survey samples and calculate measures of dispersion, such as the Gini coefficient, based on the full distribution of household disposable incomes. On the other hand, the literature on ‘top incomes’ is mainly based on tax return data and reports top income shares. One of the key challenges is that these two main sources of information for inequality measurement – income tax records and household survey data – yield substantially different inequality estimates. While tax data based top income shares indicate rising inequality, Gini coefficients from survey household net incomes have mostly stagnated between 2005 and 2016 (compare BMAS 2017 and Table 9 in the Appendix).

These gaps exist at the national level, but are even more pronounced at the regional level. Official policy reports find that the development of inequality in Germany in the 2000s is a data set-specific artefact (Bundestag 2017). Large parts of the contrasts stem from conceptual mismatches, including fundamentally different data frames, definitions of income and income-sharing units employed by tax and survey data. Underreportage and undercoverage of top incomes are further reasons for the deviations. Therefore, a growing number of studies reconciles tax and survey data and corrects top incomes in survey data to analyse income inequality. Burkhauser et al. (2018) find ambiguous developments of survey and tax based inequality measures in the UK and the US. Bartels and Schröder (2016) and Bach et al. (2009) provide evidence of systematic differences in the top 1% income series based on harmonised income data from the SOEP and the TPP for 2001 to 2011 with the top income concentration based on the tax data being much more pronounced.

Bartels and Metzing (2019) develop an integrated approach for top income corrections and apply their technique to different European countries. This approach is limited to the Pareto distribution and relies on ad-hoc assumptions for some parameters. Blanchet et al. (2018) propose a novel method using generalized Pareto

interpolation combined with reweighting. Moreover, Disslbacher et al. (2020) recommend a unified regression approach to estimate all parameters of a Pareto or generalized Pareto distribution jointly. However, all of these methods are limited to the national level.

This paper extends these estimation approaches to achieve *regional* top income corrections and examine their effects on inequality. We estimate parameters of regional Pareto (Bartels and Metzger 2019) and more flexible three-parameter generalized Pareto (Blanchet 2020) distributions. To overcome volatility issues of single top-income imputations, the use of multiple imputation methods enhances both top-correction approaches. Moreover, we suggest using the frames of the combined data sets as an additional criterion for comparison. Analysing the frames offers important information on how inequality differs between the entire population and the subgroup of taxpayers.

Creating a data base that covers sound income data is essential for sophisticated income analyses. The augmented data is valuable for different purposes since it enhances household data with all its socioeconomic information with reliable top income records from official tax data. It thereby allows to reevaluate various income-related research questions and to examine the impact of top income corrections on a range of inequality measures and income indicators to provide realistic information for economic and social policy. One field of application is building a sophisticated income data base for microsimulation purposes (Münnich et al. 2021).

The remainder of this paper proceeds as follows. Section 2 provides information on the tax and survey data in use and explains how we reconcile and combine the two aforementioned sources of information. Section 3 presents the empirical strategy applied to correct top incomes at the national and regional level based on multiple imputation techniques. Section 4 summarises the resulting spatial income inequality patterns from the imputed survey data and Section 5 concludes.

2 Income from Tax and Survey Data

When using survey data for income analyses, major challenges are undercoverage, sparseness and underreporting of top incomes. These phenomena cause an incomplete picture of inequality which has been studied extensively at the national level (cf. Lustig (2019) for a survey of causes of the “missing rich” issue in household surveys). These problems are much more eminent at the regional level due to the fact that the total sample is divided between the regions. Addressing these drawbacks of survey data, we introduce the Taxpayer Panel as a complementary data source in Section 2.1. The combination of the survey and tax data requires

harmonisation of the underlying conceptual definitions presented in Section 2.2. Finally, Section 2.3 compares the trends from the harmonised data at the national and regional level.

2.1 The Microcensus (MC) and the Taxpayer Panel (TPP)

In Germany, the household survey samples commonly used for income analyses, i.e. SILC, the SOEP and the household budget survey, are too small in size for regional evaluations. The German Microcensus (MC) allows for regional analyses thanks to its large sample size and the underlying regional sampling design. Furthermore, the Microcensus is the central official data base for reporting on structures of households, families and ways of life in Germany. The income concept recorded in the MC is net income, i.e. the sum of all income components post-tax post-transfer. Reporting units are the individual, the family and the household (for conceptual definitions, cf. Lengerer et al. 2005). The income data in the Microcensus faces several drawbacks: Income is self-reported, classified, and refers to the last month rather than an annual average. A common criticism is that the respondents do not indicate irregular income components and transfers very well (Hochgürtel 2019). This may lead to undercoverage of top incomes examined hereafter in Table 1. As of 2005, the MC is an intra-year survey. This implies that the household net incomes of all months of a year enter with approximately the same weight into the annual results. In principle, information on disposable income from the MC is available from as early as 1996. It is unclear how these

Table 1: Descriptive statistics for national inequality measures, 2014.

Data	Income Concept	Reporting unit	Frame	Mean	Median	Gini	Top 10-5%	Top 5-1%	Top 1%
TPP	Total income	Tax unit	Taxpayers	3757.0	2724.0	46.0	10.6	13.1	11.2
TPP	Uneq. net income	Tax unit	Taxpayers	3448.0	2732.0	40.4	10.1	11.2	8.4
MC	Uneq. net income	Tax unit	Taxpayers	1862.0	1535.0	34.3	9.4	10.5	7.4
MC	Uneq. net income	Household	Taxpayers	2637.0	2138.0	35.0	9.8	10.5	6.2
MC	Eq. net income	Household	Taxpayers	1823.0	1582.0	28.4	8.7	9.4	5.4
MC	Eq. net income	Household	Full population	1730.0	1482.0	30.2	9.0	9.7	5.7

The table shows different inequality measures for the year 2014 from the tax and survey data. The mean and median are reported in Euro and refer to monthly data. Data source: Microcensus 2014 and Taxpayer Panel 2014.

methodological changes of 2005 affect the MC annual income results (Hochgürtel 2019). For this reason, we restrict our analysis to years from 2005 onwards.

Moreover, income in the MC is classified and provided in 24 disjunct classes. However, the calculation of inequality measures requires continuous income values. Thus, we interpolate individual income reported for each household member using the generalized Pareto interpolation method developed by Blanchet et al. (2017). The parameters are estimated based on frequencies of observed incomes from the 24 MC income classes using the R package *gpinter* (Blanchet 2020). Compared to linear interpolation, the generalized Pareto interpolation yields a more realistic income distribution with a smooth, non-interrupted shape. The method provides the most appropriate picture of high incomes obtainable using only MC data. Nonetheless, within the observed income classes, there is uncertainty in the percentiles of the income distribution due to the interpolation process. We express all income measures in 2014 Purchasing Power Parities (PPP), i.e. being inflation adjusted based on the consumer price index (Statistisches Bundesamt 2020). 2014 builds the base year of our analysis thanks to data access to the full register of taxpayers (including observations beyond the panel population) for that specific year. As a final step, we apply the modified OECD equivalence scale to disposable income of the household to adjust for differences in household size and composition.

Tax data do not suffer from self-response bias or sampling errors. Furthermore, their large scale and their full regional coverage allows for detailed subgroup and small-scale analysis. Although underreportage of certain incomes may occur in the presence of tax avoidance and tax evasion, they provide the best picture of the distribution of top incomes that is available. They have widely been used for the correction of top incomes in surveys, for the analysis of top income shares (Alvaredo 2011; Atkinson 2005; Bartels and Metzger 2019) as well as for the analysis of small-scale income differences in an international context (Florida and Mellander 2016; Lee et al. 2016; Moser and Schnetzer 2017).

The administrative TPP provides reliable income information of high quality for all regions in Germany. The income concepts are pre-determined by the definition of the German Income Tax Act and change with it over time. Hence, conceptual adjustments are essential. The data captures, inter alia, yearly amounts of total income, income, taxable income, and the sum of income components. These four income concepts contained in the tax data are related as shown in Table 4 in the Appendix. However, due to tax reliefs, allowances, and advertising costs none of these income concepts is directly comparable to net incomes from the MC. There are two further drawbacks of the TPP. On the one hand, we lack lower parts of the income distribution. The “missing bottom” of the distribution consists of non-taxpayers as well as taxpayers who do not pay taxes in more than one year over the

panel time period. On the other hand, the reporting units are tax units rather than households. As the household context is unknown, it is impossible to calculate equivalent household income to take into account different household sizes. However, inequality trends of tax income over tax units do not coincide with the development of inequality of living standards of the entire population (Bartels and Metzing 2019). Only the use of equivalised disposable household income gives an impression of the development of comparable standards of living.

2.2 Harmonising the Two Sources

As neither survey nor tax data alone offer a comprehensive picture of income inequality, moving beyond the sole investigation of a single data source is important. Tax return data provide better top income and regional coverage, while survey data contain household structures and cover lower incomes. The combination facilitates understanding to what extent the different inequality results are due to conceptual differences e.g. in the definition and measurement of income or survey designs. In recent years, the international research agenda has thus focused on harmonising and combining the two sources of information, see e.g. Burkhauser et al. (2018), Piketty et al. (2018) and Garbinti et al. (2018) for the UK, the US and France, respectively. Other studies compare the income distributions from survey data to administrative income records to distinguish between different types of income measurement errors (cf. Angel et al. 2019 and Britton et al. 2019). To overcome the problem of flawed top income data in survey samples, one typical way of correcting the income distribution is to impute top incomes from tax data (see Lustig 2019 for an overview of correction approaches).

In Germany, a few studies have enriched data from the SOEP with top incomes from the Taxpayer Panel (Bach et al. 2009; Bartels and Metzing 2019). However, these correction approaches are limited to the national level and do not allow for regional analyses of (top) income disparities. Furthermore, these methods rely on single imputation which yields volatile point estimates and urges the use of highly sophisticated variance estimation methods that cover the imputation variability to provide reliable confidence intervals (Münlich et al. 2015).

To harmonise the tax return and survey data, we need to reconcile the three main differences, i.e. the frames, the income-sharing units and the income concept. In order to ensure that the frames of both data sources cover the same population for tax unit based evaluations of inequality, we define taxpayers in the MC as all individuals who report one of the following main income sources: (1) income from work, (2) assets, savings, dividends, renting or leasing, (3) pensions and (4) wage-replacement benefits. Two main groups of people are identified as non-taxpayers,

namely people with main income sources from either income from parents, spouses or other relatives and recipients of unemployment benefits (see Table 8 for group sizes). These individuals are excluded from the analysis of tax units, but included for the analysis of the entire population. There are also groups of people who are missing in the analysis of the entire population since they are neither reached by surveys nor tax data. These are predominantly individuals in a particularly precarious situation, such as prisoners or homeless people.

We construct the reporting units from the tax data in the MC based on legal status. As household affiliations are unknown in the tax data, the opposite direction is not feasible, i.e. it is impossible to construct MC household compositions in the TPP. In the reconciled MC data, all married couples are assumed to be one tax unit with joint taxation and all other individuals are defined as a single tax unit. This implies that a household with an unmarried couple corresponds to two tax units.¹ As a result, a tax unit's net income is equal to the individual interpolated income for single taxpayers and to the sum of wife's and husband's individual incomes for married couples. This approach neglects separate tax filing for married couples. However, descriptive statistics of the tax data reveal that there is only a total of roughly 800,000 cases of separate tax filing which corresponds to 3–4% of all tax reports. Bartels and Metzing (2019) choose a similar strategy to construct tax units in the SOEP.

To harmonise the income definitions in the survey and tax data, we construct net incomes from total income in the TPP as shown in Table 5 in the Appendix. Following Bach et al. (2009), we calculate economic gross income and subtract taxes, paid alimony, transfers and social security contributions, which are micro-simulated based on the Social Insurance Code. Finally, each tax unit in the TPP obtains *economic net income*. For comparability reasons, this analysis excludes negative and zero incomes, as they are not recorded in the survey. Calculating total income in the MC is not feasible due to the lack of information on individual income components, taxes and transfers.

2.3 Comparing Trends before and after Harmonisation

To demonstrate the effects of the previously described adjustments, Table 1 provides of a range of commonly used inequality indices at the national level derived from tax

¹ In Germany, married couples can choose between individual or joint tax assessments. Under joint tax filing in Germany, the income of the spouses is summed up and submitted to the tax office as a joint income tax return. If a married couple decides to file taxes together, they count as one tax unit and receive a joint tax assessment. Unmarried couples are not eligible to joint taxation and therefore represent two separate tax units.

and survey data before and after reconciliation. The first and the last lines show the originally reported income entities, whereas lines two to five reflect the adjusted, harmonised income measures from the TPP and the MC, respectively. The first two lines of Table 1 provide evidence that the different underlying income concepts account for the largest part of the gap between the data sources. In line with previous research (Bartels and Metzing 2019), the Table 1 indicates that the distribution of net income is far less unequal than the allocation of total income. The choice of the reporting unit has a much smaller impact on the observed inequality trends as indicated by the lines three and four. The last two lines reflect the adjustment of the frame, i.e. the addition of the non-taxpaying population, with the inequality in the full population being the ultimate figure of interest (presented in the very last line of Table 1). The Gini coefficient of equivalised household net income is higher for the full population in Germany (with a value of 30.2) than among taxpayers only (28.4). To sum up, the table disentangles the threefold counteracting effects of the different concepts from tax and survey data: (a) an inequality-reducing effect of income concept adjustment from total income to net income (lines 1 and 2), (b) an ambiguous impact of moving from tax units to households (lines 3 and 4), and (c) an inequality-increasing effect of extending the frame from the subgroup of taxpayers to the entire population. Note that the measures on tax unit level are not equivalised so that different demographic structures in the tax and survey data do not affect the reconciled results differently.

The differences between the second and the third line from Table 1 are not due to conceptual mismatches between the two data sources, but stem from discrepancies in the distribution of incomes. The TPP indicates higher values throughout all the regarded distributional measures in comparison to the MC. Potential reasons include opposing phenomena in the data-coverage of both sources: Whereas the tax data covers the top very well, it lacks lower parts of the distribution. In relation to the Microcensus, all individuals have the same probability of being included in the sample regardless of their income levels. The income distribution, however, is heavy tailed and right-skewed. As a result, the survey sample overrepresents the center of the distribution, while high incomes are not well covered. To correctly capture the upper tail, oversampling of top incomes would be required. Furthermore, the classification of incomes and, in particular, the open upper income class is problematic. Interpolating incomes without adding external data leads to an underestimation of high incomes.

In addition, underreportage of the personal income amounts may occur in both data sources. This phenomenon is potentially more pronounced at the top of the MC distribution (e.g. due to complex diversified portfolios or to reluctance) than in the TPP (where it mainly occurs because of tax avoidance). The last three columns of

Table 1 indicate that the top income shares differ drastically between the two data sources in the year 2014 with most severe issues occurring at the top 1%.

Table 2 compares the TPP and MC regional income indicators at the level of federal states. The compared sizes are the mean, the median, the Gini coefficient, the Top 10–5%, the Top 5–1% and the Top 1% share. The table shows the minimum, the mean, the median, the maximum, the standard deviation (SD) and the coefficient of variation (VC) of the 16 state-level income measures.

The results indicate that all regional inequality indicators from the TPP are larger than the MC counterparts, irrespective of the chosen concept. Using the uncorrected MC survey data, hence, leads to an underestimation of both within- and between-region inequality. The deviations between the two data sets are more pronounced at the top of the distribution. In particular, the top 1% shows very large discrepancies that we aim to correct. Since the regional Ginis also reflect substantial differences between the two data sets, we opt at re-evaluating top-corrected regional Ginis from the MC.

Figure 1 illustrates the different regional inequality patterns from the two data sources. The maps present the top 1% share from the MC (left) and the tax data (right) based on reconciled net incomes. The geographical patterns of top income shares differ between the data sources. On the one hand, the top 1% shares are generally larger measured from the tax data as opposed to the MC counterparts. On the other hand, the south–north and east–west divides are evident from both maps, being slightly more pronounced in the tax data. To obtain reliable statements on regional inequality of top incomes, combining the two data sources is necessary.

3 Methodological Framework

The top of the MC income distribution is underrepresented and thus requires correction from tax data estimates. We replace the top 1% of incomes from the MC with values following a Pareto distribution as explained in Section 3.1. In part 3.2, the more flexible three-parameter generalized Pareto distribution extends this approach. For sound predictions, both methods require performant multiple imputation techniques introduced in Section 3.3.

3.1 Pareto Top-Income Correction Approach

Income distributions are typically zero-inflated, right-skewed heavy tailed distributions with many outliers. The Pareto distribution is commonly used to estimate the degree of concentration of top incomes (Alvaredo 2011; Atkinson 2017). For the Pareto

Table 2: Descriptive statistics for federal state income measures, 2014.

	Tax data					Microcensus						
	Minimum	Median	Mean	Maximum	SD	VC	Minimum	Median	Mean	Maximum	SD	VC
Mean	2539	3397	3210	3932	495	15	1339	1813	1751	2081	262	15
Median	1996	2717	2543	3081	388	15	1178	1572	1475	1708	185	13
Gini	38	39	40	44	2	5	27	34	32	37	3	10
Top 10 – 5%	10	10	10	11	1	8	9	9	9	10	2	12
Top 5 – 1%	11	11	11	12	1	18	9	10	10	11	2	24
Top 1%	6	8	8	11	1	18	5	7	7	10	2	24

This table shows summary statistics of observed inequality measures of tax unit net income from the tax data (left) and the survey data (right) at the level of federal states for the year 2014. SD means standard deviation. VC is the coefficient of variation. The income data is reported in Euro and refers to monthly values. Data source: Microcensus and Taxpayer Panel 2014.

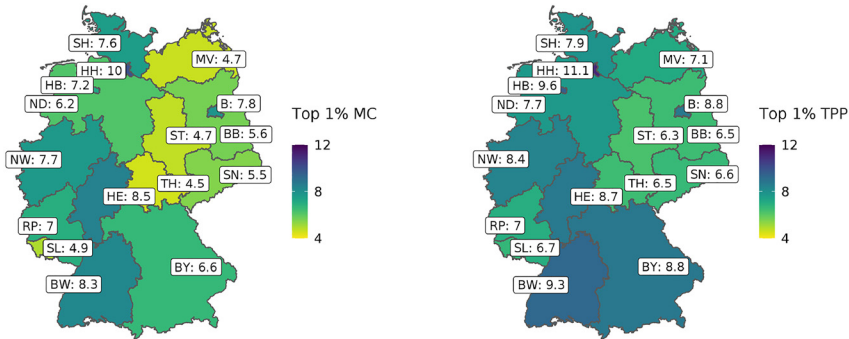


Figure 1: The top 1% incomes from survey and tax data in the federal states of Germany, 2014. The map contrasts the top 1% net income share from Microcensus survey (left side) and tax data (right side) in the year 2014 for the federal states of Germany. Data source: Microcensus and Taxpayer Panel 2014.

imputation, we follow the integrated top-correction approach developed by Bartels and Metzing (2019) to estimate the two central parameters: the Pareto coefficient α and the income threshold k above which incomes approximately follow a Pareto distribution.

We apply the Bartels and Metzing (2019) framework because it offers three major benefits: First, extending their method to the regional level is straightforward (as we will show later in this Subsection). Second, the method relies on aggregated tax-based top income shares and does not require access to the individual tax microdata. While we base our methodological development on the microdata of tax records, we intend to find a regional extension that is applicable for all researchers as soon as we publish the regional top shares. Finally, the framework yields top-corrected distributional results for any distributional measure, any income definition and any frame. Alternative correction approaches, such as combinations of inequality estimates from survey and tax data (Atkinson 2007), do not allow for comparison of different frames. These are however essential for obtaining estimates of inequality in terms of equivalised disposable household income for the entire population, i.e. the most common measure of a region's standard of living. See Jenkins (2017) or Lustig (2019) for detailed overviews of other correction approaches.

The two central assumptions of the Bartels and Metzing (2019) approach are (a) that the TPP provides an accurate picture of the upper tail of the income distribution and (b) that top incomes follow a Pareto Type I distribution. Based on these assumptions, Bartels and Metzing (2019) conduct two main steps: First, they rank tax units by their survey gross income and then replace the top 1% of the distribution with Pareto-imputed gross incomes. After the imputation, they combine tax units

into households, sum up household gross income and transform it into equivalent net household incomes based on an approximation of the tax-benefit system (by means of regression of gross on net income).

We adapt the original procedure in two ways: while Bartels and Metzing (2019) propose to impute gross earnings into the survey data; we rather construct economic net income in the tax data first. This way we can readily impute the harmonised net income values whose definition requires no further reconciliation after imputation. Due to the MC income concept, following the originally suggested order is not feasible in our case. Our careful construction of net incomes in the TPP, presented in Section 2, is at least as accurate as Bartels and Metzing's approximate transformation. Moreover, we enhance the original method by using multiple imputation techniques as described in the following Subsection.

The Pareto distribution function is given by:

$$F_{k,\alpha}(X) = 1 - \left(\frac{k}{X}\right)^\alpha \text{ if } X > k, \text{ and } 0 \text{ otherwise,} \quad (1)$$

where α is the Pareto coefficient, k is the income threshold above which incomes are Pareto distributed and X is a random variable with values x . Following Atkinson (2007), α is defined as:

$$\alpha = \frac{1}{\left(1 - \frac{\log(S_j/S_i)}{\log(P_j/P_i)}\right)}, \quad (2)$$

where P_i is the population and S_i is the income share of group i calculated from the tax data. The indices i and j refer to different fractiles of the population with $i < j$. Based on the results from Section 2, we choose to correct the top 1% which shows the largest deviations on the national level. Bartels and Metzing (2019) as well as Burkhauser et al. (2018) also adjust the top 1%. Jenkins (2017) finds that the under-coverage cut-off varies starting above 5, 1 and 0.5% over the years in the UK.

Rearranging Equation (1), the second parameter of the Pareto distribution is estimated as follows:

$$k = (1 - F(x))^{1/\alpha} * x, \quad (3)$$

where x is the 99th percentile of the MC tax unit net income distribution and $F(x)$ is 0.99.

For our analysis, the tax-data based top shares of economic net income are essential. We cannot use the top income shares provided in the WID by Bartels and Jenderny (2015) for Germany due to the differences in the chosen income concept.

However, our estimated top shares for total income indicate very similar trends to those recorded in the WID as shown in Table 10 in the Appendix.

3.2 Generalized Pareto Approach

We extend our analysis with a more flexible three-parameter generalized Pareto estimation (also referred to as the Pareto Type II distribution). The simplicity of the two-parameter Pareto distribution fits the tails of the distribution surprisingly well in many cases, however, it implies a certain rigidity that may not be appropriate for all regions. Atkinson (2017) describes Pareto's vision of an upper tail distribution with a richer functional form moving beyond the constant shape parameter α . A recent approach suggested by Blanchet et al. (2017) implements generalized Pareto curves with varying α values using a non-parametric definition of power laws. The method interpolates tabulations of exhaustive tax data with significantly higher precision compared to non-exhaustive survey data.

The cumulative distribution function of the generalized Pareto distribution is given by:

$$F_{\mu, \sigma, \xi}(X) = 1 + \xi \frac{X - \mu^{1/\xi}}{\sigma} \quad (4)$$

for $0 < \xi < 1$ where μ , σ and ξ are called the location, scale and shape parameters, respectively. The shape parameter ξ relates to Pareto's α such that $\xi = \frac{1}{\alpha}$ (Jenkins 2017). The location parameter μ is the threshold above which households' income approximately follows a generalized Pareto distribution, equivalent to Pareto's k . Rewriting Equation (4) with $\mu = k$ and $\xi = 1/\alpha$ yields the following harmonised expression:

$$F_{k, \sigma, \alpha}(X) = 1 + \frac{X - k}{\alpha \sigma} \quad (5)$$

The scale parameter σ determines the drift towards the end of the tail and defines a higher or lower income concentration compared to the two-parameter Pareto distribution. Note that the Pareto distribution is a special case of the generalized Pareto distribution with no drift by definition ($k = \alpha \sigma$).

3.2.1 Reweighting

The 99th income quantiles of the Microcensus and the Pareto distribution are aligned to the same value by definition. Since the generalized Pareto distribution is estimated purely on basis of the tax data, this is not the case. In order to avoid large gaps

between the 99th percentile of the MC and GP distributions, reweighting methods are applied. The Figures 9–12 compare the number of individuals in the 24 income classes (boundaries defined as in the Microcensus) between the tax data and the Microcensus in the status quo in the 16 federal states of Germany. The observed frequencies from the Taxpayer Panel (shown in the purple bars) provide the benchmarks for the reweighting procedure, i.e. the number of individuals in the 24 income classes calculated from the Taxpayer Panel in the year 2014. The yellow bars in Figures 9–12 illustrate the extrapolated population that is represented by the number of observations in the 24 income classes from the 1% MC sample. If the bars overlap completely, the Microcensus reflects perfect coverage of the observed tax data distribution. Note that the extrapolated population from the MC values can exceed the number of observations observed in the tax data as a result of the extrapolation. For instance, the yellow bars exceed the purple ones in the mid income range, which implies that the number of tax units observed in the MC represent more units than recorded in the tax data. The figures show that under-coverage arises in the upper tail of the MC distributions in all federal states.

To overcome these discrepancies, this analysis employs weight calibration techniques of Deville and Särndal (1992). The aim is to adjust the MC weights to represent the German taxpayer population in the best possible way. The procedure relies on a “minimum-distance” criterion which minimises the sum of differences between original and corrected weights. Truncated distance functions are used which are restricted to the range of 0–15. This well-studied calibration technique yields weights that are positive for all income recipients in the MC. The calibrated weights range from 0.0014 to 1.1672 and do thus not require any further adjustment. Brzezinski et al. (2019) follow a similar empirical strategy for the Polish taxpayer population and Blanchet et al. (2018) also suggest reweighting.

3.3 Multiple Imputation of Regional Top Incomes

We estimate Equations (1) and (5) separately for each regional entity and obtain region-specific (generalized) Pareto income distributions. For each region, we replace the top 1% of the regional MC income distributions with incomes following the regional Pareto and generalized Pareto distributions. For some regions, the top 1% of income earners consists of very few people. In the regional approach, we thus impute 50 income values in each run for every top-earning individual to receive a stable income distribution at the top and adjust the unit weights accordingly.

To map the upper tail of the (estimated) income distribution as well as possible in light of the fat-tailed distribution, we need to conduct multiple imputations to obtain stable estimates. We enhance the top correction approach by using multiple

imputation techniques with 100 draws from the estimated regional Pareto and generalized Pareto distribution, whereas other papers impute only one single income value per person (Bartels and Metzing 2019; Brzezinski et al. 2019). We have noticed that the single imputations are very volatile and do not provide a sense of the sampling variability of the estimates. To overcome these issues, we propose a multiple imputation (MI) approach to estimate inequality using top-corrected data. Jenkins (2017) is the first study that suggests multiple imputation procedures for top-coded data. Following his suggestion, we use Reiter's formulae for estimation and inference. We show how the MI approach provides consistent estimates of not only regional inequality measures but also their sampling variances, accounting for both the uncertainty introduced by the stochastic nature of the imputation process and the sampling variability.

Precisely speaking, we replace the top 1% of the MC incomes by values from the truncated (generalized) Pareto distribution with incomes above the 99th percentile by sampling quantiles z from a uniform distribution with 0.99 as the lower and 1 as the upper bound for the Pareto imputation process (because the 99th percentiles of the MC and the Pareto distribution are equal by definition). However, the 99th percentiles of the tax-data based generalized Pareto distribution and the MC differ remarkably. For this reason, the lower bound of the sampling quantiles z is defined as the quantile of generalized Pareto distribution corresponding to the 99th percentile of the MC income distribution, which ranges from 0.88 to 0.94. We rank the quantiles z and assign them to the sorted MC tax units. Finally, the quantiles are inserted into the inverse of the (generalized) Pareto distribution to obtain the imputed income value given by $y_{imp} = k \times (1 - z)^{(-1/\alpha)}$ for the quantiles $0.99 < z < 1$. Our procedure is equivalent to the one described by Jenkins et al. (2011) based on CDFs for topcoded distributions. We repeat this procedure 100 times for our regional multiple imputation approach. The variable of interest Q is the top share of the income distribution for the first part and the Gini coefficient for the second part of the analysis. We calculate the point estimators q_j of Q from the $j \in 1, \dots, 100$ partially synthetic data sets j , i.e. the top share of each imputed income distribution (Jenkins et al. 2011).

Based on Reiter's combining rules, the final multiple imputation estimate of the top share Q is the simple average of the 100 point estimates q_j : $\bar{q}_m = \sum_{i=1}^m q_j / m$. The variance of this estimate is the average of the sampling variances plus a term reflecting the finite number of imputations, $m = 100$. Reiter (2003) approach differs from Rubin (1987) rule for the combination of estimates in the fully synthetic data case. It is appropriate to follow Reiter given that we only impute the top income class, thus we deal with a partially and not a fully synthetic data case (Jenkins et al. 2011). In the fully synthetic case, the sampling variance following Rubin (1987) is

given by: $T_p = b_m + b_m/m + \bar{v}_m$, where the additional term b_m is added to account for the stochastic response mechanism.

Following Reiter (2003), the variance of the point estimator's mean is given by:

$$T_p = (b_m/m) + \bar{v}_m, \quad (6)$$

where

$$b_m = \sum_{i=1}^m \left(q_i - \bar{q}_m \right)^2 / (m - 1) \quad (7)$$

and

$$\bar{v}_m = \sum_{j=1}^m v_j / m \quad (8)$$

m is the number of imputations and v_j is the variance of the j 's imputed data set.

4 Results

This section compares our results of the multiple imputation of top incomes following Pareto Type I and II distributions at the level of federal states in Section 4.1. This first part assesses corrected top shares based on tax unit net income for the taxpaying population. Afterwards, Section 4.2 evaluates the impact of the top-correction on the Gini coefficients in terms of household net income per adult equivalent for the whole population.

4.1 Regional Top Income Distributions

For the regional imputation, the regional 1% of top-earners is assigned with new incomes following (generalized) Pareto distributions determined by region-specific parameters presented in Table 6 and Figure 5 in the Appendix. To ensure sufficient sample sizes, 50 income values are drawn per person in each run.

Figure 2 shows four versions of estimated top 1% shares of tax unit net income for the 16 federal states of Germany in the year 2014. We compare the top 1% MC income shares before imputation (yellow) to the Pareto (green) and the generalized Pareto after reweighting (blue) adjusted results from 100 multiple imputation runs. Moreover, the top 1% share from the tax data (purple) serves as a reference value. The error bars provide the confidence intervals of the shares based on bootstrapping

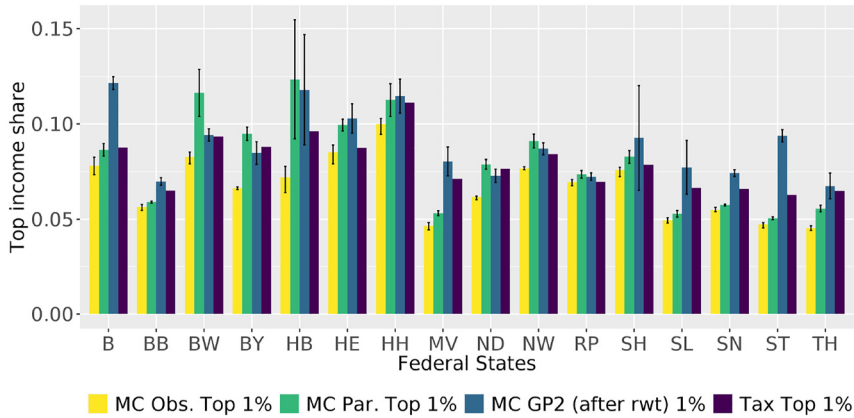


Figure 2: Regional Pareto-corrected top 1% shares in Germany, 2014. The figure shows the top-1%-share of tax unit net income from the tax data (purple), the uncorrected survey data (yellow), Pareto (green) and generalized Pareto after reweighting (blue) imputed top 1% regional shares with confidence intervals. Data source: Microcensus and Taxpayer Panel 2014.

for the MC's direct estimates and on Reiter's combining rules for the multiple imputation results.

The correction has a regionally diverging effect for the 16 federal states with three regional clusters standing out: In some states, e.g. North Rhine-Westphalia (NW) and Rhineland-Palatine (RP), the imputed shares closely resemble the tax values. In other regions, correction effects appear negligible at first sight. This group predominantly consists of states in the East of Germany, including Saxony (SN) and Thuringia (TH), inter alia. For the last group, the correction leads to large adjustments that exceed the observed tax income shares. Particularly in the states of Bremen (HB) and Hamburg (HH), these changes come along with wider confidence intervals due to the small regional sample sizes. Note that it is possible that the “real” top 1% shares are larger than observed from the tax data. This is due to the fact that the tax data does not cover the bottom of the income distribution very well. As a result, the sum of incomes of the bottom 99% is overestimated in the tax data which leads in turn to an underestimation of the share of income accruing to the top 1%. If the bottom 99% is smaller than observed in the tax data, while the sum of top 1% incomes is constant, the top 1% share increases by definition.

To shed light on potential reasons for the differences in the adjustments achieved, Figure 3 plots the observed tax data top 1% shares against the adjusted MC top 1% shares resulting from Pareto and generalized Pareto imputation on the left and right side, respectively. The values on the bisector indicate perfect

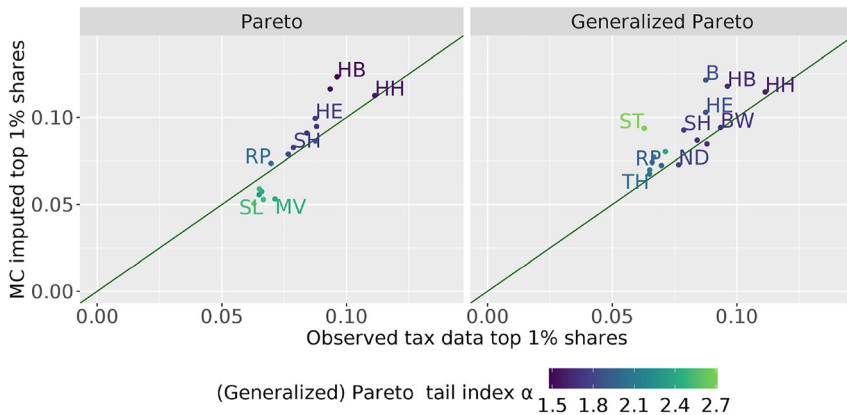


Figure 3: Correction success and Pareto indices for German regions, 2014. The scatter plot evaluates top income imputation's success by comparing the imputed MC top income shares from the Pareto (left panel) and generalized Pareto approaches (right panel) to the observed tax income shares. The lower the (Generalized) Pareto tail index α , the darker the scatter points. Data source: Microcensus 2014.

accordance of MC and TPP top shares. The color of the scatter points is darker if the (generalized) Pareto tail indices α from Equations (1) and (5) are lower.

Larger values of α go along with lower inequality within the tail (Disslbacher et al. 2020). In regions with a (generalized) Pareto tail index α substantially larger than two, the Pareto-corrected MC shares are lower than the tax shares. This is due to the fact that the proportion of very rich people, and hence the imputation effect, decreases as α increases. The regions with relatively low inequality at the top, i.e. the largest values of α from both distributions, are the five territorial Eastern states and the Saarland.

Vice versa, the lower the (generalized) Pareto index, the heavier the tail and the more unequal the distribution. Figure 3 illustrates this relation very vividly for the Pareto case. While larger Pareto α 's yield smaller adjustments, smaller α 's go along with more substantial corrections that exceed the observed tax shares. The most extensive corrections take place in Bremen (HB) as a result of the low (generalized) Pareto's α . The parameters α_{GP} at the top of the distribution (i.e. $1/\xi$ as explained above) closely resemble the α_P as illustrated in Figure 5 in the Appendix. The generalized Pareto adjustments before reweighting are generally smaller than the Pareto corrections (compare Figures 2 and 6 in the Appendix). All over-estimations observed in the GP case are consequences of the reweighting. In particular, this is true for the large GP effects observed in Saxony-Anhalt (ST) and Berlin (B).

We also examine whether the regionally varying sample sizes, especially when restricted to the upper tail of the income distribution, play a role. We define regions

with few top earners if the region’s top 1% share of income concentrates on less than 100 people in the MC sample. Table 6 in the Appendix provides sample sizes for all regions and subgroups. The 50 values per person and the 100 multiple imputation repetitions ensure sufficient variability in the data but we see that Pareto between sample variances are extremely large in regions where the Pareto tail index and the sample size are small. We correct the 5% of top earners which are based on sufficiently large samples for all regions and present the results before and after reweighting in Figures 7 and 8 in the Appendix, respectively. The top 5% correction reveals similar patterns as the top 1% imputation with the confidence intervals and the excessive adjustments being generally smaller. Overall, the top 5% correction demonstrates that our findings, in particular the three identified regional clusters, are robust.

4.2 Regional Inequality Revisited

To analyse inequality in the whole society, we calculate measures of dispersion from the raw and augmented survey data based on household disposable income per adult equivalent for the entire population including the non-taxpayers. Figure 4 contrasts the Gini coefficients before and after imputation, while Table 7 in the Appendix provides an overview of several inequality metrics for the 16 federal states.

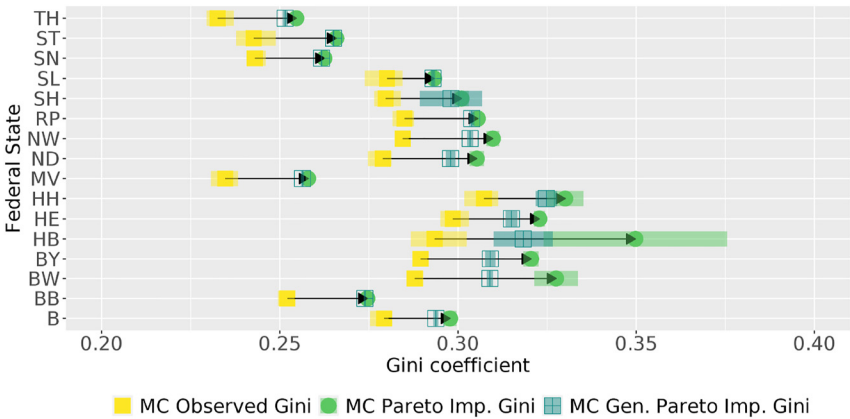


Figure 4: Top-corrected Gini coefficients for German regions, 2014. The figure compares the Gini coefficients based on household net income per adult equivalent from the uncorrected survey data (yellow), Pareto (green) and generalized Pareto (blue) imputed top 1% incomes with confidence intervals. Data source: Microcensus 2014.

Our findings indicate that the top corrected Gini coefficients are larger in all regions with the extent depending on the federal states' level of inequality in the tail. Overall, the top-corrected Gini coefficients resulting from the Pareto and the generalized Pareto approaches are very similar in magnitude. This indicates that the correction of within-region inequality is robust to distributional differences. The adjustments of the Gini coefficient range from 1 to 6 percentage points. The upward adjustment of the Gini is 8% on average. The smallest corrections (roughly 5%) occur in the Saarland (SL), while the largest increase is observed in Bremen (HB) of approximately 20% and 9% for the Pareto and generalized Pareto cases, respectively. This analysis covers the whole population including non-taxpayers, whereas the correction only affects the top 1% of the population. Hence, the confidence intervals are very narrow with the exceptions of Bremen (HB) and Hamburg (HH).

Table 3 presents the estimation results on between-region inequality derived from the Microcensus after Pareto (left side) and generalized Pareto (right side) imputation of top incomes. The table reports summary statistics for the top-corrected mean, median, Gini coefficient, and the three selected top income shares at the level of federal states based on household net income per adult equivalent. While the rise in the regional mean values is slightly more pronounced in the Pareto case, the regional medians are barely affected by the top-income imputation. Furthermore, there is more variation in the Gini coefficients and the top income shares observed after Pareto-than after generalized Pareto-correction, as indicated by the larger SD and VC values. Hence, the Pareto approach affects between-region inequality more vigorously than the generalized Pareto imputation.

Our results imply that both within- and between-states inequality is larger than survey data suggests. A potential solution to increase the accuracy of survey income data is oversampling affluent households (Disslbacher et al. 2020). The oversampling rate should ideally increase with the degree of inequality in the tail of the income distribution.

In conclusion, survey-based income inequality estimates are downward biased due to coverage problems at the top of the distribution. Ex-post imputation of high incomes from tax data is a suitable technique to derive top-corrected dispersion measures. Our results show a significant increase in inequality within and between the federal states of Germany, regardless of the specific methodological approach. The effect is more pronounced if the tail of the distribution is heavier. While top income information is limited in surveys, tax-data based correction approaches provide direly needed insights into income inequality for evidence based policies.

Table 3: Results for top-corrected federal state income measures, 2014.

	Pareto top imputed MC						GP top imputed MC					
	Minimum	Median	Mean	Maximum	SD	VC	Minimum	Median	Mean	Maximum	SD	VC
Mean	1379	1705	1679	1957	192	11	1376	1683	1662	1904	183	11
Median	1230	1454	1440	1641	134	9	1230	1454	1440	1641	134	9
Gini	25	30	30	35	3	10	25	30	29	32	2	8
Top 10 – 5%	8	9	9	9	2	13	8	9	9	9	1	9
Top 5 – 1%	9	9	9	10	2	29	9	10	9	10	1	16
Top 1%	4	6	6	11	2	28	4	6	5	7	1	16

This table shows the results of regional inequality measures of household net income per adult equivalent from the Microcensus 2014 after correction of top incomes. The left and right side refer to imputations from the Pareto and generalized Pareto (GP) distributions, respectively. SD means standard deviation, VC is the coefficient of variation. The income data is reported in Euro and refers to monthly values. Data source: Microcensus 2014 after imputation.

5 Conclusion

While the upper tail of the income distribution is particularly important for understanding economic inequality, household surveys tend to cover top incomes insufficiently. This paper adjusts data from the German Microcensus for underreporting of top percentiles by means of tax data estimates. Based on the Bartels and Metzger (2019) and the Blanchet (2020) approach, we derive region-specific Pareto and generalized Pareto distributions, respectively.

We harmonise the German Microcensus and Taxpayer Panel to examine the extent of discrepancies arising from conceptual mismatches and coverage of top incomes.

Our results indicate that the overall gap of Gini coefficients amounts to a total of 16 percentage points when comparing TPP tax unit total income of taxpayers to MC household equivalised disposable income of the entire population. Closing the conceptual gap between the two data sources reduces the difference in Gini coefficients to 6 percentage points or roughly 15%. This mismatch in inequality metrics is largely driven by the survey underreporting of top percentiles. To tackle these shortcomings, the top 1% of survey incomes is replaced by values following estimated Pareto and generalized Pareto distributions for each federal state in 100 runs. Introducing multiple imputations offer the advantages of smoothing the estimated top shares by avoiding high degrees of fluctuations in the presence of heavy tailed distributions and few observations.

Top income corrections substantially change the top shares suggesting that income inequality is vastly underestimated by raw survey data. The correction of top incomes has an average effect of 8% on the regional Gini coefficients with the most substantial increase being observed in Bremen (up to 20%). The imputation has a regionally diverging effect for the 16 federal states with three regional clusters being identified based on the magnitude of the achieved adjustment. At the same time, the inequality metrics change much less in federal states like Saxony or Brandenburg, where the tail of the distribution is less heavy. Our results imply that both within- and between-states inequality is larger than survey data suggests. To increase the accuracy of survey income data, oversampling wealthy households is recommended. While income from surveys face data constraints at the top, imputation approaches from tax data contribute essential knowledge on inequality estimates.

The paper enhances the current understanding of regional income inequality in Germany. We provide a *regional* top-income correction approach from reconciled tax and survey data based on Pareto and generalized Pareto distributions. Moreover, we extend the methodological framework by the use of multiple imputation methods to overcome volatility issues and the lack of consideration of sampling uncertainty of

single top-income imputations. We derive new estimates of regional income inequality in Germany from harmonised and combined data of the largest survey sample and tax records. Our results show that inequality within the federal states and the divergence between them are much higher than previously understood. This has important implications for policy design with respect to equal living conditions including regional planning and economic promotion of certain areas.

Acknowledgments: The research was conducted as part of the Research Unit FOR 2559 MikroSim. We appreciate the funding from the German Research Foundation. The statistical calculations were performed by the first author at the Federal Statistical Office of Germany.

Research funding: This work was supported by Deutsche Forschungsgemeinschaft under the grant no. FOR 2559.

Appendix

Table 4: Overview of income concepts in tax data.

Income component	Tax reduction
+Income from agriculture and forestry	–Allowance for agriculture and forestry
+Income from business activity	–Allowance for business activity
+Income from self-employment	–Allowance for self-employment activity
+Wage income	–(Advertising costs + pension allowance)
+Capital gains	–(Advertising costs + savings allowance)
+Income from renting and leasing	
+Other incomes	–Advertising costs lump sum
=Sum of income components	
	–Old-age lump-sum allowances
	–Exemptions for single parents
	–Income type-specific income-related expenses
=Total income	
	–Special personal deductions
	–Exceptional costs
	–Tax incentives
=Income	
	– Child allowances
=Taxable income	

The tax data contains (inter alia) the four income concepts written in bold letters. This table illustrates how these concepts are related.

Table 5: Construction of economic net income.

Income component
+Income from agriculture and forestry
+Income from business activity
+Income from self-employment
+Wage income
+Capital gains
+Income from renting and leasing
+Other incomes
+Tax-exempted foreign incomes
+Tax-exempted social transfers and child benefits
+Received alimony
=Economic gross income
–(Income tax + church tax* + solidarity surcharge)
–Social security contributions*
–Paid alimony
=Economic net income

This table shows the calculation scheme of economic net income in the tax data. The entries marked with * are partially micro-simulated.

Table 6: Estimated Pareto and generalized Pareto distributions, federal states of Germany, 2014.

Federal state	99th percentile x (in EUR)	Income threshold k (in EUR)	Tail coefficients		MC sample sizes		
			α_P	α_{GP}	Top 1%	Top 5%	Full
B	6043.90	501.00	1.85	1.86	169	841	16,320
BB	5143.80	815.90	2.50	2.14	122	625	12,453
BW	7845.80	363.70	1.50	1.69	487	2443	48,238
BY	8005.80	521.50	1.69	1.67	604	3028	61,707
HB	7032.90	311.20	1.48	1.64	28	142	2961
HE	8061.40	569.70	1.74	1.87	276	1402	27,968
HH	8143.80	420.30	1.55	1.60	79	388	8059
MV	4182.90	630.70	2.43	2.41	69	357	7227
ND	6599.60	524.00	1.82	1.77	348	1749	34,976
NW	7110.40	466.50	1.69	1.68	708	3548	73,795
RP	6919.20	699.50	2.01	1.94	172	850	18,113
SH	6736.70	493.00	1.76	1.74	123	601	12,986
SL	5525.40	881.40	2.51	2.25	46	234	4605
SN	4624.70	697.70	2.43	2.04	203	1008	20,557
ST	4187.50	739.20	2.66	2.71	109	536	11,332
TH	4436.10	599.70	2.30	2.07	110	557	11,073

The table shows the estimated Pareto distribution coefficients and corresponding sample sizes for the federal states in Germany. The income values are reported in Euro and refer to monthly data. Data source: Microcensus and Taxpayer Panel 2014.

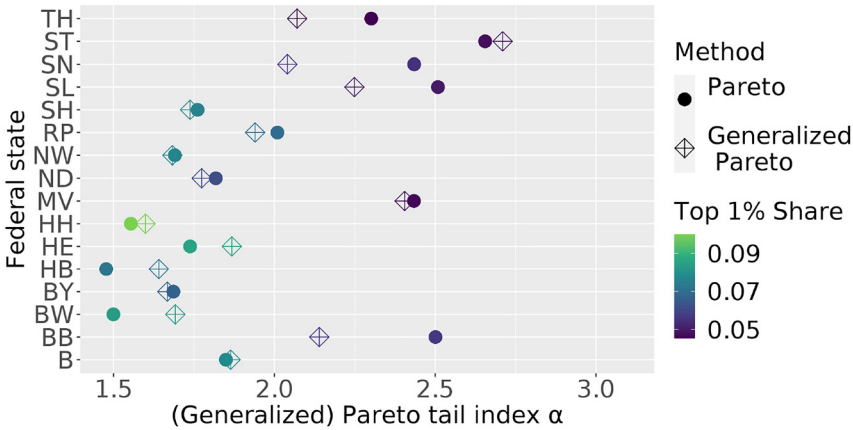


Figure 5: Estimated α coefficients for German regions, 2014. The plot shows the Pareto coefficients α_p and generalized Pareto α_{GP} from Equations (1) and (5), respectively, for the federal states of Germany. The different shapes indicate the two methods: Pareto (dots) and generalized Pareto (diamonds). The colors indicate the magnitude of the observed MC top 1% share of net equivalised income before the corrections. Data source: Microcensus 2014.

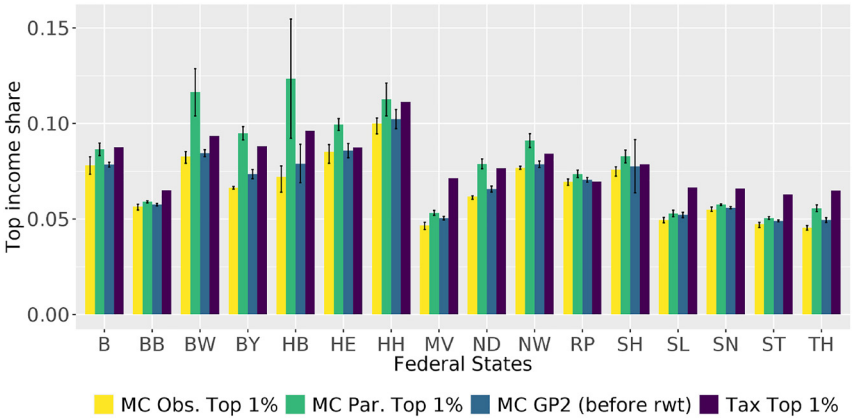


Figure 6: Corrected top 1% shares in German federal states without reweighting, 2014. The figure shows the top-1%-share of tax unit net income from the tax data (purple), the uncorrected survey data (yellow), Pareto (green) and generalized Pareto before reweighting (blue) imputed top 1% regional shares with confidence intervals. Data source: Microcensus and Taxpayer Panel 2014.

Imputation of the top 5%

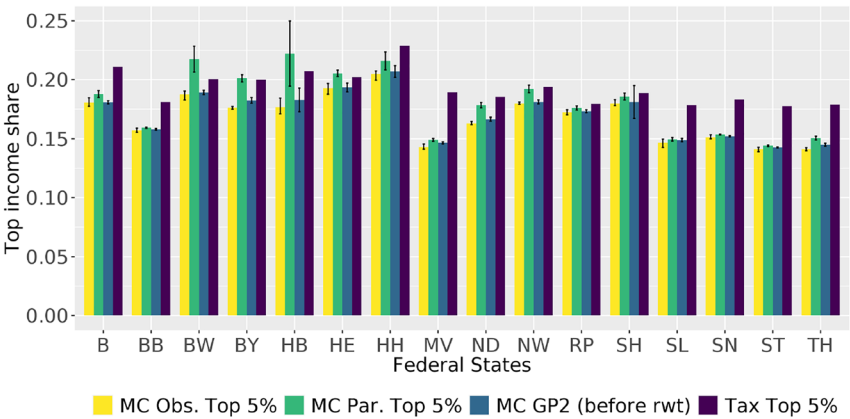


Figure 7: Corrected top 5% shares in German federal states without reweighting, 2014. The figure contrasts the regional top 5% shares based on tax unit net income observed from the tax (purple) and the survey data (yellow) with the multiply imputed Pareto (green) and generalized Pareto before reweighting (blue) results. The error bars provide confidence intervals. Data source: Microcensus and Taxpayer Panel 2014.

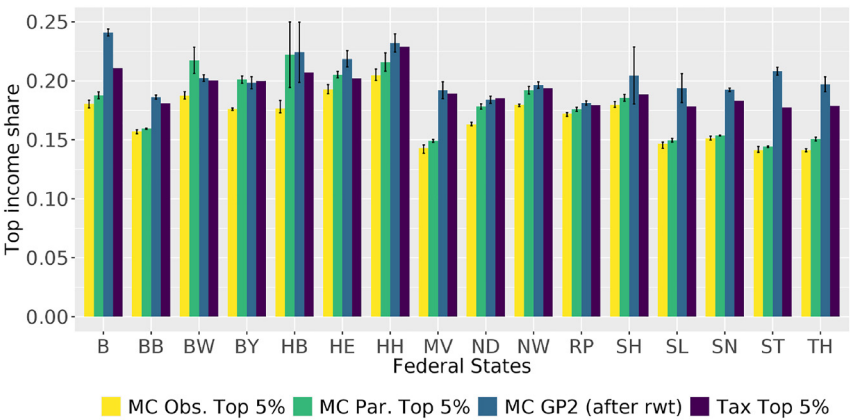


Figure 8: Corrected top 5% shares in German federal states with reweighting, 2014. The figure contrasts the regional top 5% shares based on tax unit net income observed from the tax (purple) and the survey data (yellow) with the multiply imputed Pareto (green) and generalized Pareto after reweighting (blue) results. The error bars provide confidence intervals. Data source: Microcensus and Taxpayer Panel 2014.

Table 7: Results for federal state inequality estimates, 2014.

Federal state	MC observed				Pareto			Generalized Pareto		
	Mean	Median	Gini	Top 1%	Mean	Gini	Top 1%	Mean	Gini	Top 1%
B	1628	1376	27.9	5.5	1638	29.8	6.0	1629	29.4	5.5
BB	1522	1348	25.2	4.8	1524	27.5	4.9	1523	27.4	4.8
BW	1901	1641	28.8	6.2	1957	32.8	8.8	1904	30.9	6.4
BY	1877	1613	28.9	5.7	1920	32.0	7.7	1888	30.9	6.2
HB	1589	1348	29.3	5.7	1680	35.0	10.8	1601	31.8	6.4
HE	1869	1577	29.9	6.2	1892	32.3	7.2	1870	31.5	6.2
HH	1886	1590	30.7	6.5	1904	33.0	7.4	1889	32.5	6.7
MV	1372	1230	23.5	4.2	1379	25.8	4.6	1376	25.6	4.4
ND	1705	1482	27.9	5.8	1731	30.5	7.2	1712	29.8	6.2
NW	1712	1470	28.4	5.6	1730	31.0	6.5	1714	30.3	5.7
RP	1755	1530	28.5	5.2	1761	30.6	5.4	1757	30.4	5.3
SH	1779	1545	28.0	5.8	1790	30.1	6.3	1782	29.8	6.0
SL	1650	1439	28.0	5.1	1654	29.3	5.3	1653	29.3	5.3
SN	1438	1286	24.3	4.6	1441	26.3	4.7	1439	26.2	4.6
ST	1414	1274	24.3	4.1	1417	26.6	4.3	1415	26.5	4.2
TH	1434	1295	23.3	3.7	1444	25.5	4.3	1438	25.2	3.9

This table shows our results of estimated inequality measures of equivalised household net income from the MC before and after adjustment at the level of federal states for the year 2014. The income values are reported in Euro and refer to monthly data. Data source: Microcensus and Taxpayer Panel 2014.

Table 8: Taxpayers MC and Tax Data 2014.

Definition	Microcensus		Tax data	
	Total (extrapolated in mio.)	%	Total (in mio.)	%
Population	81	100	–	–
Non-taxpayers	24	30	–	–
Taxpayers	56	70	56	100
Households	40	100	–	–
Tax units	44	100	40	100
Single tax units	27	61	24	61
Joint tax filing (married)	17	39	16	38

This table compares the population values (in million) from the Microcensus and the Tax Data for the year 2014. Data source: Microcensus and Tax Data 2014.

Detailed Comparison of the MC and TPP Income Distributions

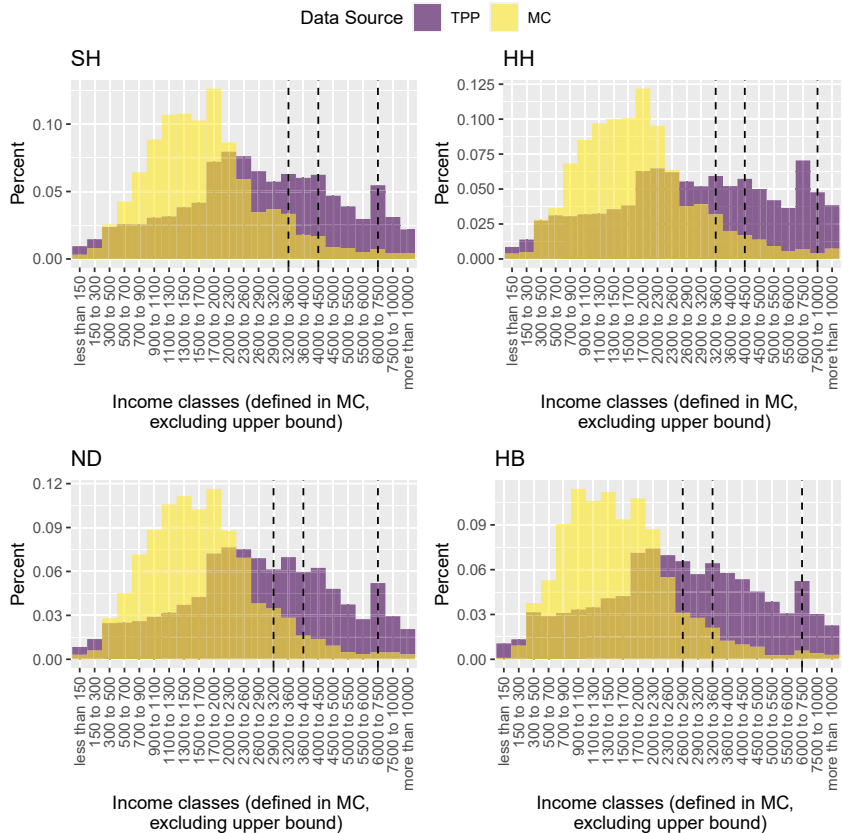


Figure 9: Comparison of the MC and TPP income distributions (1/4). The histograms compare the income distribution from the Microcensus survey data (yellow) and the Taxpayer Panel (purple) for the federal states of Germany in the year 2014. Dashed lines are 90th, 95th and 99th percentiles of the MC survey data. Data source: Microcensus and Taxpayer Panel 2014.

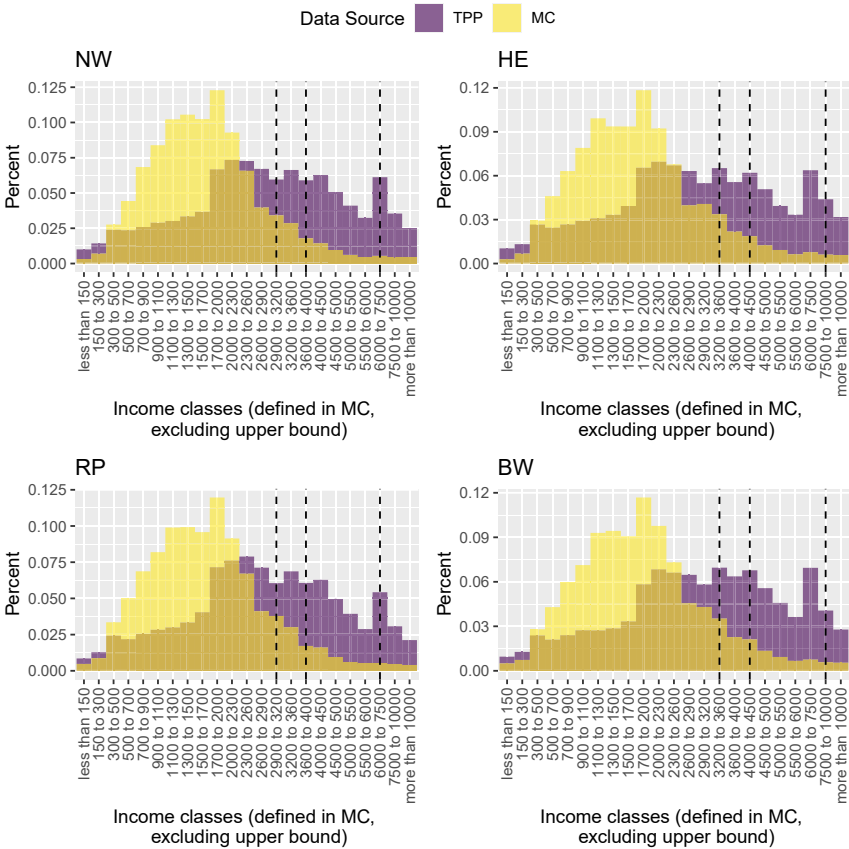


Figure 10: Comparison of the MC and TPP income distributions (2/4). The histograms compare the income distribution from the Microcensus survey data (yellow) and the Taxpayer Panel (purple) for the federal states of Germany in the year 2014. Dashed lines are 90th, 95th and 99th percentiles of the MC survey data. Data source: Microcensus and Taxpayer Panel 2014.

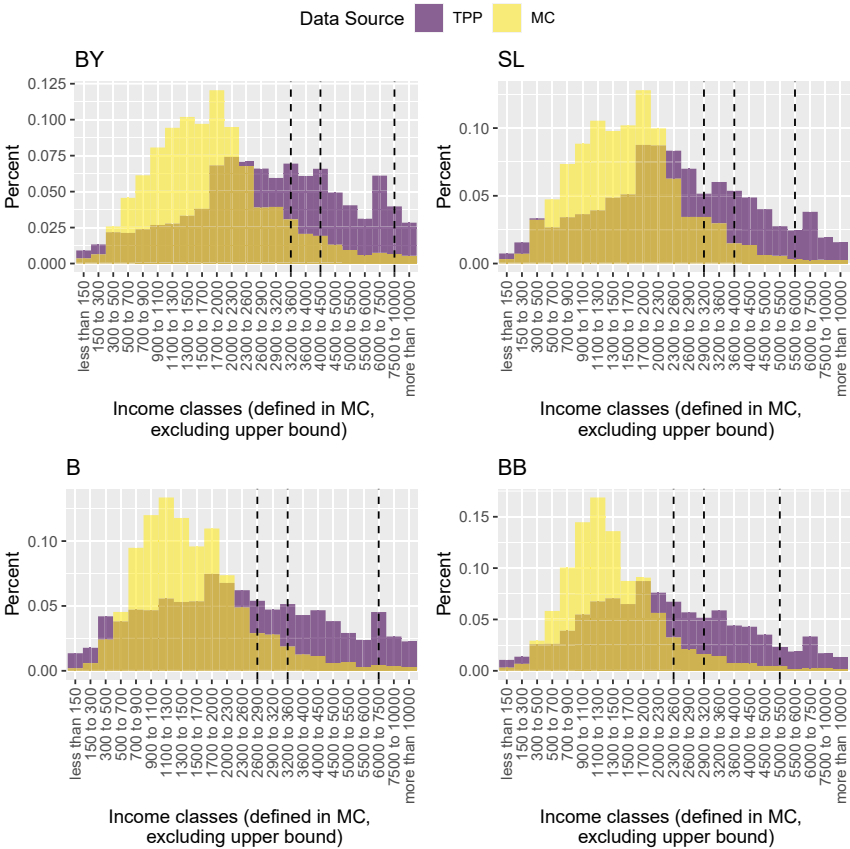


Figure 11: Comparison of the MC and TPP income distributions (3/4). The histograms compare the income distribution from the Microcensus survey data (yellow) and the Taxpayer Panel (purple) for the federal states of Germany in the year 2014. Dashed lines are 90th, 95th and 99th percentiles of the MC survey data. Data source: Microcensus and Taxpayer Panel 2014.

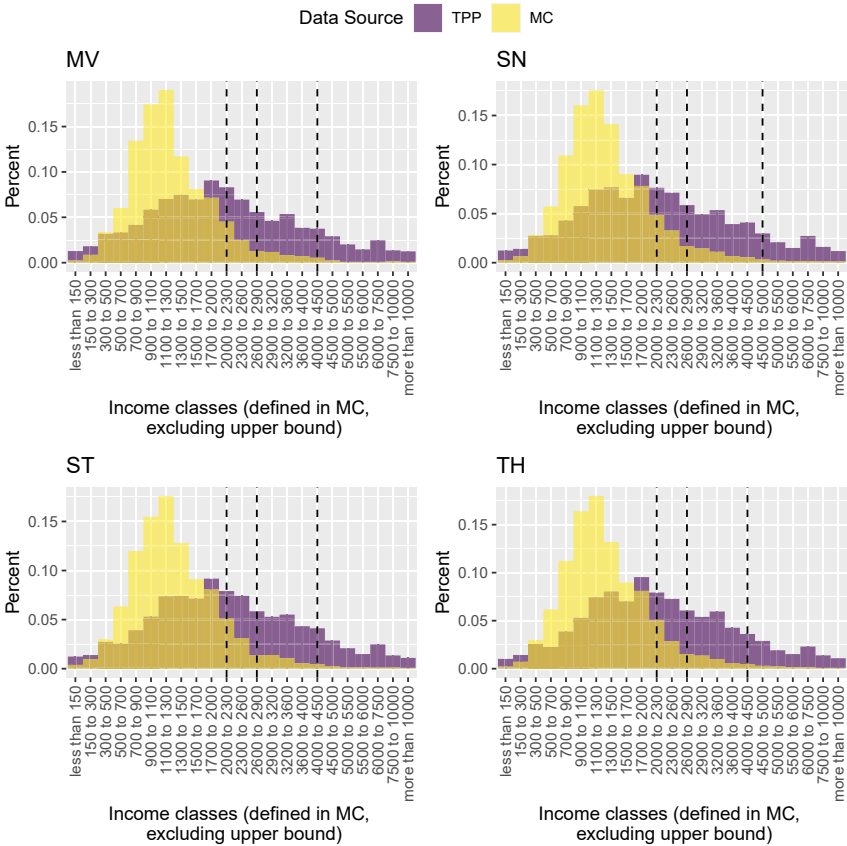


Figure 12: Comparison of the MC and TPP income distributions (4/4). The histograms compare the income distribution from the Microcensus survey data (yellow) and the Taxpayer Panel (purple) for the federal states of Germany in the year 2014. Dashed lines are 90th, 95th and 99th percentiles of the MC survey data. Data source: Microcensus and Taxpayer Panel 2014.

References

- Alvaredo, F. (2011). A note on the relationship between top income shares and the Gini coefficient. *Econ. Lett.* 110: 274–277.
- Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2019). What did you really earn last year? Explaining measurement error in survey income data. *J. Roy. Stat. Soc.* 182: 1411–1437.
- Atkinson, A.B. (2005). Top incomes in the UK over the 20th century. *J. Roy. Stat. Soc.* 168: 325–343.
- Atkinson, A.B. (2007). Measuring top incomes: methodological issues. In: Atkinson, A.B. and Piketty, T. (Eds.), *Top incomes over the twentieth century*, Vol. 1. OUP, Oxford, pp. 18–42.
- Atkinson, A.B. (2017). Pareto and the upper tail of the income distribution in the UK: 1799 to the present. *Economica* 84: 129–156.
- Bach, S., Corneo, G., and Steiner, V. (2009). From bottom to top: the entire income distribution in Germany, 1992–2003. *Rev. Income Wealth* 55: 303–330.
- Bartels, C. and Jenderny, K. (2015). *The role of capital income for top incomes shares in Germany*, 1. World Top Incomes Database (WTID) Working Paper.
- Bartels, C. and Metzing, M. (2019). An integrated approach for a top-corrected income distribution. *J. Econ. Inequal.* 17: 125–143.
- Bartels, C. and Schröder, C. (2016). Zur Entwicklung von Top-Einkommen in Deutschland seit 2001. *DIW-Wochenbericht* 83: 3–9.
- Blanchet, T., Fournier, J., and Piketty, T. (2017). Generalized Pareto Curves: theory and applications. In: *CEPR discussion paper*, Vol. DP12404.
- Blanchet, T. (2020). *Applying generalized Pareto Interpolation with gpinter*. R package version 0.0.0.9000. <https://rdr.io/github/thomasblanchet/gpinter/f/vignettes/gpinter.Rmd> (Accessed 30 November 2022).
- Blanchet, T., Flores, I., and Morgan, M. (2018). The weight of the rich: improving surveys using tax data. In: *WID. World working paper*, Vol. 12, Available at: <https://wid.world/document/the-weight-of-the-rich-improving-surveys-using-tax-data-wid-world-working-paper-2018-12/>.
- BMAS (2017). Bundesministerium für Arbeit und Soziales, Berlin. In: *Lebenslagen in Deutschland: Der Fünfte Armuts- und Reichtumsbericht der Bundesregierung*, August. Bundesministerium für Arbeit und Soziales, Berlin.
- Britton, J., Shephard, N., and Vignoles, A. (2019). A comparison of sample survey measures of earnings of English graduates with administrative data. *J. Roy. Stat. Soc.* 182: 719–754.
- Brzezinski, M., Myck, M., and Najsztub, M. (2019). Reevaluating distributional consequences of the transition to market economy in Poland: new results from combined household survey and tax return data. In: *IZA Discussion Papers*, Vol. 12734. Institute of Labor Economics (IZA), Bonn.
- Deutscher Bundestag (2017). *Sachstand Einkommensungleichheit und Armutsrisiko* quote: WD 6- 3000-071/17, Berlin: Wissenschaftliche Dienste Deutscher Bundestag. Available at: <https://www.bundestag.de/resource/blob/538870/8ca1d4131c81ce90b8af45a75381b747/WD-6-071-17-pdf-data.pdf> (Accessed 30 November 2022).
- Burkhauser, R.V., Héroult, N., Jenkins, S.P., and Wilkins, R. (2018). Top incomes and inequality in the UK: reconciling estimates from household survey and tax return data. *Oxf. Econ. Pap.* 70: 301–326.
- Cowell, F.A. (2000). Measurement of inequality. In: *Handbook of income distribution*, Vol. 1, pp. 87–166.
- Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* 87: 376–382.
- Disslbacher, F., Ertl, M., List, E., Mokre, P., and Schnetzer, M. (2020). On Top of the Top – adjusting wealth distributions using national rich lists. In: *INEQ working paper series*, Vol. 20, Available at: https://epub.wu.ac.at/7908/1/WP_20_adjusting_wealth_distributions_using_national_rich_lists.pdf.

- Florida, R. and Mellander, C. (2016). The geography of inequality: difference and Determinants of wage and income inequality across US metros. *Reg. Stud.* 50: 79–92.
- Frank, R. (2013). *Falling behind: How rising inequality harms the middle class*. Univ of California Press, Berkeley and Los Angeles.
- Garbinti, B., Goupille-Lebret, J., and Piketty, T. (2018). Income inequality in France, 1900–2014: evidence from distributional national accounts (DINA). *J. Publ. Econ.* 162: 63–77.
- Hochgürtel, T. (2019). Einkommensanalysen mit dem Mikrozensus. *Wirtsch. Stat.* 3: 53–64, <https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2019/03/> (Accessed 30 November 2022).
- Jenkins, S.P. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica* 84: 261–289.
- Jenkins, S.P., Burkhauser, R.V., Feng, S., and Larrimore, J. (2011). Measuring inequality using censored data: a multiple-imputation approach to estimation and inference. *J. Roy. Stat. Soc.* 174: 63–81.
- Lee, N., Sissons, P., and Jones, K. (2016). The geography of wage inequality in British cities. *Reg. Stud.* 50: 1714–1727.
- Lengerer, A., Bohr, J., and Janßen, A. (2005). *Haushalte, Familien und Lebensformen im Mikrozensus: Konzepte und Typisierungen, ZUMA-Arbeitsbericht 05*. Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, pp. 1–49.
- Lustig, N. (2019). The “missing rich” in household surveys: causes and correction approaches. In: *The Commitment to Equity (CEQ) Working Paper Series*, 75. Tulane University.
- Moser, M. and Schnetzer, M. (2017). The income–inequality nexus in a developed country: small-scale regional evidence from Austria. *Reg. Stud.* 51: 454–466.
- Münnich, R., Gabler, S., Bruch, C., Burgard, J.P., Enderle, T., Kolb, J.P., and Zimmermann, T. (2015). Tabellenauswertungen im Zensus unter Berücksichtigung fehlender Werte. *ASTA Wirtsch. Sozialstat. Arch.* 9: 269–304.
- Münnich, R., Schnell, R., Brenzel, H., Dieckmann, H., Dräger, S., Emmenegger, J., Höcker, P., Kopp, J., Merkle, H., Neufang, K., et al. (2021). A population based regional Dynamic microsimulation of Germany: the MikroSim model. *Methods Data Anal.* 15: 241–264.
- Piketty, T. (2015). About capital in the twenty-first century. *Am. Econ. Rev.* 105: 48–53.
- Piketty, T., Saez, E., and Zucman, G. (2018). Distributional national accounts: methods and estimates for the United States. *Q. J. Econ.* 133: 553–609.
- Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Surv. Methodol.* 29: 181–188.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in surveys*. John Wiley & Sons Inc., New York.
- Statistisches Bundesamt (2020). *Preise – Verbraucherpreisindizes für Deutschland*, Available at: <https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/Publikationen>.

Article Note: This article is part of the special issue “Micro-Data from Official Statistics” published in the *Journal of Economics and Statistics*. Access to further articles of this special issue can be obtained at www.degruyter.com/journals/jbnst.

Supplementary Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/jbnst-2022-0015>).