Mattis Beckmannshagen*, Johannes König, Isabella Retter, Christian Schluter, Carsten Schröder and Yogam Tchokni

Dealing with Censored Earnings in Register Data

https://doi.org/10.1515/jbnst-2024-0037
Received March 15, 2024; accepted April 7, 2025

Abstract: Earnings are often top-coded (right-censored) in administrative registers. The censoring threshold in the case of Germany is the limit value for social security contributions, leading to a substantial fraction of censoring: For example, about 12 % of male workers in West Germany are affected, rising to above 30 % for highly educated prime-aged workers. This missing right tail of the earnings distribution constitutes a major problem for researchers studying earnings inequality and top incomes. We overcome this challenge by taking a distributional approach and semi-parametrically modelling the right tail as being Pareto-like. Non-censored earnings survey data matched to administrative records, derived from the SOEP-RV project, let us operate in a laboratory-like setting in which the targets are known. Our approach outperforms alternative imputation methods based on Tobit regressions.

Keywords: right-censored earnings; top-coding; SOEP-RV; heavy-tailed distribution; extreme value index; imputation

JEL Classification: C13; C14; R12

Article note: This article is part of the special issue "Empirical Studies with Micro-Data from the German Pension Insurance" published in the Journal of Economics and Statistics. Access to further articles of this special issue can be obtained at www.degruyter.com/jbnst.

Johannes König and Carsten Schröder gratefully acknowledge financial support by DFG (grant 430972113) and ANR-DFG (grant ANR-19-FRAL-0006-01).

Christian Schluter gratefully acknowledges financial support by ANR (grants ANR-17-EURE-0020 and ANR-19-FRAL-0006-01).

***Corresponding author: Mattis Beckmannshagen**, DIW Berlin, SOEP, Mohrenstr. 58, 10117 Berlin, Germany, E-mail: mbeckmannshagen@diw.de. https://orcid.org/0000-0002-9991-4907
Johannes König, Isabella Retter and Yogam Tchokni, DIW Berlin, SOEP, Mohrenstr. 58, 10117 Berlin, Germany, E-mail: jkoenig@diw.de (J. König), iretter@diw.de (I. Retter), ytchokni@diw.de (Y. Tchokni)
Christian Schluter, Aix Marseille School of Economics, Marseille, France; and University of Southampton, Southampton, UK, E-mail: christian.schluter@univ-amu.fr
Carsten Schröder, DIW Berlin, SOEP, Mohrenstr. 58, 10117 Berlin, Germany; and Freie Universität Berlin, Berlin, Germany, E-mail: cschroeder@diw.de

1 Introduction

Administrative earnings data are increasingly important for empirical research. They offer the promise of unrivalled accuracy and vast sample sizes (even the universe of workers), being produced by government agencies as primary input for the purpose of individual-level tax or benefit calculations. However, given the purpose for which these data are produced, frequently there are strict constraints on their utility. For instance, in Germany, given that social security contributions are made up to an assessment ceiling, earnings in *all* registers are top-coded (right-censored) at this threshold. The resulting incidence of right-censoring is substantial and cannot be ignored (e.g. about 12 % for male workers in West Germany, rising to above 30 % for highly educated prime-aged workers, see e.g. Drechsler and Ludsteck (2025) for an assessment).

This missing right tail of the earnings distribution constitutes a major problem for researchers seeking to study earnings inequality and top incomes. In light of this, we address two natural questions: (1) How can the missing right tail of the earnings distribution be credibly estimated given censored data? (2) How good is the proposed tail estimate?

The usual way to deal with top-censored administrative earnings data is to impute them. However, since the true earnings above the censoring threshold are unknown, the researcher cannot assess the quality of the imputation. Our first principal contribution is to deploy a new unique data source to this end: The SOEP-RV project record-matches administrative (and hence right-censored) earnings data from the German Pension Insurance to uncensored self-reported earnings of SOEP survey respondents. More precisely, the match is made with the individual's social security biography (the individual's Insurance Account), as maintained by the German Pension Insurance, and used by the latter to determine pension entitlements. Since we demonstrate that administrative earnings and survey earnings are closely aligned *below* the censoring threshold, we have a laboratory-style setting where, *above* the censoring threshold, we can compare uncensored survey earnings (labeled below the 'target') with imputed earnings from censored administrative data.

Our second contribution is our imputation method for the missing right tail of the censored administrative earnings distribution. We take a distributional approach, modeling the tail of the earnings distribution as being Pareto-like. More specifically, we assume that, in line with the literature, the tail of the earnings distribution decays like a power function (Emmenegger and Münnich 2023; Jenkins 2017; Schluter and Trede 2019).[1]

1 There is a long-established body of theoretical literature addressing the question of why the top tails of income and wealth distributions are Pareto-like. See, for example, Champernowne (1953), Simon (1955), Kesten (1973), and for a review König et al. (2020)). A more general overview of power laws in economics and finance is provided by Gabaix and Ibragimov (2011) and Gabaix (2016). For

We estimate this power (more precisely, the extreme value index), using a rank-size regression estimator and select the number of upper order statistics entering the computation optimally by minimizing the asymptotic mean-squared error of the estimator. The statistical theory is presented in Schluter (2018) and we demonstrate how the generalization accommodating complex survey weights successfully deals with the censoring issue when we include a point mass at the censoring threshold. Our suite of Stata functions, entitled beyondpareto, makes this procedure publicly available and our companion paper (König et al. 2025) describes in detail its core functionality.[2] We then demonstrate the performance of our estimation approach, first on synthetic data, subsequently on SOEP and SOEP-RV data. In all our examples, our procedure, using only right-censored data, produces an extreme value index estimate and top earnings shares that are close to the target values. Finally, we demonstrate that our estimation approach outperforms alternative imputation procedures, such as the popular Tobit imputation.

The outline of the paper is as follows. Section 2 briefly reviews the rank-size regression estimator, relegating statistical detail to Appendix C, and verifies the performance of the approach using simulations and a parametric earnings model that happens to fit the actual earnings distribution very well (the GB2 model), both for uncensored and artificially censored data. Section 3 not only describes the database and SOEP-RV project, but also verifies the close correlation of survey earnings and administrative earnings below the censoring threshold. Section 4 deploys beyondpareto and shows that our estimates, obtained from censored administrative data, are close to the target values. Furthermore, it compares our distributional approach to the popular Tobit imputation method and shows that, in terms of the distributional tail as well as in terms of top earnings shares, our approach outperforms the latter. As an illustration, Section 5 shows the implications of the censoring and subsequent imputation for gender earnings gaps. Section 6 concludes.

2 Heavy-Tailed Distributions and the Extreme Value Index

Earnings, income, and wealth distributions are heavy-tailed but not exactly Pareto, see e.g. Schluter and Trede (2019) for a discussion of the theory, statistics, and

empirical applications regarding the wealth distribution see Disslbacher et al. (2023); Wildauer and Kapeller (2022) and Karlsson et al. (2024) for the distribution of health expenditure.

2 All our models are estimated using our Stata suite beyondpareto; see the vignette posted at https://christianschluter.github.io/beyondpareto/. The replication kit for the present paper is also available.

empirics. First, we provide a brief summary and illustration of how to model and estimate such tails in the absence of right-censoring. Let Y_{base} denote the lowest value in the Pareto-distributed tail. Consider then a regularly varying cumulative distribution function F of earnings or wealth, so for sufficiently large $y > Y_{\text{base}}$

$$F(y) = 1 - y^{-1/\gamma} l(y), \qquad (\gamma > 0), \tag{1}$$

where l denotes a slowly varying nuisance function that is asymptotically constant ($l(ty)/l(y) = 1$ as $y \to \infty$). This nuisance function captures the fact that the distributional tail is not exactly Pareto but only *eventually* so. $\gamma > 0$ is called the extreme value index and the Pareto or tail index ($\alpha \equiv 1/\gamma$) is its reciprocal. The objective is to estimate the parameter γ.[3]

We use the rank-size regression estimator of the extreme value index, which measures the *ultimate* slope of the Pareto QQ-plot.[4] The challenge is that the Pareto QQ-plot becomes linear only *eventually*, requiring selecting a threshold value for the number of upper order statistics to enter the estimation. For our empirical setting this is evidenced below in Figure 5. Schluter (2018) provides the distributional theory for the rank-size regression estimator in the distributional model (1) and considers an optimal data-dependent threshold choice based on the minimization of the asymptotic mean-squared error (AMSE). Details of the statistical theory are collected in Appendix C. Our estimation command beyondpareto implements this theory and generalizes the approach to accommodate complex survey designs. In particular, it involves the computation of the extreme value index estimator $\widehat{\gamma}$ as a function of the upper order statistics k, considering all observations up to the k-th largest as part of the upper tail. Its output includes the optimally selected upper order statistic k^*, corresponding to the rank of Y_{base}, and the estimator $\widehat{\gamma}$ of k^*. A plot option allows the user to call our command pqqplot and to display a Pareto QQ-plot as on the left in Figure 2. The accompanying command beyondpareto_topshares reports the top income or wealth shares.[5] For details on beyondpareto, see our companion paper (König et al. 2025). Throughout, we use beyondpareto as shorthand for the rank-size regression estimator of the extreme value index with AMSE-minimized threshold selection.

---

3 To help interpret the size of γ, note that the pth raw moment of the distribution exists only if $p < 1/\gamma$. In the applications to German earnings data below, γ assumes values around 0.3, so the first three raw moments are finite. By contrast, Konig, Schluter, and Schroder (2024) show that the German wealth distribution has a much heavier tail and that top wealth is much more concentrated, with γ being estimated at 0.6, thus indicating that the variance of the wealth distribution does not even exist.

4 This Pareto QQ-plot describes the sample analogue of the asymptotic behavior of the log of the tail quantile function U, $\log\ U(x) \sim \gamma \log\ x$ as $x \to \infty$, where $U(x) \equiv F^{-1}(1 - 1/x)$.

5 The command is provided in the replication files.

2.1 In the Lab: a Parametric Earnings Distribution

We illustrate the main procedure (and a proof of concept and performance evidence) using artificial data. So that the model be realistic, we fit the parametric model to monthly earnings data taken from the SOEP for male workers in West Germany in 2018. Specifically, we assume a heavy-tailed GB2 model, the generalized beta distribution of the second kind, with the density function.

$$GB2\,(y\,;a,b,p,q) = \frac{|a|y^{ap-1}}{b^{ap}B\,(p,q)\,(1+\,(y/b)^{a})^{p+q}}, \qquad (y,a,b,p,q>0). \tag{2}$$

The GB2 parameters are estimable by maximum likelihood. It is well-known that the GB2 model is a member of family (1) and that the extreme value index equals $\gamma = 1/(ap)$. See e.g. Schluter and Trede (2024) for an application of the GB2 model. Maximum likelihood then yields the following estimates: $(\widehat{a} = 4.037\,;\widehat{b} = 3633.295\,;\widehat{p} = 0.757\,;\widehat{q} = 0.825)$. Figure 1 reveals that the fitted parametric GB2 model provides a close approximation to the earnings data.

We then use the fitted GB2 model as our data generating process (DGP), with implied $\gamma = 0.327$ and an earnings share for the top 1 % of 4.7 %.

Figure 2 illustrates the estimator of γ for one such random sample using beyondpareto. Panel (A) depicts the Pareto QQ-plot for the upper earnings tail and shows its approximate linearity. We fit a straight line with slope $\widehat{\gamma}$; the vertical

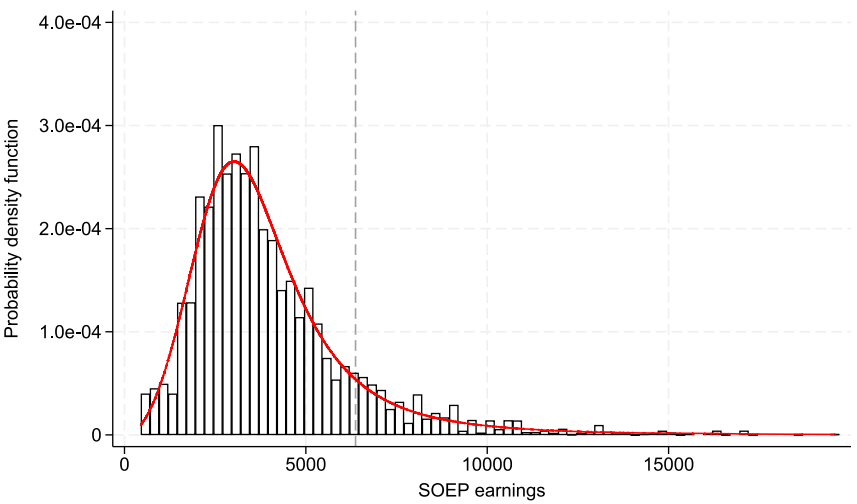

Figure 1: GB2-fit for earnings of West German men in 2018. Displayed is the empirical earnings distribution in the sample of West German men in 2018 (in bars) and the according gb2 fit (red curve). The dashed vertical line represents the earnings assessment ceiling at € 6,370, which would induce a censoring incidence of 11.22 %. The Stata package gb2fit yields the following parameters: $\widehat{a} = 4.037, \widehat{b} = 3633.295, \widehat{p} = 0.757, \widehat{q} = 0.825$.

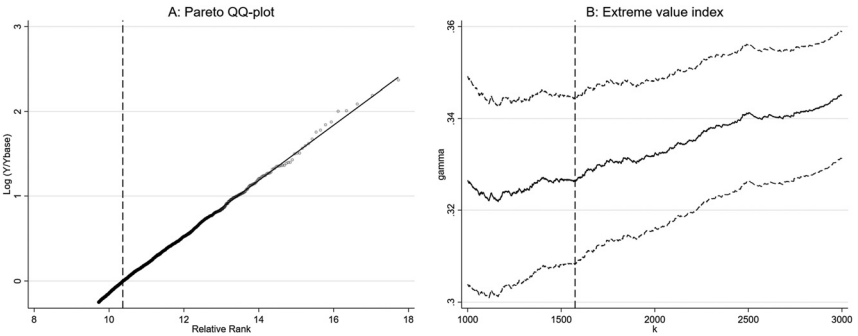

Figure 2: Diagnostic plots of beyondpareto in GB2 example. Panel A depicts the Pareto QQ-plot for one random sample based on the GB2 model. Panel B depicts the corresponding extreme value index plot. The dashed vertical lines represent the optimal Y_{base}.

dashed line depicts the value corresponding to the optimally determined Y_{base}. In panel (B), we depict $\widehat{\gamma}$ as a function of the upper order statistics k, and the optimal k^* corresponds to Y_{base}. Beyond this k^*, the estimator is increasingly distorted.

Is the good fit for this one synthetic distribution just a coincidence? To answer this question, we perform a Monte Carlo study, generating 1,000 synthetic distributions of size 10,000. For each synthetic distribution, we estimate the extreme value index γ. Row 1 of Table 1 displays tail index estimate $\widehat{\gamma}$, the optimally AMSE-based chosen earnings threshold Y_{base}, and the earnings share of the top 1 %, including the standard deviations across all 1,000 samples. In this lab setting, beyondpareto performs very well: the estimated values are very close to the true population values and the empirical standard deviations across simulations suggest tight estimates.

Regarding the top earnings shares, Row 3 of Table 1 reports the empirical sample values that exhibit substantial distortions, illustrating the need for appropriate statistical modelling: The observed downward bias reflects the nature of heavy-tailed distributions, since top earnings in a random sample tend to be under-represented (even, as in our case, for samples of size 10,000).

2.1.1 The Effect of Right-Censoring

Next, we artificially right-censor the earnings data. Since our DGP is based on the actual earnings distribution, we now censor the simulated data at 0.98 times the German social security limit, i.e. € 6,370.[6] This implies a top-censoring incidence of

6 The 2018 censoring threshold for West Germany was at € 6,500. We follow a common practice of scaling it by 0.98 to account for imprecision.

Table 1: GB2 Monte Carlo simulations.

	Tail metrics		Top earnings shares	
	y_{base}	$\widehat{\gamma}$	1 %	At cens.%
Simulation, uncensored	6,369	0.31	4.89	26.34
	(1,138)	(0.02)	(0.21)	(0.65)
Simulation, censored	6,314	0.32	5.30	26.79
	(40)	(0.07)	(1.56)	(2.49)
Empirical sample value			3.95	26.53

The Monte Carlo simulations are based on the GB2 parameters that weobtained when running gb2fit on the SOEP sample for West German men (see Figure 1). The censoring incidence is 11.22 %. The theoretical population values are $\gamma = 0.31$, and top earnings share equal to 4.89 for the 1 %, 26.34 at the censoring threshold ('at cens.%'). We report the empirical means across all 1,000 simulation runs, and in parentheses the empirical standard deviations.

11.22 % (which is close to the cross-sectional censoring incidence in the integrated employment biographies provided by the Federal Employment Agency). In the estimation, we place a mass-point corresponding to the censoring incidence at the censoring point. beyondpareto accommodates complex survey weights and, thus, can take into account the weight of all censored observations. Censoring results in information loss, which inevitably reduces the quality of the estimates. However, Row 2 of Table 1 reveals that the induced distortion is small in the current setting: The estimate $\widehat{\gamma}$ is 0.32 compared to a true value of 0.31. The variability of the estimates is also slightly increased. Compared to a theoretical top 1 % earnings share of 4.89 %, the estimate is now 5.30 %.

In sum, this lab exercise evidences the performance of our approach. The next sections work with actual data. It shows that while the parametric GB2 model is a decent approximation, it is easily outperformed by our estimator for the semi-parametric model given by equation (1).

3 Data

Our empirical analyses rely on two data sources: (a) The German Socio-Economic Panel (SOEP), and (b) the SOEP linked with German social-security register data, SOEP-RV (Forschungsdatenzentrum der Rentenversicherung 2024).

3.1 German Socio-Economic Panel

The German Socio-Economic Panel, SOEP, is one of the largest and longest-running multidisciplinary household surveys worldwide (Goebel et al. 2019). It is a random

draw of all private households in Germany. In every household, all adults are asked to provide, amongst many other items, incomes from all sources. Most importantly for our purposes, it asks for individual gross labor earnings from dependent employment, the determinant of pension contributions and entitlements. Accordingly, the data provide, in contrast to German register data, information on the distribution of earnings below and also *above* the assessment ceiling.

3.2 SOEP-RV

SOEP-RV is based on a record linkage project with the German Pension Insurance (*Deutsche Rentenversicherung Bund*), in which the survey data of SOEP respondents have been linked on a 1:1 basis with their individual social security biographies.

The insurance data have a very broad coverage, as the vast majority of employees in Germany are mandatorily insured. The pension insurance keeps an account for each of these employees – for both the employment phase and retirement phase. As contributions are closely linked with pension entitlements in the German system, contributors' earnings biographies are carefully recorded and contained in individual insurance accounts at the German Pension Insurance. However, social insurance contributions in Germany are capped at an annually adjusted earnings threshold, leading to right-censored (topcoded) earnings in *all* administrative social security datasets.

The SOEP data have been linked with administrative records from the German Pension Insurance, creating the SOEP-RV dataset (Lüthen et al. 2022). For all SOEP respondents who gave explicit consent,[7] SOEP-RV links, on an individual level, the SOEP data with the pension account biographies. The present study is based on the longitudinal dataset, called SOEP-RV.VSKT2020.[8] The longitudinal data comprise exact information on pension-relevant status, such as pension recipient, education, or dependent employment, and in case of the latter the associated earnings points and gross earnings for up to 624 months starting from January in the calendar year in which an individual turns 14 years old (Forschungsdatenzentrum der Rentenversicherung 2024). Throughout, we will use interchangeably 'administrative earnings' and 'Insurance Accounts (IA) earnings.' In the next section, we use SOEP-RV to assess the differences between self-reported earnings in the SOEP and the administrative records *below* the censoring threshold. Their close correspondence

7 The consent rate, taking into account individuals who were asked for consent in 2018 or 2020, is 53.1 %

8 With respect to its structure, the longitudinal data of SOEP-RV resemble the *Versicherungskontenstichprobe* or Insurance Accounts Sample, a dataset provided by the research data center of the German Pension Insurance.

then enables our research design in a lab-style setting: *Above* the censoring threshold, we use uncensored self-reported earnings to assess the performance of imputation techniques for censored administrative earnings.

3.3 Comparability of Earnings in SOEP and Register Data

As with all surveys, respondents' self-reported earnings may be subject to measurement errors, whereas register earnings are, presumably, reliable. Another difference between the two data sources concerns the accounting period: SOEP earnings refer to the month before the interview, while register earnings refer to employment spells based on which monthly earnings are derived. Another issue are one-time payments: Such payments are subject to social security and are part of register earnings – unless these earnings already exceeded the assessment ceiling. In SOEP, one-time payments are contained in a separate variable that refers to the full year and that is surveyed retrospectively. Therefore, for comparability reasons, we divide annual one-time payments by 12 and add it to the reported regular earnings.

To use the SOEP as an external validation tool, SOEP and register earnings must be sufficiently "close," allowing credible comparisons between the top-tail of SOEP earnings and the tail of censored register data. Figure 3 provides an initial descriptive assessment. It is a scatter plot of self-reported SOEP earnings versus administrative IA earnings in the year 2018 for record-linked men in West Germany with IA earnings below the assessment ceiling of € 6,370. The black line is the 45-degree bisector, the red functional form shows the result of a locally weighted regression – and thus a highly flexible form. The figure shows that both earnings variables are highly correlated with a correlation coefficient (ρ) of 0.92. The in-depth analysis in Schröder et al. (2023) supports that the SOEP data are very suitable as comparative statistic for the register data.

3.4 Working Samples Restrictions

The literature on earnings distributions in Germany using administrative data usually focuses on male workers in West Germany (see Section 4.3 below for examples). We follow this practice, but present in Appendix A the analyses for women in West and all workers in East Germany.

We focus on 2018, the year with the largest number of linked observations. The 2018 West German assessment ceiling is € 6,500 per month and we follow common practice of scaling it by 0.98 to account for imprecision in the register earnings. Thus,

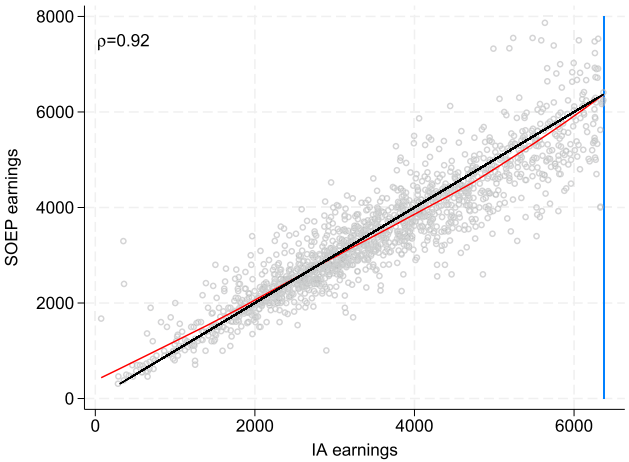

Figure 3: Scatter plot of IA and SOEP earnings *below* the censoring threshold. Linked SOEP-RV data for men in West Germany with IA earnings below the assessment ceiling. The correlation coefficient is $\rho = 0.92$. The black line marks the 45-degree line. The blue vertical line indicates the assessment ceiling. The red line results from a locally weighted regression. Four outliers with self-reported earnings above 8,000 Euro are excluded in the scatter plot but included in the regression. Cases with statistically imputed SOEP earnings are excluded. Furthermore, individuals who work less than half of the month are excluded, as part of these individuals receive wage subsidies that might be included in the SOEP but not in the register data.

the effective censoring threshold is € 6,370 and, based on the weighted SOEP data, the censoring incidence among West German men is 12.2 %.

To construct the SOEP-RV sample, we proceed as follows: As SOEP earnings refer to the pre-interview month, we check the employment status in the register data in that exact month. If the status indicates employment subject to social security in West Germany for at least half of that month, we consider the registered earnings for that month. For the larger SOEP sample, we follow the same restrictions, except for two minor deviations: 1) For the SOEP sample, we cannot rely on the pension status of employment subject to social security in West Germany and, thus, consider all dependent employees earnings above € 450 living in West Germany (excluding Berlin). Thus, instead of the workplace location, the sample restriction is based on residence. 2) Since the exact pension records of days employed in a month are not available for the large SOEP sample, we cannot restrict the sample to individuals who worked for at least half of the reference month. Both of these small differences in sample construction have a negligible – if any – impact for our empirical exercises. Table 2 summarizes these selection restrictions, and reports the resulting sample sizes.

Table 2: Sample restrictions.

	SOEP-RV sample	SOEP sample
Year	2018	2018
Location	Workplace in West Germany	Residence in West Germany
Employment	In employment subject to social security for at least half a month	In dependent emp. at the time of the survey, earnings > € 450
Income concepts	(a) Self-reported (1/12 of yearly OTP added, no imputed values) (b) Register data (incl. OTP)	Self-reported (1/12 of yearly OTP added, no imputed values)
Obs. (total)	4,979	11,302
Men, West	1,920	4,444
Women, West	1,997	4,227
Men, East	514	1,284
Women, East	548	1,347

OTP denotes one-time payments.

4 From the Lab to the Field

We have documented the close correspondence between survey-reported earnings and matched administrative reports in the SOEP-RV *below* the administrative censoring threshold. This suggests that *above* the censoring threshold, the survey reported earnings are good approximations to the true values that are top-coded in administrative data. Thus, we have a unique laboratory-style setting within which to assess the performance of imputation methods for right-censored administrative data.

Based on our distributional approach, we focus on the extreme value index and top earnings shares as performance metrics, computed by beyondpareto, that accommodate the complex survey design involving sampling weights[9]. In particular, the uncensored survey earnings yield the target values (which would correspond to the theoretical population values in a true lab setting) against which the metrics obtained from censored administrative data will be compared.

Our estimation method includes a threshold parameter Y_{base} that is the earnings threshold above which values enter the estimator's computation (i.e. the number of upper order statistics considered). In the presence of administrative censoring, Y_{base}

9 In all our estimations, we apply standard SOEP weighting factors (phrf) to adjust for different sampling probabilities across SOEP subsamples and other selectivities. The purpose is not to obtain "exact" point estimates of top earnings shares, which would require the incorporation of the self-employed and civil servants, whose earnings are not subject to social security and, hence, not contained in the register data. See Bartels and Metzing (2019) for a discussion. For the same reason, we also refrain from correcting SOEP weights for potential selection into the SOEP-RV sample.

will be forced to be below the censoring threshold. For transparency, we consider three settings where we build up restrictions, so that any performance loss can be traced.

1. **Target (unrestricted):** Estimation relies on the complete uncensored survey earnings distribution without any restrictions on Y_{base}.
2. **Restricted:** Estimation relies on the complete uncensored survey earnings distribution, but Y_{base} is restricted to be below the censoring threshold, i.e. the administrative assessment ceiling.
3. **Censored:** Estimation uses earnings right-censored above the assessment ceiling.

Figure 4 illustrates the three settings. The left-hand panel shows the distribution of uncensored SOEP earnings. For the target setting, beyondpareto determines the optimal Y_{base} within the whole range depicted by the shaded areas A and B. In the restricted scenario, beyondpareto will only consider area A, below the censoring threshold. The right-hand panel shows the artificially censored SOEP earnings, including the mass point at the censoring threshold. The area under consideration for beyondpareto in the censored setting is again displayed by the shaded area.

4.1 Top Earnings: the SOEP Sample

We use the SOEP sample as a first benchmark, before considering the smaller SOEP-RV subsample.

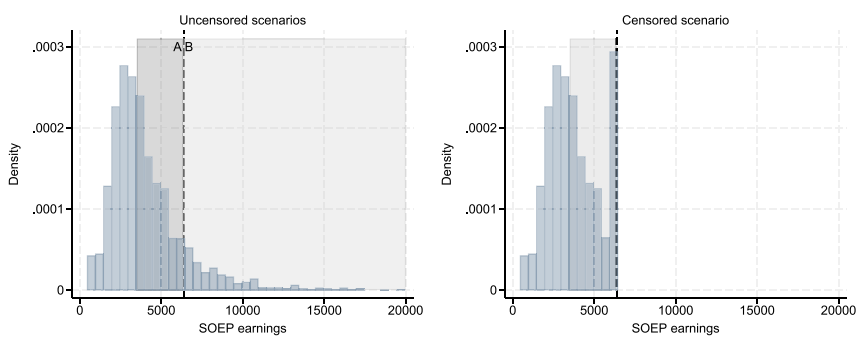

Figure 4: The three scenarios. The left-hand panel displays the complete empirical (uncensored) SOEP earnings distribution (probability density function). In the 'unrestricted' case Y_{base} can fall anywhere in the shaded areas A and B. In the 'restricted' case Y_{base} is constrained to fall into area A. The right-hand panel displays the SOEP earnings distribution censored at the administrative contribution ceiling; hence, the 'censored' setting. *Source:* SOEPv38.1.

Table 3 summarizes the results for the SOEP sample. The target value of the extreme value index based on the uncensored earnings data and an optimal choice of Y_{base} (Row 1) is $\widehat{\gamma} = 0.24$. The implied estimate of the top 1 % earnings share is 4.01 % and, for those above the administrative censoring limit, 26.57 %. These results also substantiate our claim that the parametric GB2 approach (see Row 1 of Table 1) is outperformed. The optimally chosen Y_{base} = € 7,992 exceeds the assessment ceiling of € 6,370. Figure 5 depicts the Pareto QQ-plot of the upper order statistics of these uncensored SOEP earnings and the fitted line with slope $\widehat{\gamma} = 0.24$ at the optimal Y_{base} = 7, 992 choice. Panel (b) depicts $\widehat{\gamma}$ as a function of the upper order statistics k, while the optimal k^* corresponds to Y_{base}. Row 2 of Table 3 shows the effect of restricting Y_{base} to be below censoring ceiling, leading to a restricted Y_{base} of 6,200. Consequently, $\widehat{\gamma}$ increases slightly to 0.26 and the top 1 % earnings share increases marginally to 4.16 %. This modest increase in $\widehat{\gamma}$ is further illustrated in panel (b) of Figure 5, where the restricted choice of k associated with Y_{base} = 6,200 is identified by the solid grey vertical line. Finally, in Row 3, we artificially censor the SOEP earnings data at the assessment ceiling and incorporate the resulting mass point in our estimation. Thanks to the weighting procedure of our estimator, the estimate $\widehat{\gamma}$ is slightly above the target value of 0.24. However, since Y_{base} is smaller, the implied earnings share of the top 1 % is now 4.21 %.

Thus, we conclude that censoring of the data at the assessment ceiling does not prevent us from achieving a very good approximation of the tail of the earnings distribution.

Finally, we verify that the results of our analysis of top earnings in the focal year of 2018 are qualitatively representative for other recent years. To this end, Table 4 summarizes the results for the years 2014–2017 and 2019. For all years, the estimates of γ using the censored data get close to the target values, slightly overestimating the latter and implying a slight overestimation of the 1 % and 5 % top earnings shares. Overall, as regards the evolution of top earnings, the 2014-19 period is a period without major changes or trends: measured earnings concentration fluctuates without direction within a narrow band.[10]

4.2 Top Earnings: The SOEP-RV Sample

We now consider the smaller SOEP-RV sample. Table 5 reports the results. Row 1 considers again the uncensored SOEP survey earnings to establish the target values. The extreme value index estimate is $\widehat{\gamma} = 0.27$, which slightly exceeds the value for the

10 This is in line with other assessments of earnings inequality in Germany in the 2010s (Beckmannshagen and Schröder 2022).

Table 3: Tail estimation in the SOEP sample.

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted)	7,992	0.24	4.01	13.71	23.13	26.57
Restricted	6,200	0.26	4.16	13.59	22.63	26.19
Censored	6,239	0.27	4.21	13.70	22.78	26.35

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to the share of all earnings at the assessment ceiling. All estimates obtained by beyondpareto. *Source:* SOEPv38.1.

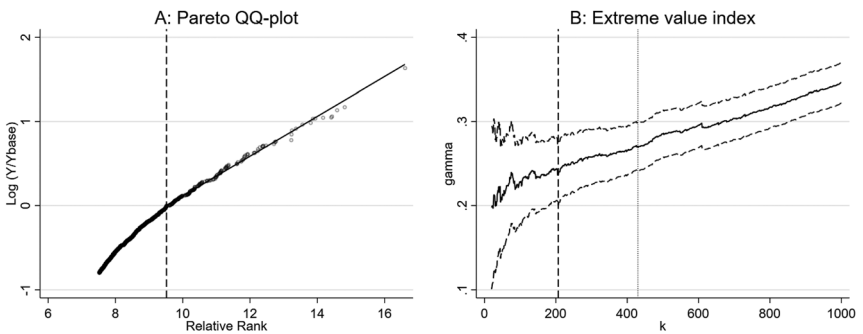

Figure 5: Diagnostic plots of beyondpareto in the SOEP sample. Panel A depicts the Pareto QQ-plot in the SOEP sample of West German men. Panel B depicts the corresponding extreme value index plot. The dashed vertical lines represent the optimal Y_{base}. The dotted line (panel B) represents the optimal Y_{base} in the restricted setting.

complete SOEP sample. Hence, potentially selective consent into the record-linkage has little effect. The earnings share of the top 1 % is 4.27 %. Y_{base} is at € 6,200 and thus below the administrative censoring threshold. As a consequence, Row 2 where Y_{base} is forced to be below the administrative censoring threshold equals Row 1. Row 3 censors the reported survey earnings at the social security assessment ceiling. In line with the results of Table 3 (and the simulation setting of Table 1), our estimation approach accommodates the censoring very well, with only minor effects on $\widehat{\gamma}$ or the top earnings share. Specifically, $\widehat{\gamma} = 0.26$. The earnings share of the top 1 % is somewhat reduced to 4.10 %. Finally, in Row 4 we consider the matched administrative but right-censored earnings reports, i.e. the usual setting a researcher using administrative data would confront. The table reveals that our estimation approach gets close to the target value of $\widehat{\gamma} = 0.27$! The estimated top earnings shares are reduced by small amount.

Table 4: Tail estimation in the SOEP sample, other years.

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
2014						
Target (unrestricted)	6,655	0.26	4.18	13.88	23.24	28.37
Restricted	5,700	0.27	4.28	13.84	22.94	28.02
Censored	5,392	0.32	5.14	15.31	24.50	29.52
2015						
Target (unrestricted)	7,500	0.22	3.78	13.36	22.89	29.17
Restricted	5,800	0.25	3.96	13.31	22.42	28.68
Censored	5,700	0.33	5.34	15.73	25.05	31.20
2016						
Target (unrestricted)	6,456	0.28	4.49	14.34	23.64	27.79
Restricted	5,917	0.29	4.56	14.35	23.50	27.94
Censored	6,008	0.32	5.16	15.35	24.56	28.96
2017						
Target (unrestricted)	7,333	0.23	3.93	13.52	22.96	27.54
Restricted	6,139	0.26	4.09	13.55	22.70	27.24
Censored	6,190	0.25	4.02	13.45	22.60	27.16
2019						
Target (unrestricted)	6,483	0.29	4.66	14.52	23.68	27.30
Restricted	5,800	0.30	4.70	14.48	23.50	26.43
Censored	6,400	0.31	4.98	15.05	24.22	27.18

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to the share of all earnings at the assessment ceiling. All estimates obtained by beyondpareto. See Figure A.1 for a graphical illustration. *Source:* SOEPv38.1.

We conclude that our estimation approach enables us to provide a credible estimate for the right tail of the earnings distribution based on censored administrative earnings. Figure 6 visualizes the setting and the results: The histogram (brown bars) depicts the earnings density below the administrative censoring ceiling alongside the mass point of the latter (transparent brown bar). We then smoothly paste a Pareto tail (blue line) based on our estimate of $\widehat{\gamma} = 0.24$.

4.3 Tobit Imputations

A popular way of dealing with top-censored earnings in German administrative data is to impute based on individual-level Tobit regressions. In the regressions' predictions, practitioners generally add an error term drawn from a group-specific (log) normal distribution (see e.g. Dustmann et al. (2009) and Card et al. (2013) in case of the well-known Integrated Labour Market Biographies (SIAB), provided by the Institute

Table 5: Tail estimation in the SOEP-RV sample.

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted, SOEP earn.)	6,200	0.27	4.27	13.74	22.73	26.03
Restricted (SOEP earn.)	6,200	0.27	4.27	13.74	22.73	26.03
Censored (SOEP earn.)	6,200	0.26	4.10	13.43	22.39	25.70
Censored (IA earn.)	6,266	0.24	3.81	12.86	21.72	26.13

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to the share of all earnings at the assessment ceiling. All estimates obtained by beyondpareto. *Source:* SOEPv38.1, SOEP-RV.VSKT.2020-v2.

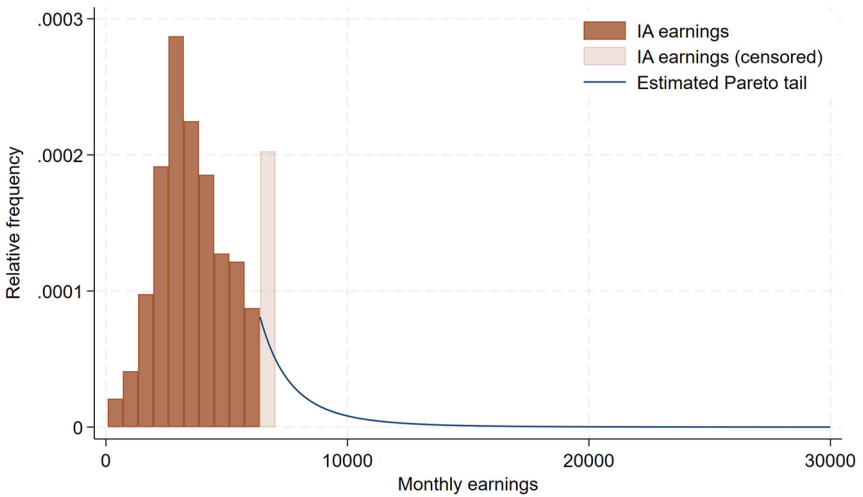

Figure 6: IA earnings distribution: censored and fitted tail. Displayed is the empirical distribution of IA earnings in the SOEP-RV sample (probability density function), with the mass point at the assessment ceiling (transparent), and the fitted tail. *Source:* SOEP-RV.VSKT.2020-v2.

for Employment Research in Germany (IAB)). Dauth and Eppelsheimer (2020) provide the *de facto* standard data wrangling code for IAB data, including the Tobit-based imputation procedure. Drechsler and Ludsteck (2025), assessing the performance of these imputations, observe a sharp-discontinuity of the earnings density at the censoring threshold for prime-aged university educated male workers (for whom the censoring incidence is above 30 %) and offer several strategies how to alleviate the deficiencies of Tobit-based imputations.

For the implementation of the Tobit procedure, we stick to a more simple approach and follow Dauth and Eppelsheimer (2020) as closely as our SOEP-RV data

allow us to.[11] For our experiments, we consider the setting of Table 5 and focus on the top earnings shares. The Tobit model implies a tail decay that is faster than the power function of equation (1), so γ in this case equals the limit value of 0 (see e.g. Schluter and Trede (2019) for a discussion and tests of power function behavior against subexponentiality).

Figure 7 compares the resulting empirical distribution function (EDF) for Tobit-imputed earnings to the target EDF of uncensored SOEP earnings. The vertical line marks the censoring threshold. Also included is the cumulative distribution function (CDF) based on beyondpareto's extreme value estimate. While the Tobit imputation does produce an approximation for the missing right tail of the censored earnings distribution that captures its heaviness fairly well, our CDF is always closer to the target EDF. The Tobit imputation systematically assigns earnings that are too large.

This is further illustrated in Figure 8, which compares the implied top earnings shares with the target shares based on the SOEP sample (Table 3, row 1). For maximum comparability, we imputeartificially censored SOEP earnings, to prevent deviations between target and imputed shares due to differing earnings concepts. For the vast majority of top earnings shares, beyondpareto's estimates are closer to the target than the empirical shares of the Tobit imputation. For example, the target earnings share of the top 5 % (10 %) is 13.71 % (23.12 %). The estimated share based on beyondpareto, is very close at 13.70 % (22.78 %), while Tobit imputation yields slightly higher shares of 15.03 % (24.52 %). Generally, the beyondpareto procedure slightly under-predicts, while the Tobit imputation over-predicts. Only in the extreme part of the tail do all three shares become, by necessity, very close.

5 Implications for Calculating Gender Earnings Gaps

What are the implications of the data censoring and subsequent Pareto imputation for applied empirical analyses? This is demonstrated below by examining gender

11 The main difference stems from the absence of firm identifiers in the SOEP-RV data. Instead of tenure with a specific firm, we include labor market experience, which we define as the sum of experience in full-time employment and 0.5 times experience in part-time employment. Instead of the imputed education variable in the integrated labor market biographies, we rely on a more fine-grained surveyed education variable that comprises four categories of the highest obtained educational degree: neither *Abitur* nor vocational training; *Abitur* and/or vocational training; Bachelor's degree; Master's degree. As SOEP-RV has no firm identifiers, we cannot compute leave-one-out means of log wages. However, the evidence presented in Dauth and Eppelsheimer (2020) (their Figure 1), the implications of implementing this step seem negligible for the estimated tail of the wage distribution.

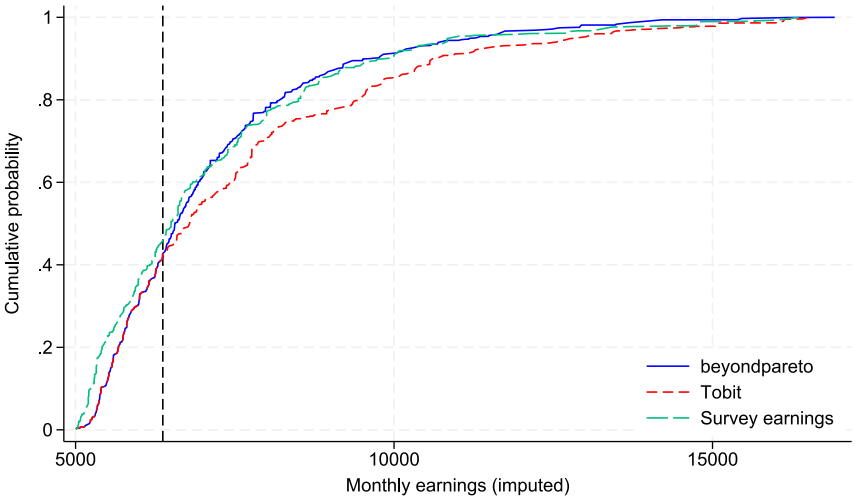

Figure 7: IA earnings distribution: censored and fitted tail. Displayed are the distribution functions based on beyondpareto, Tobit estimations and the empirical survey earnings in the SOEP-RV sample. *Source:* SOEP-RV.VSKT.2020-v2.

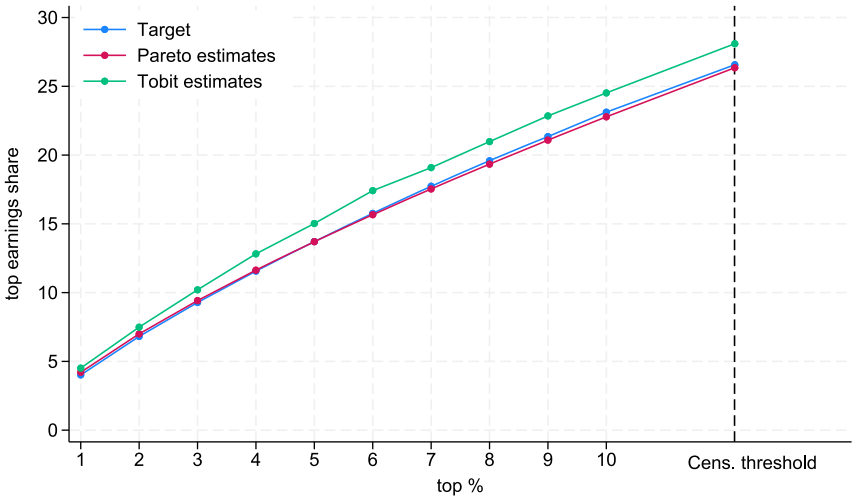

Figure 8: Comparison of top earnings shares based on different imputation methods. Displayed are the cumulative earnings shares of top percentiles as well as the share of all earnings above the censoring threshold in the target estimates (see Table 3, row 1), based on the beyondpareto procedure and based on Tobit estimations. For the years 2017 and 2019, the Pareto estimates are even closer to the target, see Figures A.2 and A.3 in the Appendix. *Source:* SOEPv38.1.

earnings gaps both for all employees and disaggregated by educational attainment.

The findings, derived from the SOEP sample for West Germany 2018, are presented in Table 6. Male (female) employees comprise approximately 54 (46) % of the total workforce. About 30 (27) % are male (female) employees holding a low, and approximately 24 (20) % holding a high educational degree (Row 1). The incidence of earnings above the ceiling exhibits a marked gender disparity (Row 2): while around 12 % of male employees exceed the assessment ceiling, only about two % of female employees experience the same. Furthermore, the incidence is considerably higher among employees with higher education in both groups, with 23 % of men and under five % of women affected.

When earnings exceeding the contribution assessment ceiling are set to this ceiling (as in the raw released social security data), male employees earn approximately € 3,743, approximately 33.37 % more than their female counterparts, who earn around € 2,500 (Row 3). In the low-education group, this relative gap is 34.32 %, whereas in the high-education group, the gap is 31.42 %. Following the application of Pareto imputation of the pseudo-censored earnings (Row 4), the relative gender earnings gap increases to 37.40 %; for employees with lower education, the gap rises to 35.08 %, and for those with higher qualifications, it reaches 39.50 %. These adjusted values (Row 5) closely align with the empirical figures obtained by removing the pseudo-censoring of the SOEP earnings data.

In summary, the illustration demonstrates, that the incidence of censoring, is quantitatively relevant, and that it diverges substantially across different groups of employees. Additionally, the illustration highlights the importance of appropriate imputation in accurately quantifying the true earnings differences between these groups.

Table 6: Implications of top-tail imputation for earnings by gender and education.

	Men			Women		
	All	**Low ed.**	**High ed.**	**All**	**Low ed.**	**High ed.**
% population	53.96	29.75	24.21	46.04	26.53	19.51
% above ass. ceiling	12.34	3.36	23.38	2.03	0.25	4.44
Avg. earn. censored	3,743	3,214	4,395	2,494	2,111	3,014
Avg. earn. imputed	4,029	3,255	5,088	2,522	2,113	3,078
Avg. earn. observed	4,038	3,296	4,949	2,532	2,115	3,100

Row 1 contains the gender- and education-specific shares of employees. Row 2 gives the subgroup-specific shares of earnings (as reported in SOEP) exceeding the assessment ceiling. The next three rows provide group-specific average earnings for three scenarios: (a) when earnings above the assessment ceiling are set to the level of the ceiling (pseudo-censoring); (b) when earnings above the ceiling are Pareto-imputed group-specific with beyondpareto; (c) when observed SOEP earnings are not pseudo-censored (and, hence, earnings above the ceiling are observed). All results are presented separately for men and women overall, as well as low and high educated men and women. Low educated comprises individuals who obtained secondary education or less (ISCED 2011 categories 0–3), while high educated comprises post-secondary and university education (ISCED 2011 categories 4–8). Note that we ran beyondpareto group-specific (all men/women, low educated men/women, high educated men/women), which is why higher (lower) average earnings among low educated and high educated not necessarily add up to higher (lower) average earnings among all men or women. *Source:* SOEPv38.1, weighted with SOEP weighting factors.

6 Conclusions

Scientific interest in register data is immense, due to its unrivaled accuracy for earnings data and its large sample sizes (or even data for the complete population). Examples of such data in Germany are the frequently used social security data, provided by the Federal Employment Agency and the Federal Pension Insurance. In these and many other register data worldwide, however, a relevant part of earnings and wealth is top-coded (right-censored), which, depending on the research question, can severely undermine its usefulness. For instance, how can inequality and top earnings be credibly studied if the right tail of the earnings distribution is missing?

Taking a distributional approach that is based on the semi-parametric modelling of the right tail being Pareto-like, we show how the missing tail can be successfully estimated using the administrative censored data. Our validation of this approach exploits a unique feature of the SOEP-RV project, in which we can compare the estimated or imputed tail to estimates based on uncensored survey data. Such validation is normally not feasible in research that exclusively relies on censored administrative data. Our methods are made available as a suite of functions entitled beyondpareto; these are described in detail in our companion paper (König et al. 2025).

The presented distributional approach should not be confused with individual-level imputations of censored data. The latter requires, in addition to the Pareto parameter, the assignment of a rank of to each censored observation. Due to the censoring, these ranks are unobserved. Bönke et al. (2015), for example, suggest to assign ranks as a function of years that an individual earned above the assessment ceiling and their last observed uncensored earnings. For researchers who are interested in individual-level imputations, the calculation of individual-level earnings after running beyondpareto is technically straightforward given a rank assignment: For rank j, with $j = 1$ being the highest upper-order statistic, the imputed earnings $\widetilde{Y}_j$ is given by:

$$\widetilde{Y}_j = Y_{\text{base}} \left(\frac{k+1}{j} \right)^{\widehat{\gamma}}, \tag{3}$$

where Y_{base} is the value of $Y_{n-k,n}$, $k+1$ the rank associated with Y_{base}, and $\widehat{\gamma}$ the estimated tail parameter (see Appendix C). As seen in equation (3), the prior assigned ranks are crucial determinants of the individual-level imputed earnings. Generally, these ranks must be assigned by the researcher and may be based on assumptions that take into account the specific circumstances of the research question. Depending on the used data source, there may be no useful information for rank assignment of the censored observations available. In this case, the calculation of top shares within the estimated tail presents the preferable option to assess the distribution of high earnings (see Appendix C).

Acknowledgments: We would like to thank Adam Lederer for his thorough linguistic editing of the manuscript. We would also like to thank the colleagues at the Research Data Center of the German Pension Insurance for their technical support in merging and preparing the data and their helpful advice on working with the pension insurance data. We also thank the editor, Peter Winker for the efficient handling of the submission, and two anonymous referees for their constructive feedback.

(Web) Appendix

A Empirical Evidence for Alternative Samples

The following appendix contains analyses analogous to those presented in Table 3 and 5 for the samples of West German women (Tables A.1 and A.4), East German men (Tables A.2 and A.5) and East German women (Tables A.3 and A.6). Note that the presented results differ from our main results due to one major difference. Compared to West German men, in all of the alternative samples, the distribution of earnings is shifted to the left. Accordingly, also the censoring incidence is much lower (based on the weighted SOEP sample: West German women: 3.2 %; East German women: 1.0 %; East German men: 7.2 %). Put differently, there are only few observations with high earnings in these alternative samples. In these cases, even in the unrestricted scenario, beyondpareto yields target values that lie below the respective censoring threshold. Thus, the "restriction" that we introduce in our second analysis scenario has no effect in the analyses based on these samples.[12]

SOEP sample (corresponding to Table 3)

Table A.1: Tail estimation in the SOEP sample (women, west).

	Tail metrics		Top earnings shares			
	Y_{base}	$\hat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted)	4,785	0.22	3.81	13.38	22.92	6.57
Restricted	4,785	0.22	3.81	13.38	22.92	6.57
Censored	5,492	0.18	3.47	12.97	22.53	6.15

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to the share of all earnings at the assessment ceiling. All estimates obtained by beyondpareto. *Source:* SOEPv38.1.

12 One exception is Table A.3. Here, Y_{base} in the target scenario (€ 4,883 corresponds to the 20th ranked earnings. In the restricted scenario, we force the estimation to contain at least 20 cases below the censoring threshold. The target Y_{base} is thus not included in the range of beyondpareto.

Table A.2: Tail estimation in the SOEP sample (men, east).

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted)	3,383	0.32	4.85	14.56	23.38	17.04
Restricted	3,383	0.32	4.85	14.56	23.38	17.04
Censored	5,300	0.37	5.65	15.64	22.45	18.09

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to the share of all earnings at the assessment ceiling. All estimates obtained by `beyondpareto`. *Source:* SOEPv38.1.

Table A.3: Tail estimation in the SOEP sample (women, east).

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted)	2,956	0.26	3.82	12.66	21.21	6.70
Restricted	2,956	0.26	3.82	12.66	21.21	6.70
Censored	4,975	0.21	3.44	12.24	20.65	6.26

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to the share of all earnings at the assessment ceiling. All estimates obtained by `beyondpareto`. *Source:* SOEPv38.1.

SOEP-RV sample (corresponding to Table 5)

Table A.4: Tail estimation in the SOEP-RV sample (women, west).

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted, SOEP earn.)	3,225	0.25	3.83	12.83	21.58	5.43
Restricted (SOEP earn.)	3,225	0.25	3.83	12.83	21.58	5.43
Censored (SOEP earn.)	4,950	0.22	3.67	12.95	21.99	5.28
Censored (IA earn.)	3,744	0.28	4.42	14.01	23.04	7.65

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to the share of all earnings at the assessment ceiling. All estimates obtained by `beyondpareto`. *Source:* SOEPv38.1, SOEP-RV.VSKT.2020-v2.

Table A.5: Tail estimation in the SOEP-RV sample (men, east).

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted, SOEP earn.)	3,197	0.29	4.38	13.67	22.31	10.81
Restricted (SOEP earn.)	3,197	0.29	4.38	13.67	22.31	10.81
Censored (SOEP earn.)	3,125	0.31	4.73	14.28	22.99	11.37
Censored (IA earn.)	4,990	0.19	3.39	12.53	22.54	12.69

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to theshare of all earnings at the assessment ceiling. All estimates obtained by `beyondpareto`. *Source:* SOEPv38.1, SOEP-RV.VSKT.2020-v2.

Table A.6: Tail estimation in the SOEP-RV sample (women, east).

	Tail metrics		Top earnings shares			
	Y_{base}	$\widehat{\gamma}$	1 %	5 %	10 %	At cens.%
Target (unrestricted, SOEP earn.)	4,883	0.12	2.77	11.21	20.02	1.80
Restricted (SOEP earn.)	3,167	0.18	2.99	11.10	19.55	2.00
Censored (SOEP earn.)	4,679	0.14	2.83	11.22	20.03	1.85
Censored (IA earn.)	3,598	0.21	3.44	12.17	20.99	3.27

1 %, 5 %, 10 % refers to the respective earnings share, while 'at cens.%' refers to theshare of all earnings at the assessment ceiling. All estimates obtained by `beyondpareto`. *Source:* SOEPv38.1, SOEP-RV.VSKT.2020-v2.

B Empirical Evidence for Alternative Time Periods

The evolution of top earnings shares based on beyondpareto (corresponding to Table 3 and 4) (Figures A.1–A.3)

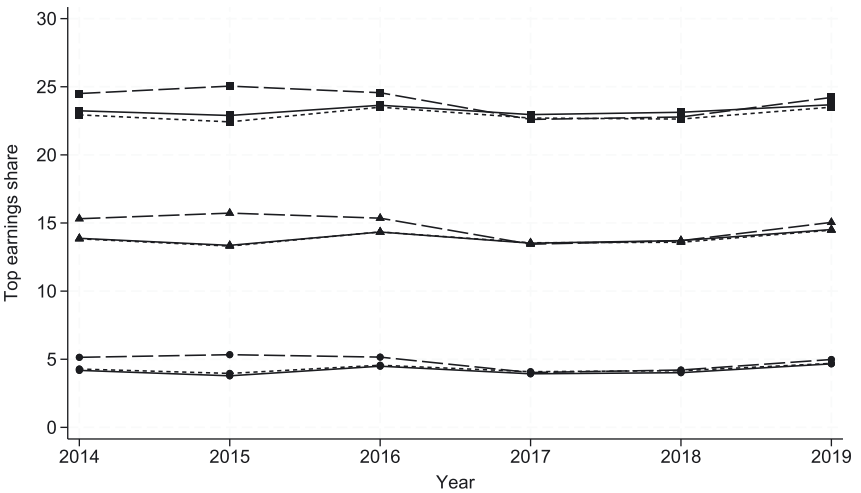

Figure A.1: The evolution of top earnings shares. Displayed are the earnings shares of top percentiles in the target estimates, in the restricted scenario, and in the censored scenario. Squares (triangles) [dots] refer to the earnings share of the top 10 (5) [1] percent of earnings. Solid (short-dashed) [long-dashed] lines present the estimates in the target (restricted) [censored] scenario. Displayed shares correspond to those in Table 3 and 4. *Source:* SOEPv38.1.

Comparison of imputation methods in selected years (corresponding to Figure 8)

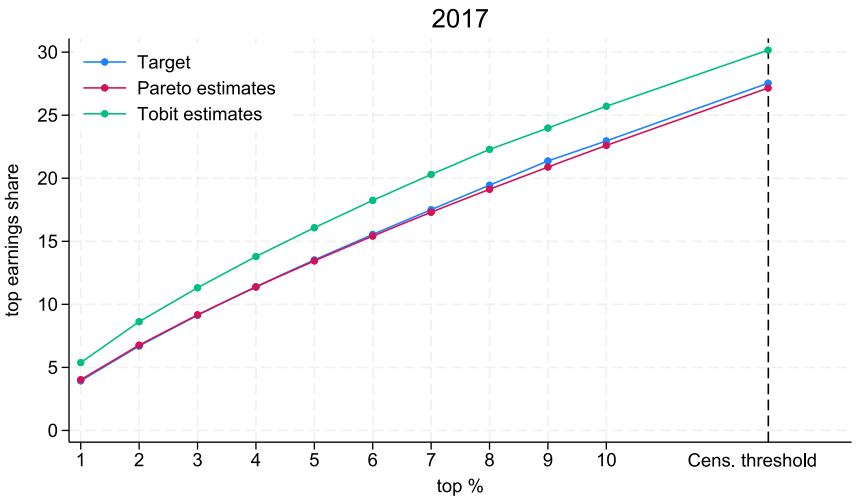

Figure A.2: Comparison of top earnings shares based on different imputation methods, year 2017. Displayed are the cumulative earnings shares of top percentiles as well as the share of all earnings above the censoring threshold in the target estimates (see Table 3, row 1), based on the beyondpareto procedure and based on Tobit estimations. *Source:* SOEPv38.1.

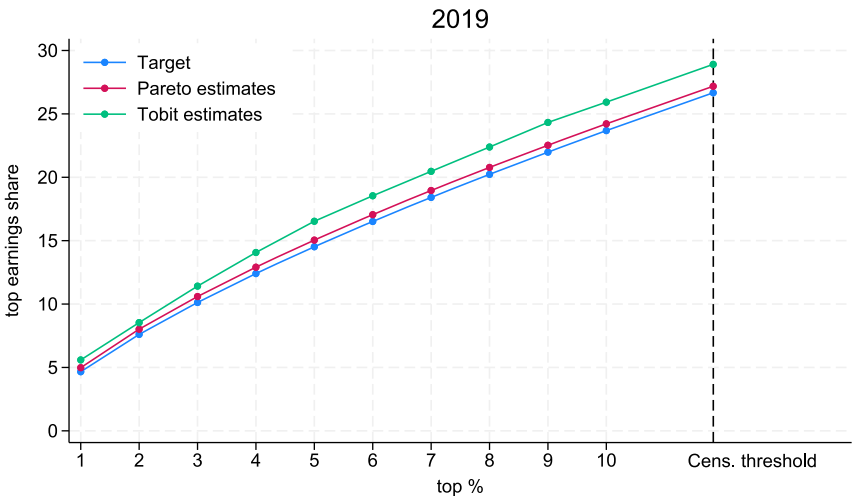

Figure A.3: Comparison of top earnings shares based on different imputation methods, year 2019. Displayed are the cumulative earnings shares of top percentiles as well as the share of all earnings above the censoring threshold in the target estimates (see Table 3, row 1), based on the beyondpareto procedure and based on Tobit estimations. *Source:* SOEPv38.1.

C Statistical Theory

Consider a regularly varying cumulative distribution function F, so for sufficiently large y and $\gamma > 0$

$$F(y) = 1 - l(y) \cdot y^{-1/\gamma},$$

where l denotes a slowly varying nuisance function that is asymptotically constant ($l(ty)/l(y) = 1$ as $y \to \infty$). $\gamma > 0$ is called the extreme value index, and the Pareto or tail index ($\alpha \equiv 1/\gamma$) is its reciprocal. The objective is to estimate the parameter γ.

C.1 The Rank-Size Regression and the Pareto QQ-Plot

The rank-size regression estimator of the extreme value index measures the *ultimate* slope of the Pareto QQ-plot. This follows since the tail quantile function for above model is

$$U(y) = \inf\{t : \Pr(Y > t) = 1/y\} = y^{\gamma} \cdot \tilde{l}(y)$$

where $\tilde{l}(y)$ is another slowly varying function, which then implies $\log U(y) \sim \gamma \log(y)$ as $y \to \infty$. Replacing these population quantities with their empirical counterparts gives the Pareto QQ-plot, and γ is its ultimate slope.

If the tail of the distribution were strictly Pareto, then the Pareto QQ-plot would be linear and a linear regression would estimate its slope coefficient. In the above model, it will become linear only *eventually*, and a slow decay of the nuisance functions $l(y)$ and $\tilde{l}(y)$ will then induce asymptotic distortions in the estimator of the slope coefficient. Below, such slow convergence will be considered in the form of second-order regular variation.

Let $Y_{1,n} \leq \ldots \leq Y_{n,n}$ denote the order statistics of the given sample $Y_1, \ldots, Y_n$ of, for example, wealth or earnings, and consider the k upper order statistics. The Pareto quantile plot (QQ-plot) has coordinates

$$(-\log(j/(n+1)), \log Y_{n-j+1,n})_{j=1,\ldots,k}$$

where the relative rank is given by $-\log(j/(n+1))$ and $j = 1$ for the highest upper-order statistic.

The OLS estimator of the slope parameter in the Pareto QQ-plot is obtained by minimizing the square sum

$$\sum_{j=1}^{k} \left(\log \frac{Y_{n-j+1,n}}{Y_{n-k,n}} - \gamma \log \frac{k+1}{j} \right)^2 \qquad (1 \leq j \leq k < n)$$

with respect to γ, which corresponds to a regression of log sizes on the log of relative ranks for sufficiently large values given by $Y_{n-k,n}$, also denoted by Y_{base}. Note that $\frac{Y_{n-j+1,n}}{Y_{n-k,n}}$ is a normalized size equal to one at the threshold. The resulting OLS estimator is

$$\widehat{\gamma} = \frac{\frac{1}{k}\sum_{j=1}^{k} \log\left(\frac{k+1}{j}\right)\left[\log Y_{n-j+1,n} - \log Y_{n-k,n}\right]}{\frac{1}{k}\sum_{j=1}^{k}\left[\log\frac{k+1}{j}\right]^2}. \tag{4}$$

To compute the estimated values $\widetilde{Y}_{n-j+1,n}$ for rank j, $j = 1, \ldots, k$ as vertical projections on the line estimate, we note that in the Pareto QQ-plot, for the estimated slope holds

$$\widehat{\gamma} = \frac{\Delta y}{\Delta x}, \tag{5}$$

where

$$\Delta x = -\log\left(j/(n+1)\right) - \left(-\log\left((k+1)/(n+1)\right)\right) = \log\left((k+1)/j\right)$$

and

$$\Delta y = \log\widetilde{Y}_{n-j+1,n} - \log Y_{n-k,n}.$$

Solving eq. (5) for $\widetilde{Y}_{n-j+1,n}$, we obtain

$$\widetilde{Y}_{n-j+1,n} = Y_{n-k,n}\left(\frac{k+1}{j}\right)^{\gamma}. \tag{6}$$

In the context of censored data, this expression can be used to impute earnings above a censoring threshold on the individual level, given an assignment of ranks.

C.2 Distributional Theory

The distributional theory for $\widehat{\gamma}$ requires imposing more structure on the behavior of nuisance functions. Schluter (2018) demonstrates, using a framework of second-order regular variation, that as $k \to \infty$ and $k/n \to 0$, this estimator is weakly consistent, and if $\sqrt{k}A(n/k) \to 0$

$$\sqrt{k}(\widehat{\gamma} - \gamma) \to^{d} N\left(0, \frac{5}{4}\gamma^2\right).$$

where $A(t)$ is a rate function that is regularly varying with index ρ, with $A(t) \to 0$ as $t \to \infty$.

Asymptotically, the estimator is thus unbiased if $\sqrt{k}A(n/k) \to 0$. However, if this decay is slow, the estimator will suffer from a higher order distortion in finite samples given by

$$b_{k,n} \equiv \frac{1}{2}\frac{\gamma}{\rho}\frac{2-\rho}{(1-\rho)^2}A(n/k) \qquad (\gamma > 0, \rho < 0).$$

C.3 The Choice of the Threshold k for the Upper Order Statistics

Any tail index estimator requires a choice of how many upper order statistics, given by k, should be taken into account. This choice invariably introduces a trade-off between bias and precision of the estimator that is typically ignored by practitioners. However, this mean-variance trade-off suggests that it is unwise to set the threshold level mechanically (e.g., a wealth level of 1 million euros or 10 % of the sample). By contrast, we determine this threshold level in a data-dependent manner by using the residuals in the rank-size regression in order to estimate non-parametrically the asymptotic mean-squared error (AMSE).

Following Beirlant, Vynckier, and Teugels (1996) and Schluter (2018, 2021), we observe that the expectation of the mean-weighted theoretical squared deviation,

$$\frac{1}{k}\sum_{j=1}^{k} w_{j,k} E\left(\log\left(\frac{Y_{n-j+1,n}}{Y_{n-k,n}}\right) - \gamma \log\left(\frac{k+1}{j}\right)\right)^2, \tag{7}$$

equals, to first order, $c_k \mathrm{Var}(\widehat{\gamma}) + d_k(\rho) b_{k,n}^2$ for some coefficients c_k depending only on k, and $d_k(\rho)$ depending on k and $\rho < 0$. For an explicit statement of the coefficients c_k and d_k, see Schluter (2018).

The procedure then consists of applying two different weighting schemes $w_{j,k}^{(i)}$ (i = 1, 2), estimating the corresponding two mean weighted theoretical deviations using the residuals of rank-size regression and computing a linear combination thereof such that $\mathrm{Var}(\widehat{\gamma}) + b_{k,n}^2$ is obtained. We proceed in this manner for weights $w_{j,k}^{(1)} \equiv 1$ and $w_{j,k}^{(2)} = j/(k+1)$ for a set of pre-selected values of ρ.

In particular, based on the experiments reported in Schluter (2018, 2021), we set a very conservative value of $\rho = -0.5$ (implying a slow decay of the slowly varying nuisance function l).

C.4 Complex Surveys

Survey data often come with sampling weights to allow inference on the population level. The aforementioned theory and methods are easily adapted to this setting if we define the weighted empirical distribution function as

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} w_i 1(Y_i \le y),$$

where w_i is the sampling weight associated with the i's observation Y_i with $\sum_{i=1}^{n} w_i = n$. Examples are a scheme of unity weights ($w_i \equiv 1$ for all i), or $w_i = \widetilde{w}_i n$ with $\widetilde{w}_i < 1$ and $\sum_i \widetilde{w}_i = 1$. Then, for the j's largest observation, we have $F_n(Y_{n-(j-1),n}) = \frac{n-\sum_{i=1}^{j} w_{(i\le j)}}{n}$ with the implicit notation convention that $\sum_{i=1}^{j} w_{(i\le j)}$ denotes the summation of the survey weights corresponding to the j largest upper-order statistics. The resulting Pareto QQ-plot has coordinates

$$\left(-\log\left(\sum_{i=1}^{j} w_{(i\le j)} \Big/ (n+1)\right), \log Y_{n-j+1,n}\right)_{j=1,} \ldots, k,$$

and the resulting survey-weights-adjusted estimator of γ then becomes

$$\widehat{\gamma} = \frac{\frac{1}{k}\sum_{j=1}^{k} \log\left(\frac{\sum_{i=1}^{k+1} w_{(i\le k+1)}}{\sum_{i=1}^{j} w_{(i\le j)}}\right)\left[\log Y_{n-j+1,n} - \log Y_{n-k,n}\right]}{\frac{1}{k}\sum_{j=1}^{k}\left[\log \frac{\sum_{i=1}^{k+1} w_{(i\le k+1)}}{\sum_{i=1}^{j} w_{(i\le j)}}\right]^2}. \tag{8}$$

For the estimated value $\widetilde{Y}_{n-j+1,n}$ of the j's largest observation, $j = 1, \ldots, k$, as vertical projection onto the line estimate, we obtain a weights-adjusted version of eq.(6),

$$\widetilde{Y}_{n-j+1,n} = Y_{n-k,n}\left(\frac{\sum_{i=1}^{k+1} w_{(i\le k+1)}}{\sum_{i=1}^{j} w_{(i\le j)}}\right)^{\gamma}. \tag{9}$$

C.5 Computation of Top Shares

Assuming that the Pareto QQ-plot becomes approximately linear from the k's largest observation, $Y_{n-k+1,n} \equiv Y_{\text{base}}$, the complete distribution F is, for $y > Y_{\text{base}}$,

$$F(y) = p + (1-p)\left(1 - \left(\frac{y}{Y_{\text{base}}}\right)^{-1/\gamma}\right) \tag{10}$$

with $p = \widehat{F}_{\text{SOEP}}(Y_{\text{base}})$ and $\widehat{F}_{\text{SOEP}}$ being the empirical CDF of the survey data. Upon inversion, an upper quantile is

$$Q(u) = \left(\frac{1-u}{1-p}\right)^{-\gamma} Y_{\text{base}}, \qquad (u > p). \tag{11}$$

In the unweighted case, the resulting well-known (see, e.g. Embrechts, Klüppelberg, and Mikosch 1997, p. 348) estimates of the tail of the distributions and the top quantile estimate are, then, with $1 - p = k/n$,

$$1-\widehat{F}(y)=\left(\frac{k}{n}\right)\left(\frac{y}{Y_{n-k+1,n}}\right)^{-1/\widehat{\gamma}},\qquad \widehat{y}_u=\left(\frac{n}{k}(1-u)\right)^{-\widehat{\gamma}}Y_{n-k+1,n}.$$

Taking into account the survey sampling weights ω_i for household i enumerated from the poorest to the richest, we have

$$p=\frac{\sum_{i=1}^{n-k}\omega_i}{\sum_{i=1}^{n}\omega_i}.$$

The expected value is then simply $E(Y)=pE_{\text{SOEP}}+(1-p)\left(\frac{\alpha}{\alpha-1}\right)Y_{\text{base}}$ with $\alpha=1/\gamma$ and E_{SOEP} being the empirical mean in the survey data conditional on Y not exceeding Y_{base}. The share of the top t% then is, with $1-t=u>p$,

$$t^{1-1/\alpha}\left(\frac{\alpha}{\alpha-1}\right)Y_{\text{base}}(1-p)^{1/\alpha}\Big/E(Y). \tag{12}$$

The inverted Pareto coefficient $E(Y|Y>y)/y$ with y equal to the top t quantile y_{1-t} and $1-t=u>p$ is $\alpha/(\alpha-1)$.

To facilitate the computation of top shares, we provide the Stata command beyondpareto_topshares which can be called after estimation of the upper tail index using beyondpareto.

References

Bartels, C., and M. Metzing. 2019. “An Integrated Approach for a Top-Corrected Income Distribution.” *The Journal of Economic Inequality* 17: 125–43.

Beckmannshagen, M., and C. Schröder. 2022. “Earnings Inequality and Working Hours Mismatch.” *Labour Economics* 76: 102184.

Beirlant, J., P. Vynckier, and J. L. Teugels. 1996. “Tail Index Estimation, Pareto Quantile Plots Regression Diagnostics.” *Journal of the American statistical Association* 91 (436): 1659–67.

Bönke, T., G. Corneo, and H. Lüthen. 2015. “Lifetime Earnings Inequality in germany.” *Journal of Labor Economics* 33 (1): 171–208.

Card, D., J. Heining, and P. Kline. 2013. “Workplace Heterogeneity and the Rise of West German Wage Inequality.” *Quarterly Journal of Economics* 128 (3): 967–1015.

Champernowne, D. G. 1953. “A Model of Income Distribution.” *The Economic Journal* 63 (250): 318–51.

Dauth, W., and J. Eppelsheimer. 2020. “Preparing the Sample of Integrated Labour Market Biographies (SIAB) for Scientific Analysis: a Guide.” *Journal for Labour Market Research* 54 (10): 1–14.

Disslbacher, F., M. Ertl, E. List, P. Mokre, and M. Schnetzer. 2023. “On Top of the Top: A Generalized Approach to the Estimation of Wealth Distributions.” *Available at SSRN 4499915*.

Drechsler, J., and J. Ludsteck. 2025. “Imputation Strategies for Rightcensored Wages in Longitudinal Datasets.” *arXiv preprint arXiv:2502.12967*.

Dustmann, C., J. Ludsteck, and U. Schönberg. 2009. “Revisiting the German Wage Structure.” *Quarterly Journal of Economics* 124 (2): 843–81.

Embrechts, P., C. Klüppelberg, and T. Mikosch. 1997. *Modelling Extremal Events for Insurance and Finance. Vol. 33 of Stochastic Modelling and Applied Probability*. Berlin: Springer.

Emmenegger, J., and R. Münnich. 2023. "Localising the Upper Tail: How Top Income Corrections Affect Measures of Regional Inequality." *Jahrbucher für Nationalokonomie und Statistik* 243 (3–4): 285–317.

FDZ-RV (Forschungsdatenzentrum der Rentenversicherung). 2024. SOEP-RV Versichertenkontenstichprobe 2020 (Version 2.0). https://doi.org/10.5684/soep.v37-RV.VSKT2020.V2.0

Gabaix, X. 2016. "Power Laws in Economics: An Introduction." *The Journal of Economic Perspectives* 30 (1): 185–206.

Gabaix, X., and R. Ibragimov. 2011. "Rank – 1/2: A Simple Way to Improve the Ols Estimation of Tail Exponents." *Journal of Business & Economic Statistics* 29 (1): 24–39.

Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp. 2019. "The German Socio-Economic Panel (SOEP)." *Jahrbucher für Nationalokonomie und Statistik* 239 (2): 345–60.

Jenkins, S. P. 2017. "Pareto Models, Top Incomes and Recent Trends in uk Income Inequality." *Economica* 84 (334): 261–89.

Karlsson, M., Y. Wang, and N. R. Ziebarth. 2024. "Getting the Right Tail Right: Modeling Tails of Health Expenditure Distributions." *Journal of Health Economics* 97: 102912.

Kesten, H. 1973. "Random Difference Equations and Renewal Theory for Products of Random Matrices." *Acta Mathematica* 131 (1): 207–48.

König, J., C. Schröder, and E. N. Wolff. 2020. *Wealth Inequalities*, 1–38. Cham: Springer International Publishing.

Konig, J., C. Schluter, and C. Schroder. 2024 In press. "Routes to the Top." *Review of Income and Wealth*.

König, J., C. Schluter, C. Schröder, I. Retter, and M. Beckmannshagen. 2025. "The Beyondpareto Command for Optimal Extreme-Value Index Estimation." *STATA Journal* 25 (1): 169–88.

Lüthen, H., C. Schröder, M. M. Grabka, J. Goebel, T. Mika, D. Brüggmann, S. Ellert, and H. Penz. 2022. "SOEP-RV: Linking German Socio-Economic Panel Data to Pension Records." *Jahrbucher für Nationalokonomie und Statistik* 242 (2): 291–307.

Schluter, C. 2018. "Top Incomes, Heavy Tails, and Rank-Size Regressions." *Econometrics* 6 (1): 10.

Schluter, C. 2021. "On Zipf's Law and the Bias of Zipf Regressions." *Empirical Economics* 61 (2): 529–48.

Schluter, C., and M. Trede. 2019. "Size Distributions Reconsidered." *Econometric Reviews* 38 (6): 695–710.

Schluter, C., and M. Trede. 2024. "Spatial Earnings Inequality." *The Journal of Economic Inequality* 22: 531–550.

Schröder, C., M. M. Grabka, J. König, O. Morales, M. Priem, C. Schluter, J. Seebauer, and A. Winkler. 2023. "Sonderauswertungen des Sozio-oekonomischen Panels (SOEP) 2020 und 2021 zu Löhnen und Arbeitszeiten in der Pandemie." *Studie im Auftrag der Mindestlohnkommission*.

Simon, H. A. 1955. "On a Class of Skew Distribution Functions." *Biometrika* 42 (3/4): 425–40.

Wildauer, R., and J. Kapeller. 2022. "Tracing the Invisible Rich: A New Approach to Modelling Pareto Tails in Survey Data." *Labour Economics* 75: 102145.