DIW BERLIN

2066

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2023

Routes to the Top

Johannes König, Christian Schluter and Carsten Schröder

IMPRESSUM

DIW Berlin, 2023

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
https://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
https://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
https://ideas.repec.org/s/diw/diwwpp.html
https://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Routes to the Top*

Johannes König† Christian Schluter‡ Carsten Schröder§

December 19, 2023

Abstract

Who makes it to the top? We use the leading, socio-economic survey in Germany supplemented by extensive data on the rich to answer this question. We identify the key predictors for belonging to the top 1 percent of income, wealth, and both distributions jointly. Although we consider many, only a few traits matter: Entrepreneurship and self-employment in conjunction with a sizable inheritance of company assets is the most important covariate combination across all rich groups. Our data suggest that all top 1 percent groups, but especially the joint top 1 percent, are predominantly populated by intergenerational entrepreneurs.

Keywords top wealth, top income, intergenerational transfers, rich-group classification modelling, predictions

JEL Classification D31 · C38 · D63

*Acknowledgements: We would like to thank the German Federal Ministry of Education and Research and the German Federal Ministry of Labor and Social Affairs for financial support of our field work to collect the new subsample, P, of the Socio-Economic Panel. Johannes König and Carsten Schröder gratefully acknowledge financial support by Deutsche Forschungsgemeinschaft (project "Wealth holders at the Top" (WATT), project number: 430972113), Christian Schluter from ANR-17-EURE-0020, and all authors from the ANR-DFG (grant ANR-19-FRAL-0006-01). We thank Charlotte Bartels, Mattis Beckmannshagen, Emmanuel Flachaire, Jonathan Goupille-Lebret, Markus M. Grabka, Jana Hamdan, Jonas Jessen, Robin Jessen, Isabel Martinez, Clara Martinez-Toledano, Lukas Menkhoff, Thomas Piketty, Jacob Robbins, Johannes Seebauer and Mark Trede for helpful comments and suggestions. Further, we thank the participants of the MPIFG conference on wealth, LAGV 2022, the economics research seminar at Freie Universität Berlin, the conference of DFG priority program 1764, the DFG-supported WATT workshop "Taxation and Inequality", ECINEQ 2023, and ESEM 2023. We thank Paul Brockmann, Hannah Penz, and Thomas Rieger for outstanding research assistance.

†DIW Berlin, Mohrenstr. 58, 10117 Berlin, Germany (jkoenig@diw.de)

‡Aix-Marseille Université (Aix Marseille School of Economics), CNRS & EHESS, 5 Boulevard Maurice Bourdet CS 50498, 13205 Marseille Cedex 01, France, and Department of Economics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK (christian.schluter@univ-amu.fr and https://christianschluter.github.io/)

§Corresponding author, DIW Berlin, Mohrenstr. 58, 10117 Berlin, Germany, and Freie Universität Berlin, Department of Economics, (cschroeder@diw.de)

1. Introduction

Wealth and income concentration continue to be at the top of the public and academic debate (see, e.g., Piketty and Saez 2003; Atkinson et al. 2011; Bricker et al. 2016; Saez and Zucman 2016; Piketty et al. 2018; Kuhn et al. 2020; Smith et al. 2023). While much has been learned about the levels and long-run dynamics of wealth and income inequality, many substantive questions still remain unanswered. A key open question addressed in this paper is: Who makes it to the top? We are the first to use new data on top earners and wealth holders from a just-introduced subsample in the German Socio-Economic Panel (SOEP) to shed light on this question. In addition to income and wealth 'from bottom to top,' our data contain multiple crucial covariates including asset-specific information on inheritances, career histories, rich demographics, and personality traits. With this data we identify and quantify the key *predictors* for being among the top 1 percent.

A better understanding of the people at the top is relevant for recurrent debates about income and wealth taxes as well as their design in targeting specific sources of income and wealth. Further, it is important in determining how top wealth is perpetuated across generations and how this results in intergenerational mobility or immobility (see, e.g., Kopczuk and Zwick (2020)). Intergenerational immobility at the top is partly a policy choice, since policy makers can employ tools, like the inheritance tax, to shape it. Yet, in many countries inheritance tax law exempts bequests of company assets.

Empirical research on the most important predictors of belonging to the top 1 percent has been out of reach due to data limitations. Many important contributions in the inequality literature rely on large-scale restricted-access administrative data which have been collected for the purpose of levying taxes. As a consequence, these data provide detailed information relevant for the calculation of a tax unit's tax burden, but usually lack covariates—from household and individual characteristics to educational choices to employment biographies and detailed inheritance data—that would enable a comprehensive study of the routes to the top of the distribution. By contrast, survey data are easily accessible, often provide a rich set of covariates, but usually fail to sample adequately the top wealth tail (the well-known problem of the "missing rich").

Our data come from the Socio-Economic Panel (wave of 2019), which was augmented by a new and fully-integrated subsample of the wealthy (sample P). We label the combined data SOEP+P. The sampling universe of sample P is the population of substantial shareholders residing in Germany who are invested in at least one company globally. We used business register data from around the world to construct the sample (Schröder et al. 2020). The combined data are most suited for our research question for two reasons. First, SOEP+P provides detailed information on both income and wealth while also covering well

the top tails of both distributions. We therefore analyze income and wealth separately as well as *jointly*. Wealth is reported directly and does not have to be inferred by capitalizing capital income. Since the new subsample of the wealthy is fully integrated in the SOEP and all the variables are exactly comparable across the rich and the non-rich population, we do not have to pursue any data harmonization. Second, SOEP+P provides, unlike many administrative datasets, a broad set of variables and thus potential predictors in four broad domains: Socio-demographic characteristics (such as age, education, gender, household composition), labor-market characteristics (labor-market experience, entrepreneurship, job characteristics, etc.), intergenerational transfers (types and levels of inheritances, such as real estate or company assets), and personality traits (the so-called Big 5 personality traits and risk tolerance).

Our analysis proceeds in three steps. First, we use our new data to assess the current extent of wealth concentration in Germany. In doing so, we update previous estimates and also show that we overcome the "missing rich" problem from previous studies. Specifically, in 2019 the top 1 percent, who hold at least 1.9 million euros, own about 23% of total wealth, roughly as much as the bottom 70%. This wealth share is consistent with the estimate of 23% reported in Albers et al. (2022) who use data from the Income and Expenditure Survey (EVS) after uprating and top-correcting them. We also note that there are no administrative wealth registers for Germany.

Secondly, we show the concentration of the second core determinant of economic well-being, income, and how it correlates with wealth. In terms of yearly household income the top 1 percent make at least 203,000 euros, and hold a share of 7% of total income. This 7% income share for the top 1% corresponds closely to the 6.6% income share reported in Drechsel-Grau et al. (2022) who merge confidential data from the German Taxpayer Panel to individual-level administrative income data from the Sample of Integrated Labor Market Biographies. Top wealth and top income are usually studied in isolation because of data constraints. As a substantive contribution, we show that wealth and income at the top are strongly correlated: Roughly half of those in the top 1 percent of wealth are also in the top 1 percent of household income (0.5% of the population). This joint top 1 percent stand out as a group of extreme wealth: they hold about 21% of total wealth and thus roughly 90% of the wealth share that the top 1 percent of wealth hold. These results underscore the observation made in Saez and Zucman (2016, p. 525) that understanding the link between the wealth and income is vital for assessing wealth taxation proposals.

As our third and core contribution we study membership of the top 1 percent in terms of wealth, income, and both jointly using state-of-the-art *nonparametric* classification models taken from the statistical learning literature (James et al. 2021). These models are designed

to fit complex data relationships, in particular non-linearities and covariate interactions, without simply overfitting and they perform well on not-yet-seen data. Our research design focuses on *prediction* and the measurement of predictor importance, and enables us to identify key correlations and predictor interactions that classic estimation techniques would not have uncovered. Our predictive modelling and its clear empirical findings complement research efforts to isolate causal relationships through random variation in covariates.[1] In the analysis of the determinants of wealth, the identification of credible natural experiments is still rare,[2] and our predictive analysis contributes to this endeavor by guiding this search and revealing which sources of variation are most important.

In particular, we estimate random forests, a technique that grows an ensemble of data-driven hierarchically structured classification trees by binary data splitting, which enables the study of non-linearities and variable interaction.[3] With this ensemble in hand, one averages over the predictions of the individual trees to arrive at final predictions. The resulting ensemble estimator has a lower variance than any individual tree, and we show that this method clearly outperforms classic logit modelling for all our outcome variables. Random forests are more difficult to interpret as they lack direct analogues to model coefficients. We pursue a thorough interpretation of the random forest using several model-specific and model-agnostic variable importance metrics which paint a coherent picture of the top predictors for rich group membership.

Our key empirical finding is that our approach reduces the large set of potential predictors we feed into the random forests to a very small number of key interacting predictors: entrepreneurship *in conjunction* with a large inheritance of company assets (as opposed to real estate or financial assets) is the most important covariate combination to predict top rich group membership. Other covariates play clearly subordinate roles. This

[1] Mullainathan and Spiess (2017) label the different objectives of prediction and causal inference as $\hat{y}$ and $\hat{\beta}$, and observe that "the success of machine learning at intelligence tasks is largely due to its ability to discover complex structure that was not specified in advance. It manages to fit complex and very flexible functional forms to the data without simply overfitting; it finds functions that work well out-of-sample" and "Machine learning provides a powerful tool to hear, more clearly than ever, what the data have to say." The complementarity (as opposed to conflict) between the approaches is further argued in Athey and Imbens (2019). The merit of prediction, and specifically the use of ML techniques to this end, have been successfully demonstrated in other fields such as oncology and bioinformatics, the Cancer Genome Atlas (TCGA) of the US National Cancer Institute being but one very prominent example. Our objective of identifying the top predictors in an interpretable manner is in the same spirit.

[2] An exception is Nekoei and Seim (2023), who study the impact of random timing of the receipt of inheritances.

[3] The use of machine learning is also becoming increasingly widespread in distribution analyses. Regression trees and random forests are used, for example, to study the relationship between inheritances and wealth (Salas-Rojo and Rodríguez 2022) and inequalities of opportunities (Brunori and Neidhöfer 2021).

adds nuance to the debate about whether the richest individuals are passive recipients or active creators of their fortunes (inheritors and rentiers vs. entrepreneurs): It is a combination of both.

The link between inheritances and top rich group membership may appear mechanical and straightforward. However, inheritances are generally anticipated and behavioral adjustments of labor supply, human capital accumulation, and other choice variables may result, which is why considerable effort has been put toward understanding the impact of inheritances on wealth and the wealth distribution (Boserup et al. 2016, 2018; Adermon et al. 2018; Black et al. 2020; Fagereng et al. 2021; Black et al. 2022; Nekoei and Seim 2023). Further, we show that the most predictive feature for top rich group membership is the inheritance of company assets in conjunction with entrepreneurship. This points away from the idea of a mere mechanical effect, which any type of inheritance would be able to provide. Rather, it speaks to the hypothesis that we are looking at the intergenerational transmission of entrepreneurship.

We can also rule out some clear cases of what does not pin down rich group membership. For example, although, education is likely to be important for economic success, it is not a strong predictor of top rich group membership. So while our models do not estimate causal effects, they deliver important information about the processes in our society that lead to high income and wealth and thus constitute a basis for a) the specification of models of individual wealth accumulation, and b) informed discussion about intergenerational mobility and (in)equalities of opportunities. Moreover, the results of our predictive model are informative about who should be targeted in prospective wealth or inheritance tax reforms. Our results highlight that current exemptions of company assets may be detrimental to the maximization of revenues from these taxes.

To illustrate quantitatively the role of the key predictors we estimate predictive margins from the random forest models, that is partial dependence plots. The combination of being self-employed and having received a firm inheritance of at least 2.5 million euros, raises the base probability of belonging to the top 1 percent income group by 26 percentage points, to the top 1 percent wealth group by 38 percentage points, and to the joint top 1 percent group by 26 percentage points. Being self-employed but having received an equally sized inheritance that is not a firm does not raise predicted probabilities nearly as much. The respective predicted change in probability for the top 1 percent of income is 16 percentage points, for the top 1 percent of wealth 30 percentage points, and 13 percentage points for the joint top 1 percent. Thus, a non-firm inheritance confers about half the benefit of a firm inheritance with respect to the inclusion probability in joint top 1 percent. This shows that it is the combination of predictors that explains membership

in the top 1 percent.

By contrast, if one lacks one of the covariates in this combination, predicted probabilities of top 1 percent membership fall drastically. Take membership in the top wealth group as another example. Having received a firm inheritance of at least 2.5 million euros but not being self-employed reduces the predicted change in membership probability from 38 to 31 percentage points. Conversely, being self-employed but cutting the firm inheritance to zero, reduces the change in probability from 38 to 7 percentage points.

Further complementary descriptive analyses show that the top 1 percent differ from the rest of the population in terms of their portfolio composition and position in the labor market. The joint top 1 percent hold their wealth predominantly in closely-held businesses (42%, with 62% being held in a single firm), while this share is only 35% for those in the top tails of income or wealth, and quickly declines with net wealth. This distinction in portfolio composition along the wealth distribution is also present in US data as Kuhn et al. (2020) show. Further, the joint top 1 percent tend to work in small to medium-sized firms in the financial, real estate, and the skilled services sectors. Remarkably, the joint top 1 percent share many characteristics highlighted in recent work on top income recipients in the US based on administrative data (see, e.g., Smith et al. 2019) and the household portfolio compositions reported in Norwegian administrative data (Ozkan et al. 2023). However, the key difference we highlight is the importance of intergenerational transfers of company assets in Germany.

Taken together, our data suggest that the top 1 percent groups —especially the joint top 1 percent— are populated by a class of *intergenerational entrepreneurs*. The predictions from our classification models unanimously show that entrepreneurship in conjunction with sizable firm inheritances are the strongest predictors for being in the top 1 percent groups. This is in stark contrast to the rich groups just below them, the top 10-1 percent, who belong to the top 10 but not the top 1 percent. For this group entrepreneurship and education are important predictors, but firm inheritances far less so. A further stark contrast comes in the form of the portfolio composition of the top 10-1 percent groups: For all of these groups more than half of their portfolio is held as real estate and less than 15% as firms.

Our findings relate to several strands of literature. First, our validation exercise shows that survey data can provide convincing estimates of wealth and income concentration without the need to augment the data with external rich lists (Bricker et al. 2016; Vermeulen 2016, 2018; Bach et al. 2019). This is particularly important for countries in which no register data are available (e.g., because there is no wealth tax) or where these data are not easily accessible to the research community. Further, our survey gives

important insights into the dependence between wealth and income, which deepens our understanding of economic inequality. Further, we gain information with respect the joint distribution of two of the most important tax bases and of the potential for the rich to shift between these bases (Saez 2002; Christiansen and Tuomala 2008; Saez and Zucman 2019).

Second, our results have important implications for the literature on intergenerational wealth transmission and social mobility (Piketty et al. 2014a; Boserup et al. 2016, 2018; Kopczuk and Zwick 2020; Fagereng et al. 2021; Black et al. 2022; Ozkan et al. 2023). The group with the most extreme wealth concentration, the joint top 1 percent, who hold 21% of all wealth, is generally comprised of entrepreneurs that have received substantial firm inheritances. Tax law in Germany, like in other European countries, codifies and thus exacerbates firm inheritances' impact on intergenerational immobility through partial or full exemption. Thus, the route toward this top group tends to be paved by an intergenerational transmission as opposed to an independent career path, facilitated by an enabling tax regime. The top 10-1 percent receive predominantly other inheritances (e.g. real estate and tangibles), which tend to have smaller long-run returns than equity (Jordà et al. 2019).

Third, the theoretical literature on wealth concentration puts entrepreneurship as one of the probable mechanisms by which wealthy individuals manage to both receive high incomes and hold large shares of aggregate wealth (Cagetti and De Nardi 2006; DeNardi and Fella 2017; Benhabib et al. 2019; Kopczuk and Zwick 2020). Auray et al. (2022) tested the entrepreneurship mechanism in a dynamic heterogeneous agent model of a closed economy and showed that it produces a good fit for income and wealth inequality levels as well as for dynamics in France. Our results show that the combination of firm inheritances and entrepreneurship has tremendous predictive power for top group membership, offering supportive evidence of the entrepreneurship mechanism.

The outline of this article is as follows. After a brief summary of the sampling framework for the new sample of the wealthy, we conduct two extensive validations, focusing in Section 2.2 on wealth and in Section 2.3 on income. In Section 3.1 we model the dependence between wealth and income, which leads us to consider, in addition to the wealthy and the income rich, the top of the joint distribution. In the descriptive analysis of Section 3.2 we look at the six rich groups. In the key Section 4 we use nonparametric classification models in order to identify the top predictors using interpretable machine learning techniques. Section 4.3 provides a summary and discusses our findings in the context of the literature. Section 5 concludes. The Appendix contains extensive supplementary material.

2. New estimates of wealth and income concentration and survey validation

We briefly summarize how the SOEP-P sample was generated. The companion paper Schröder et al. (2020) provides a comprehensive methodological exposition of the sampling strategy without providing an analysis of the SOEP-P data.

The sampling universe of the SOEP-P subpopulation is defined as shareholders residing in Germany who have invested in at least one company globally to the extent that the person's business investment is listed in relevant business registers. Note that, even when an individual's wealth is invested predominantly in another type of asset (for example, mutual funds) as long as that individual owns some business assets, they may be included in our sample. Thus, the sampling frame is not limited solely to entrepreneurs.

The threshold for having substantial shares and thus being listed is 0.1% of all shares of a company. The motivation for this sampling procedure is an empirical regularity observed in many countries: The percentage of wealthy individuals who are invested in companies is very high, and so is the monetary value of their investments (see, e.g., Bucks et al. 2009; Bricker et al. 2017; Martínez-Toledano 2020; Wolff 2021; Smith et al. 2023). From all these shareholders, a probabilistic sample of individuals was drawn, stratified according to the value of their shareholdings. The resulting SOEP-P sample is therefore a stratified random sample of 1,960 top shareholders residing in Germany drawn from the top 600,000 Germans with the highest monetary values of investments. The Data Appendix C details the composition of individual personal balance sheets, and Table C.1 reports how SOEP-P populates the right tail of the wealth distribution.

The sampled individuals—and their household members—were then surveyed using the standard SOEP questionnaire. Hence, wealth portfolios were measured *directly*, and we do not have to infer the wealth from income flows. SOEP-P is a fully integrated SOEP subsample, which means that all variables are fully comparable across SOEP and SOEP-P, and the SOEP weights are adjusted to account for the inclusion of the new sample (Siegers et al. 2021). Accordingly, the SOEP-P sample and all other SOEP samples can be analyzed jointly, enabling a comprehensive analysis of the marginal and *joint* wealth and income distributions in one unified data framework. We refer to **SOEP+P** as a shorthand for the combined and integrated survey.

2.1. The "missing rich" and the top tail of the wealth distribution

We proceed by illustrating, first, how household wealth (net of debts) is underrepresented in the standard Socio-Economic Panel (SOEP). This is problematic since SOEP is the

leading panel for Germany and one that is frequently used for inequality analysis.[4] We then demonstrate how SOEP-P successfully populates the upper tail of the wealth distribution, making it a *key data innovation* that enables reliable top wealth and income measurement.

Our first benchmark uses external rich list data. In recent contributions to the "missing rich" problem, researchers have addressed the problem pragmatically by augmenting survey data with external rich lists (for instance, Bricker et al. (2016) and Vermeulen (2016, 2018) use the Forbes list). In the case of Germany, the leading national rich list is published by Manager Magazin (MM), and was used, for instance, in Bach et al. (2019). We take the MM data here at face value, that is, we do not consider the question of how sampling weights should be redefined, as SOEP's design is complex, nor do we address the issue of potentially inconsistent wealth measurement across data sources.[5]

To visualize the top tail of the wealth distribution, we use a Pareto quantile-quantile (QQ) plot. This is a diagnostic device that correlates the empirical top quantiles of the wealth distribution and the corresponding theoretical or population quantiles of the Pareto distribution (see Statistical Appendix A.1 for a formal exposition). Figure 1.(a) depicts this Pareto QQ plot for approximately the richest 10% of households in the SOEP in 2019. The estimated slope using the rank-size regression methods (explained below) is .55. As is evident in Figure 1.(a), adding the "missing rich" from MM effectively appends a disconnected right tail to the Pareto QQ plot. Although this new right tail is approximately linear, its slope of about 1 is substantially larger than the tail slope based on SOEP alone. Further, we find a large vertical jump in the plot. This is a consequence of the fact that top wealth in SOEP and MM do not overlap. The household with the highest wealth in the SOEP is considerably less wealthy than the household with the lowest wealth in the MM list. Combining the two datasets to estimate a common slope is also problematic in the presence of such a large vertical jump, leading to a distorted overall slope and potentially distorted wealth share predictions.

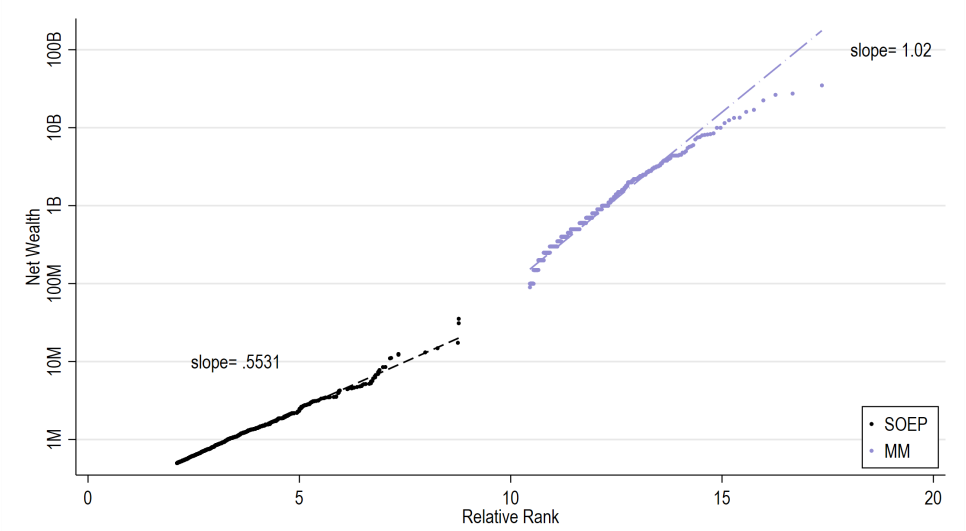
Panel (b) of Figure 1 shows that SOEP-P populates the right tail of the survey's wealth distribution. In particular, it *adds extremes* to SOEP and *fills in* the gap between SOEP and the MM list. The result is a "dovetail joint", so the vertical jump evident in panel

[4] For instance, the SOEP is the principal data source for the periodic Poverty and Wealth Report of the German government; see `https://www.armuts-und-reichtumsbericht.de`. According to SOEP records, 1,317 peer-reviewed papers were published using SOEP data between 2011 and 2020, 234 of these on the topic of inequality.

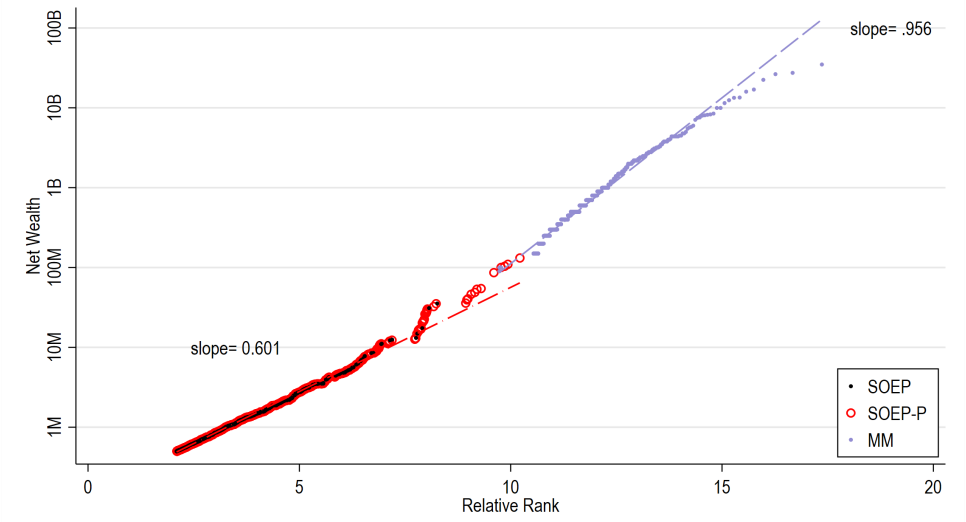
[5] As discussed in Bach et al. (2019), the wealth concept used by MM is based on expert valuation and mainly captures business and real estate wealth. These valuations generally refer only to gross wealth and not to net wealth, ignoring the liabilities in the balance sheets of the richest. Raub et al. (2010) report a substantial difference between wealth reported in Forbes and estate tax filings. Finally, we note that the reference unit in MM data refers sometimes to individuals, sometimes to families, and sometimes to "family clans." Smith et al. (2023) raise the same concerns about the use of rich lists.

(a) of the figure has disappeared. The plot is now approximately linear for the depicted high wealth levels, which are now well connected. Because of this observed linearity, extrapolation methods based on the model given by equation (1) below will then enable us to dispense completely with the MM list.

Figure 1: Pareto QQ-plots for top wealth: Filling in and appending extremes



(a) Without SOEP-P



(b) With SOEP-P

Notes. Panel (a): Upper wealth order statistics of the German Socio-Economic Panel (SOEP, black) in 2019, and the national rich list (Manager Magazin, MM, purple). Panel (b) includes SOEP-P (red). The unit of observation is the household. Wealth is in 2019 euros. Slope estimates are based on rank-size regressions. For a detailed exposition of the Pareto quantile quantile (QQ) plot and the estimation approach, see the Statistical Appendix A.1. Appendix Table C.1 reports the exact number of observations over high thresholds. *Source*: SOEP, SOEP-P, and Manager Magazin in 2019.

2.2. Validation metrics 1: Top wealth shares and the tail index

As validation metrics we follow the literature and report wealth shares and estimates of the top tail index. More specifically, the consensus view among researchers is that the top of the wealth (and the income) distribution is Pareto-like, that is, for sufficiently large wealth or income

$$F_X(x) = 1 - x^{-1/\gamma_X} l_X(x), \tag{1}$$

where X indicates either wealth or income, F denotes the associated cumulative distribution, and l a slowly varying nuisance function that is constant asymptotically. $\gamma > 0$ is called the extreme value index, and the Pareto or tail index ($\alpha \equiv 1/\gamma$) is its reciprocal. This semi-parametric model enables us to confidently study tail areas of the distribution that are less densely populated by the sample data and to extrapolate beyond them.

We will estimate $\gamma > 0$ below using rank-size regressions that are based on the behavior of the Pareto QQ plot,[6] the distributional theory of which has recently been developed in Schluter (2018). The Statistical Appendix A provides a detailed exposition, including our generalization to account for the complex survey design. Top wealth shares are computed based on the estimated tail index (see the Statistical Appendix A.6 for details).

We consider as benchmarks some estimates reported in the literature.[7] For instance, Vermeulen (2018) uses the Household Finance and Consumption Survey (HFCS) for Germany in 2010 and sets wealth thresholds at 0.5, 1, and 2 million euros. The rank-size regression without augmentation results in estimates of the tail index of 1.54, 1.64, and 1.87, respectively, while Bach et al. (2019) report estimates of 1.53, 1.61, and 1.77, respectively. Vermeulen then appends 52 Forbes billionaires, which reduces his estimates to 1.40, 1.39, and 1.38, respectively. By contrast, Bach et al. (2019) append the tail provided by the MM national rich list, and using the 300 largest entries, reduces their threshold-specific estimates to 1.37, 1.36, and 1.34, respectively. Turning to the implied top 1 percent wealth share, Vermeulen (2018) reports threshold-dependent wealth shares of 34, 33, and 32%, respectively, while the respective estimates in Bach et al. (2019) are 33, 32, and 31%, respectively. Albers et al. (2022) use data from the Income and Expenditure

[6] For Pareto-like distributions, the Pareto QQ plot becomes linear only *eventually*, and $\gamma > 0$ is its *ultimate* slope. That is, this Pareto QQ plot describes the sample analogue of the asymptotic behavior of the the log of the tail quantile function U, $\log U(x) \sim \gamma \log x$ as $x \to \infty$, where $U(x) \equiv F^{-1}(1 - 1/x) = x^{\gamma}\tilde{l}(x)$ and $\tilde{l}$ is a slowly varying nuisance function.

[7] There are no administrative wealth records in Germany, so that all studies of top wealth in Germany rely on either survey data, possibly with augmentation from rich lists, or capital income capitalization. Both augmentation with rich lists and capital income capitalization may come with serious uncertainties discussed in Kopczuk (2015) and König et al. (2020).

Survey (EVS), after uprating and top-correcting them, to arrive at a top 10 percent wealth share of about 57% and a top 1 percent share of about 23% in 2018.

How do SOEP and SOEP+P compare with this? Table 1 reports the results. Panel A of Table 1 considers SOEP alone, to quantify the distortions resulting from underrepresented top wealth. For a wealth threshold of 0.5M, the Pareto index α estimate is 1.8 (and 1/1.8=.56 as reported in Figure 1.A), implying a top 1 percent wealth share of about 20%. Increasing the wealth threshold to 2 million euros decreases the estimate of α (1.66) and raises this wealth share by less than a percentage point. Turning to the precision of the estimates, Table 1 reveals that the variability of the Pareto parameter can be large, for instance, .097 when the threshold is 1 million euros and .167 for the 2 million euros threshold. The confidence limits for the top 1 percent wealth shares and are (19%,21%) for the former and (17%,28%) for the latter (see Appendix Table D.2).

Table 1: Top wealth in Germany in 2019

		tail index		wealth share	
threshold	k	$\hat{\alpha}$	$SE(\hat{\alpha})$	Top 10%	Top 1%
A. SOEP alone (under-represented wealth)					
0.5M€	1457	1.802	0.052	55.04	19.76
1.0M€	442	1.850	0.097	55.02	19.41
2.0M€	117	1.657	0.167	55.43	20.43
B. SOEP+P					
0.5M€	2735	1.683	0.031	57.49	22.59
1.0M€	1307	1.672	0.039	57.63	23.00
2.0M€	626	1.522	0.042	58.14	24.40
C. Optimal wealth threshold selection					
0.402M€	3370	1.665	0.032	57.45	22.90

Notes. The wealth distribution is given by equation (1). Tail estimates ($\alpha \equiv 1/\gamma$) are obtained from standard rank-size regressions of wealth above the stated wealth threshold; see Statistical Appendix, Section A.2. k denotes the number of upper-order statistics corresponding to the fixed threshold. Statistical Appendix Section A.6 gives details for the computation of the wealth shares, Empirical Appendix Table D.2 reports the confidence limits for the wealth shares. *Source*: SOEP+P.

Next, for Panel B, we re-estimate the tail indices for the fixed thresholds and the associated wealth shares using SOEP+P. Using the fixed arbitrary wealth threshold of 0.5 million euros, the Pareto index α estimate is now 1.68 and the associated top 1 percent wealth share 23%. Increasing the wealth threshold has a small effect on this wealth share, as it rises to about 24% at 2 million euros. As regards precision of the estimates, the

standard error of the Pareto index for the case of the 2 million euros threshold has been substantially reduced to .042 (instead of .167), which is the result of using more data in the estimation.

Finally, in Panel C, we present our preferred estimates based on the optimal data-dependent threshold selection which optimally trades off variability and bias of the estimator, i.e., the asymptotic mean-squared error of the estimator is minimized. The statistical theory, developed in Schluter (2018, 2020), is summarized in the Statistical Appendix A.4.[8] The optimal wealth threshold is estimated to be about 0.4 million euros, leading to an estimate of 1.66 for α and a top 1 percent wealth share of about 23%. The precision of the Pareto parameter estimate is much improved (.032 compared to, e.g., .042 for a 2 million euros threshold). The left and right 95% confidence limits are 1.60 and 1.72, respectively.[9]

We conclude that our point estimates of both the tail index and the top wealth shares are comparable to those reported in the literature using alternative data sets supplemented by rich lists. However, we have innovated by using new statistical methods that yield greater precision. Our preferred Pareto index estimate is 1.66, which is much more precise than in the previous literature, and is robust.[10] As a substantive empirical observation, we note that $\hat{\alpha} = 1.66 < 2$, so the second moment of the wealth distribution does not exist. This implies that tail of the distribution is very heavy, which manifests itself observationally in heavily concentrated top wealth: The top 10 percent wealth share is 57% and the top 1 percent share 23%. These estimates are especially congruent with the most recent estimates of top wealth shares for Germany presented in Albers et al. (2022). Overall, we conclude that SOEP+P passes this validation test for top wealth.

2.3. Validation metrics 2: Top income shares and the tail index

The consensus in the established literature is that top income, like top wealth, is under-represented in leading surveys,[11] and that an appropriate model is given by equation (1). Data augmentation using national rich lists cannot remedy this problem, however, since

[8] The choice of the threshold is not innocuous for two reasons: (i) a bad (typically data-independent, blind) choice falling outside the appropriate tail area will bias the estimation; (ii) the number of observations exceeding the chosen threshold will determine the precision of the estimator.

[9] In the Empirical Appendix D.1 we rationalize the similarities between point estimates reported in panels B and C of the table using Hill-type plots, and further show that our threshold selection is robust against alternative data-dependent methods.

[10] For instance, in Appendix D.2 we show that the well-known Hill estimator yields, at our optimal threshold, a Pareto index estimate of 1.67, which is almost identical to ours.

[11] In Appendix Section D.4 we depict the diagnostic Pareto QQ plots which show that the inclusion of SOEP-P successfully appends extremes and fills in the upper tail of the income distributions.

their focus is on wealth and not income. Researchers are therefore constrained to using confidential tax data in conjunction with imputation techniques. In the case of SOEP-P, however, the usual SOEP income questionnaire is submitted to survey respondents. Therefore, as our next validation exercise, we ask: Does SOEP+P overcome the "missing rich" problem for income?

Only a few benchmark estimates for Germany exist, and comparison across these estimates is difficult because of differences in assessment units (e.g., tax units vs. households), types of data sources, imputation methods, and time points. Drechsel-Grau et al. (2022) merge confidential data from the German Taxpayer Panel to individual-level administrative income data from the Sample of Integrated Labor Market Biographies and find that the top 1 percent labor income share is 6.6% in 2016. Bartels and Waldenström (2022) report a top 1 percent income share for Germany in 2014 of 13% using the methodology of the World Income Database Alvaredo et al. (2021). Piketty et al. (2014b) report a top 1 percent income share of about 11%, which they averaged between 2005 and 2009. Bach et al. (2009) report a top 1 percent income share of about 12% for 2001. Despite the caveats about comparability, we will take these results as benchmark values.

We turn to our estimates of the top income shares and the underlying tail indices. Table 2 reports the results for our four household income concepts: Yearly market income (labor and capital incomes including pensions), capital income (income from dividends, interest, rent and leasing payments, and capital gains), labor market income, and post-government income. See Appendix C for detailed income definitions.

Table 2: Top incomes in Germany in 2019

income			tail index		income share	
concept	k	quantiles	$\hat{\alpha}$	$SE(\hat{\alpha})$	Top 10%	Top 1%
A. SOEP+P and fixed thresholds at P90						
MktInc	3072	95376	2.772	0.056	35.93	8.25
PostInc	3226	67460	3.025	0.060	26.72	5.72
LabInc	2927	90946	3.203	0.066	36.60	7.51
CapInc	2622	10242	1.587	0.035	58.42	24.94
B. Optimal thresholds						
MktInc	3998	82339	2.772	0.049	31.53	7.24
PostInc	5047	53817	3.002	0.047	26.59	5.73
LabInc	1770	116300	3.347	0.089	36.88	7.41
CapInc	3504	7822	1.590	0.030	58.46	24.88

Notes. Income concepts: *MktInc* is household market income, *PostInc* is post-government household income, *LabInc* is household labor income, *CapInc* is household capital income; see Appendix C for detailed definitions. As per wealth analysis, the tail index estimate $\hat{\alpha} \equiv 1/\hat{\gamma}$ is based on the rank-size regression estimator, and the optimal income threshold is obtained by minimizing the AMSE (as detailed in the Statistical Appendix). *Source*: SOEP+P.

In Panel A, the income threshold is fixed at the 90th percentile (P90) of the respective income distribution, a conventional choice in the top income literature. In Panel B, our optimally chosen income threshold is used. As it turns out, the estimates are similar across the two threshold choices. A closer look at the Hill-type plots of the tail index estimator (see Empirical Appendix D.4.2) reveals that the estimators exhibit extended horizontal section in which P90 has the good luck to fall. However, our optimal estimator picks the end point of the extended horizontal section, resulting in less variability. For instance, for market income, the standard error for $\hat{\alpha}$ falls from .056 to .049. The tail index estimate for capital income is reassuringly of the same order of magnitude as the estimates for wealth. Our preferred estimate of the top tail of the market income distribution is 2.77, which, also reassuringly, indicates that the upper tail of the pre-government income distribution is heavier than that for post-government income (estimated to be 3.0), which is as expected given the progressiveness of the German tax-benefit system. Except in the case capital income, the income tail indices are about twice as large as the tail index of wealth, leading to lower income concentration. Specifically, the top 10 percent income shares for market, post-government, and labor market income range from 27 to 37%, which is considerably smaller than for wealth. Only top capital income shares are of a similar order to the top wealth shares we estimate: about 58% for the top 10 percent

share, and roughly 25% for the top 1 percent share.[12]

Our estimated top income share for household market income and the top 1 percent is of the same order of magnitude as the estimate of Bartels and Waldenström (2022) despite different units of analysis, years, and data sources. The internal consistency of our estimates for capital income and wealth is also reassuring. We therefore conclude that our SOEP+P data also pass this validation test for top incomes.

3. Who are the rich? Top wealth and income descriptors

The literature is compartmentalized in its analysis of "the rich": The focus is either on wealth or on income, and the classification is simply based on being in the top tail of the respective distribution. Little is known about the extent to which wealth and income overlap, due primarily to a lack of suitable data (Saez and Zucman (2016, p. 525), see Martinez (2021) for an exception).

We first show, by means of dependence analysis, that such compartmentalization is problematic since the overlap between top wealth and top income is large but not one to one. This is particularly true for the top 1 percent. Thus, not everyone at the top of the wealth distribution is also at the top of the income distribution and vice versa.

We then turn to our key question: What are the routes to the top? Answering this question requires access to an extensive set of covariates, and SOEP+P enables this for the first time for Germany. In line with the public debate and related literature we focus on three rich groups: the top 1 percent of wealth, of income, and, because of the findings of the dependence analysis, those *jointly* in the top 1 percent of wealth and income. After a first look at descriptors, we proceed to identify formally the key *predictors* of being in a top group using state-of-the-art classification techniques from the field of machine learning. This enables us to quantify the relative importance of various predictors, like education, work experience, inheritances, and entrepreneurship.

3.1. How much do wealth and income overlap?

3.1.1. Rank correlations

We start by examining the dependence structure of wealth and income non-parametrically using (Spearman's) rank correlations for the four income concepts. Throughout, the

[12] In the Appendix, Section D.5, we collect our results for wealth and income concentration and depict them using Lorenz curves. Finally, we compare our results for Germany to other countries as reported in the literature.

marginal distribution of wealth is denoted by F_W, and that of generic income by F_Y. The rank correlation is then $\rho = cor(F_W, F_Y)$. These empirical rank correlations between wealth and income are all fairly high. In particular, the rank correlation between household wealth and capital income is .72, and between wealth and market income is .58. It is slightly lower for labor income at .41 unconditionally and at .55 when conditioning on working (see Appendix Table B.1 for more results). These results harmonize with the findings in Garbinti et al. (2021) about the joint distribution of wealth and income, as they report that the top 1 percent wealth group predominantly consists of top capital income earners and not top labor income earners.

3.1.2. A parsimonious copula model for wealth and income

Next, we seek to describe the relation between wealth and income as parsimoniously as possible using parametric copula models.[13] It turns out, as detailed in Statistical Appendix B and our extensive goodness-of-fit analysis, that the one-parameter Gumbel copula, say C_θ, describes the dependence structure across the entire distribution very well in the German case.[14] This parametric copula model enables us to confidently study tail areas of the joint distribution that are less densely populated by our sample data and to extrapolate beyond them. Using the copula, we can easily compute wealth and income shares for jointly defined top wealth and income groups (see Statistical Appendix B.2 for the detailed computation). Table 3 reports these wealth and income shares.

[13] Recall the definition of a copula C and Sklar's theorem. Let the two-dimensional random vector $[W, Y]$ have joint distribution H, then $H(w, y) = C(F_W(w), F_Y(y))$ and $C(u) = H(F_W^{-1}(u_1), F_Y^{-1}(u_2))$ where $u \in [0, 1]^2$. If the margins F_i are continuous, the copula is unique. Also recall that Spearman's rank correlation can be written as $\rho = 12 \int_{[0,1]^2} C(u)du - 3$, so ρ depends only on the underlying copula and can be interpreted as a moment of the copula. See, e.g., Nelsen (2006) or Hofert et al. (2017) for an extensive textbook treatment.

[14] In this Appendix, we compare the empirical (non-parametric) copula, which is the empirical joint distribution function of the empirical ranks of wealth and income, to the fitted model copula. The estimate of the scalar copula parameter θ of the Gumbel copula is obtained by inverting the theoretical mapping of the rank correlation ρ and θ and evaluating it at the empirical ρ, thus yielding a method-of-moments estimate. Appendix Table B.1 reports the estimates of θ.

Table 3: Wealth and income: The joint top shares

wealth & income	Population shares of joint		Income shares of joint		Wealth shares of joint	
	Top 10%	Top 1%	Top 10%	Top 1%	Top 10%	Top 1%
MktInc	5.40	0.51	14.30	3.12	40.81	20.99
PostInc	5.63	0.53	13.73	2.83	41.74	21.53
LabInc	4.14	0.37	13.36	1.70	34.65	17.40
CapInc	6.49	0.63	34.56	14.44	44.85	23.34

Notes. For the joint top shares, we consider the group of households that are in the top $s \times 100$ percent of the marginal wealth and income distribution. The wealth share of this group is $E\{W|W > F_W^{-1}(1-s), Y > F_Y^{-1}(1-s)\}/E(W)$ where $E(W)$ denotes average wealth. An analogous expression holds for the income share. Values were calculated using the fitted Gumbel copula and the fitted marginal distributions for incomes and wealth. For the detailed computation of the shares, see Statistical Appendix B.2. *Source*: SOEP+P.

Population shares for the joint top. The population shares at the top in the joint distribution are shown in columns 2 and 3 of Table 3. Top income does not coincide with top wealth, but many top income households are also members of the top wealth group. For instance, with respect to market income, the population share of the top 10 percent in both marginal distributions of wealth and income is 5.4% (and 6.5% for capital income). This share lies midway between the case of complete dependence (10%) and complete independence ($1\%{=}100 * (.1)^2\%$). To complement these numbers, we provide visualizations in Appendix B.3: A plot of the population shares across the entire joint distribution (i.e. the joint survival copula along the main diagonal), and the ridge plot, which evaluates for a selected wealth decile the copula density across income ranks.

Income and wealth shares in the joint top group. The population shares measure how dense the top of the joint distribution is. What are the associated wealth and income shares, and how do these shares compare to the top shares in the marginal distributions? Table 3 columns 4-7 reports the results.

For brevity, we focus on market income and the top 10 percent. The income share of the joint top 10 percent group is 14% (compared to 32% for the top 10% in the marginal income distribution; see Table 2), and the respective wealth share is 41% (compared to 57% for the top 10 percent in the marginal wealth distribution; see Table 1). Being in the joint top 10 percent predicts much higher wealth than income: This group captures about 43% of the market income accruing to the top 10 percent in the marginal income distribution but about 72% (= 100*41/57) of the wealth accruing to the top 10 percent in the marginal wealth distribution.

Turning to the joint top 1 percent, their wealth share of 21% is close to wealth share in the marginal distribution (23%), whereas their income share of 3% is slightly less than half of the size of that in the marginal distribution (8%).

3.1.3. Summary: Top wealth, income, and joint wealth and income

In view of the results of this dependence analysis, we conclude that it is important to extend the rich groups from two to three: namely, top wealth (W), top income (I), and those being in the top of wealth and income simultaneously (W+I). The copula has already revealed that the top 1 percent W+I group is highly influential as they capture about 91% of the wealth that accrues to the top 1 percent W group. In subsequent analyses we show that the top 1 percent W+I group not only shows much greater wealth concentration, but also differs systematically in terms of firm inheritances and entrepreneurship. In line with the literature and the public debate, we continue to focus on the top 1 percent. Throughout, we contrast the results with those in the top 10 percent but not the top 1 percent, that is, the "Top 10-1" Percent groups.

3.2. A descriptive view at the top: Who are the rich?

We exploit the depth of information in SOEP+P to examine the principal characteristics of "the rich". The established literature to date could not do so for lack of data. In a first step, the present section provides an informal analysis of *descriptors*. In a second step, Section 4 pursues a formal classification analysis in order to identify the top *predictors* of membership in the top 1 percent groups.

Our covariates can be divided into four groups: socio-demographics, personality items, labor market related variables, and an area in which our data are unique, measures of intergenerational transfers (gifts and inheritances). Below, we will use the terms intergenerational transfers and inheritances interchangeably. More specifically, (i) the demographics include age, sex, years of schooling,[15] a marriage dummy, the number of children, and a dummy for growing up in East Germany. (ii) Regarding personality, SOEP+P contains what are referred to as the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) (McCrae and Costa Jr 1997; John et al. 2008) and a survey measure of risk tolerance (Dohmen et al. 2011). The Big Five personality traits have been shown to be relevant for entrepreneurship (Caliendo et al. 2009) and to be markedly different for high-wealth individuals (Leckelt et al. 2022). Risk

[15] Years of schooling is the number of years required to complete the highest level of education that is recorded for the respondent. See SOEP Group (2021) for details on the generation of this variable.

tolerance has been shown to matter for entrepreneurial survival (Caliendo et al. 2010) and entrepreneurial investment (Fossen 2011; Fossen et al. 2020) and to vary strongly across the wealth distribution (Leckelt et al. 2022). (iii) Labor-market-related variables are past labor market experience in years (full-time, part-time, and unemployment) and self-employment status. In the context of our study, which focuses on the rich, self-employment often covers firm ownership and entrepreneurship. Finally (iv), measures of intergenerational transfers (gifts and inheritances) are explicitly split between transfers of closely held company assets, or in short, firm inheritances (" I1k_firm"), and other transfers ("I1k_other"), which include cash, financial assets, tangibles, and real estate. Transfers are in thousands of euros and capitalized using the CPI-adjusted bond rates provided by Jordà et al. (2019).[16] Firm inheritances are likely to play an important role in the transmission of wealth status since they have preferred inheritance tax treatment in Germany. Further, inheriting a firm provides not only wealth but also opportunities for generating income. Our analysis below will focus on this.

Up to now, we have used the household as the unit of our analysis to ensure that our results are readily comparable with the literature. To maintain the internal consistency of our analysis, we therefore report individual characteristics for the household head. Alternative units of analysis have little impact on the qualitative results, as our robustness checks in Appendix E.1 show.[17]

3.2.1. Intra- and inter-group comparisons of descriptors

Table 4 reports the mean of the covariates of each rich group. Here, W refers to a wealth group, I to an income group, and $W+I$ refers to those jointly in wealth and income groups.

The rich groups vs. the non-rich groups. As a benchmark, we have also included everyone not in the top 10% group, labeled the bottom 90%. It is evident that members of this latter group are, compared to the rich, significantly less well educated, tend not to be self-employed, have less stable labor market histories, and are less tolerant of risk. Most importantly, the incidence of intergenerational transfers is considerably lower, as are their mean values.

[16] Full details on capitalization are given in Appendix C.

[17] The household head is defined as the person who completes the household questionnaire. This is usually the household member with the most detailed knowledge about household affairs. We have re-run our analysis in Appendix E, alternatively, (i) using the sample of household heads and their partners, and (ii) changing the unit of analysis to the individual, thus also using individual labor income. Neither change of unit suggests important differences from our main results.

Table 4: The rich: Descriptors

	Top 1%			Top 10-1%			Bottom 90%		
	W	I	W+I	W	I	W+I	W	I	W+I
	Demographics								
Age	60.15	54.03	55.75	62.95	51.54	55.43	56.19	57.42	56.90
Female	0.26	0.31	0.22	0.36	0.39	0.28	0.53	0.53	0.52
SchoolYrs	14.33	15.40	15.02	13.82	14.60	14.81	12.29	12.20	12.36
Married	0.67	0.82	0.74	0.67	0.79	0.82	0.43	0.42	0.45
NumChildren	0.40	0.59	0.49	0.30	0.56	0.56	0.31	0.28	0.30
East_Soc	0.03	0.09	0.03	0.06	0.11	0.07	0.20	0.20	0.19
	Personality								
Risk_Tol	5.94	6.06	6.48	5.22	5.28	5.46	4.85	4.85	4.87
B5_Open	0.08	0.12	0.35	-0.04	0.02	-0.06	-0.06	-0.07	-0.06
B5_Cons	-0.06	-0.04	-0.10	-0.01	0.09	0.06	0.00	-0.01	-0.00
B5_Extra	0.06	0.07	0.15	-0.06	0.04	0.11	-0.06	-0.07	-0.06
B5_Agree	-0.18	-0.09	-0.24	-0.14	-0.12	-0.20	0.03	0.03	0.02
B5_Neuro	-0.33	-0.38	-0.40	-0.20	-0.28	-0.36	-0.01	0.00	-0.02
	Labor Market and Income								
SelfEmp	0.51	0.52	0.64	0.17	0.16	0.29	0.05	0.05	0.05
TopManag	0.03	0.04	0.05	0.01	0.01	0.02	0.00	0.00	0.00
CivServ	0.02	0.03	0.05	0.06	0.10	0.10	0.03	0.03	0.03
ExpFT	28.34	23.82	25.97	26.81	21.66	25.59	20.75	21.33	21.23
ExpPT	3.61	3.15	2.92	4.32	3.62	3.28	4.39	4.47	4.42
ExpUE	0.13	0.12	0.16	0.36	0.22	0.13	1.48	1.50	1.41
LabInc	139,448	276,956	312,039	60,774	116,585	134,030	33,266	26,109	32,668
CapInc	84,067	97,597	197,605	15,386	10,218	22,734	2,998	3,399	3,690
	Intergenerational Transfers								
Heir	0.48	0.44	0.40	0.43	0.27	0.41	0.19	0.21	0.21
Heir_firm	0.21	0.11	0.30	0.04	0.04	0.06	0.01	0.01	0.01
I1k_firm	719.49	606.52	1945.80	18.14	35.97	62.90	1.28	5.03	5.02
I1k_other	716.16	390.33	721.11	254.83	183.47	307.31	96.80	127.70	124.43
N	663	697	317	2016	2360	1230	13514	13136	14646

Notes. Weighted means with groups defined by their household-level position in the wealth and income distributions. Sample restricted to heads of household. *W* refers to a wealth group, *I* to an income group, and *W+I* refers to those jointly in wealth and income groups. top 10-1% are the groups in the respective top 10 percent but not the top 1 percent. Bottom 90% are the groups not in the respective top 10 percent. *SchoolYrs* are years of schooling. *NumChildren* is the number of children in the household. *Heir* is a dummy for having received an intergenerational transfer, while *Heir_firm* is a dummy for having received a firm transfer. *I1k_firm* are capitalized, intergenerational transfers (gifts and inheritances) of the type business, or for *I1k_other* the types cash, financial assets, tangibles, and real estate. The means for intergenerational transfers are conditional on receiving a transfer. *SelfEmp* is a dummy for being self-employed, *TopManager* is a dummy for being a CEO or a C-level executive, and *CivServ* is a dummy for being a civil servant. *ExpFT*, *ExpPT*, and *ExpUE* are full-time, part-time, and unemployment experience in years. *LabInc* is yearly household labor income in 2019 Euros, *CapInc* is yearly household capital income. *Risk_Tol* is risk tolerance measured on an 11-point Likert scale. *B5_Open*, *B5_Cons*, *B5_Extra*, *B5_Agree*, *B5_Neuro* are the z-standardized Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. *Source*: SOEP+P.

Top 1 percent vs. top 10-1 percent. Starting with demographics and personality traits, across each rich group, the top 1 percent are on average more risk tolerant, more educated, and less likely to have grown up in the East. Furthermore, the top 10-1 percent

income group is younger than the respective top 1 percent group, suggesting that the life-cycle appears relevant for the remaining differences in demographics.

Turning to labor market characteristics, the top 1 percent groups have a much larger incidence of self-employed/entrepreneurs (e.g., 51% vs. 17% in the W group and 64% vs. 29% in the W+I group). Capital income is a related stand-out feature: mean capital income of 197,000 euros for the top 1 percent in W+I group exceeds by a factor of at least two that of the other top 1 percent groups, and by a factor of nine that in the top 10-1 percent group. Entrepreneurship, however, is not the exclusive feature of top 1 percent membership. Mean labor market earnings indicate that many in the top 1 percent are top earners, often in dependent employment, as the means are more than twice as large in the top 1 percent group relative to the top 10-1 percent group (e.g. for the W+I group 312,000 vs. 134,000 euros, compared to 32,000 euros in the bottom 90 percent group). This importance of earnings is consistent with our dependence analysis of Section 3.1 above.

A more detailed look at inheritances reveals firm inheritances to stand out as a further "separator" between the top 1 percent and top 10-1 percent. For instance, 30% in the top 1 percent W+I group have received such inheritances, compared to only 6% in the top 10-1 Percent W+I group. However, not all firm inheritors are automatically part of the rich groups. For instance, among all firm inheritors, 39% are in the top 1 percent W group, 23% in the top 10-1 percent W group, and the remaining 38% belong to the bottom 90 percent W group. The mean value of firm inheritances the top 1 percent W+I group is about 2.7 times larger than for the top 1 percent W group, and 30 times larger than in the top 10-1 percent W+I group. The mean values of other inheritances in the top 1 percent groups tend to be about three to two times as large as for the respective top 10-1 percent groups. We conclude that a) the characteristics of the three top 1 percent groups are systematically different from the remaining population, that b) the three top 1 percent groups also differ from each other; and c) that no single characteristic in isolation can explain top rich group membership. Instead, the top 1 percent groups include inheritors, capitalists, entrepreneurs, as well as top managers. The incidence and mean values of inheritances also sets the top 1 percent apart from the top 10-1 percent (and of course the bottom 90%). Qualitatively, this might not surprise, but quantitatively the differences are remarkable. For instance, the mean value of firm inheritance for the top 1 percent W+I group is 1,945,000 euros compared to 62,000 euros in the respective 10-1 percent group, and the incidences are 30% vs. 6%. The mean values of other inheritances are considerably smaller.

4. Routes to the top: Predicting rich group membership

Which variables or variable combinations best predict top rich group membership and so might indicate the routes taken to reach the top? Our descriptive analysis has concluded that this question cannot simply be reduced to a single (set of) predictor(s); there are no sufficient statistics.

Therefore to answer our question, we deploy state-of-the-art *nonparametric* statistical learning techniques[18] that are designed to fit complex data relationships, in particular non-linearities and covariate interactions, without simply overfitting and perform well on not-yet-seen data. blue Our statistical approach, detailed in Section 4.1, enables us to identify essential correlations through sophisticated resampling and cross-validation strategies, which is an important complement to (often infeasible) causal inference, and can help direct the search for appropriate natural experiments. In order to *interpret* the role of the top predictors we use model-agnostic and model-specific importance metrics (variable importance scores, partial dependence plots, accumulated local effects, counterfactual simulations). Presented in Section 4.2 below, these metrics paint a coherent picture of the top predictors for rich group membership. Section 4.3 summarises our empirical results.

4.1. Classification trees and random forests

We use Random Forests (RFs), an ensemble technique that averages across many classification trees for the purpose of variance reduction. Since this approach is fairly new in economics, we start with a brief primer on the subject (where we also demonstrate that approach clearly outperforms classic parametric logit modelling).

A primer on classification trees. We start by explaining how to grow a single classification tree and optimally prune it to avoid overfitting. Figure 2 depicts these for the top 1 percent groups. For simplicity and ease of interpretation, consider panel (a), the top 1 percent in the joint wealth and income (W+I) group. The classification tree algorithm implements binary splits of data. A split of the data produces a node, and at each node a single predictor is used to partition the data into two homogeneous groups. Splitting is straightforward for categorical data (such as the self-employment indicator), whereas continuous data is discretized in a data-dependent manner (such as receiving an inheritance valued at more than 1.9 million euros or not). The procedure selects the best threshold value for this discretization. At each potential node, the dissimilarity of the

[18] See e.g. Friedman et al. (2001) or James et al. (2021) for textbook treatments of statistical learning.

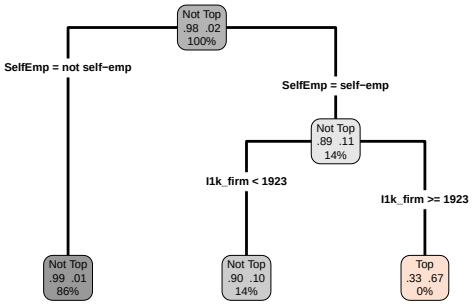
sample is computed using the Gini impurity measure.[19] The goal of the algorithm is to minimize the weighted sum of dissimilarity measures. At each node, the selected predictor is the one with the largest reduction of dissimilarity. Hence the tree is *hierarchical*, and produces a ranking of predictor importance (for instance, the self-employment indicator has the largest importance). The classification algorithm terminates when further splits would not reduce dissimilarity sufficiently.

A tree grown in this manner may be complex, thus risking overfitting the data. Hence, for predictive power, it is important to prune back the tree by recursively eliminating the least important splits. Specifically, this is done using a *cross-validation* procedure.[20] All trees presented in Figure 2 are optimally pruned in this way. A single node reports the overall population share in the line 3, the partition into not-top / top in line 2, and the dominant group for this node in line 1. For instance, in Panel (a) for node 2, 86% of the sample are not self-employed, and of these 99% are not in the top 1 percent W+I group; its complement is node 3 with 14% of the sample of which 11% are in the top 1 percent W+I group.

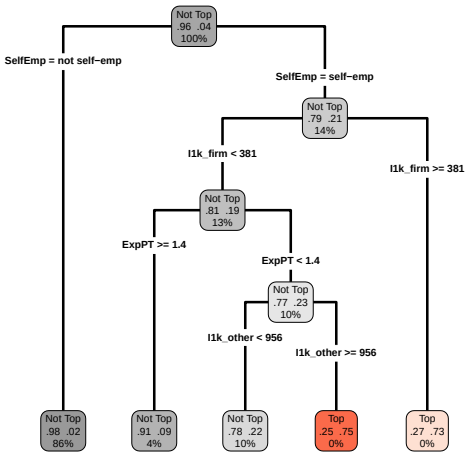
[19] The Gini impurity is one minus the sum of the squared probabilities of class occurrence with a given node. Thus, if a node consists of only one class, the Gini impurity is equal to zero.

[20] We use 10-fold cross validation, and stratified sampling to address the inherent group imbalances. In each iteration (fold), the data are randomly split into a training set used for estimation and a hold-out set used for prediction. Model risk is assessed by the proportion of observations misclassified, and this is averaged across all 10 folds; an observation is predicted to be in the top group if the predicted probability exceeds .5. The complexity of a tree depends on how much dissimilarity reduction the modeler is willing to permit. Define the so-called complexity parameter as the required minimal dissimilarity reduction at each split in a tree. The pruning algorithm will optimize over this complexity parameter using a grid search. We start by setting a smallest value for the complexity parameter. We produce several trees associated with several values of the complexity parameter up to this minimum, and calculate the cross-validation error associated with each tree. Finally, we choose the tree that is associated with the smallest cross-validation error.

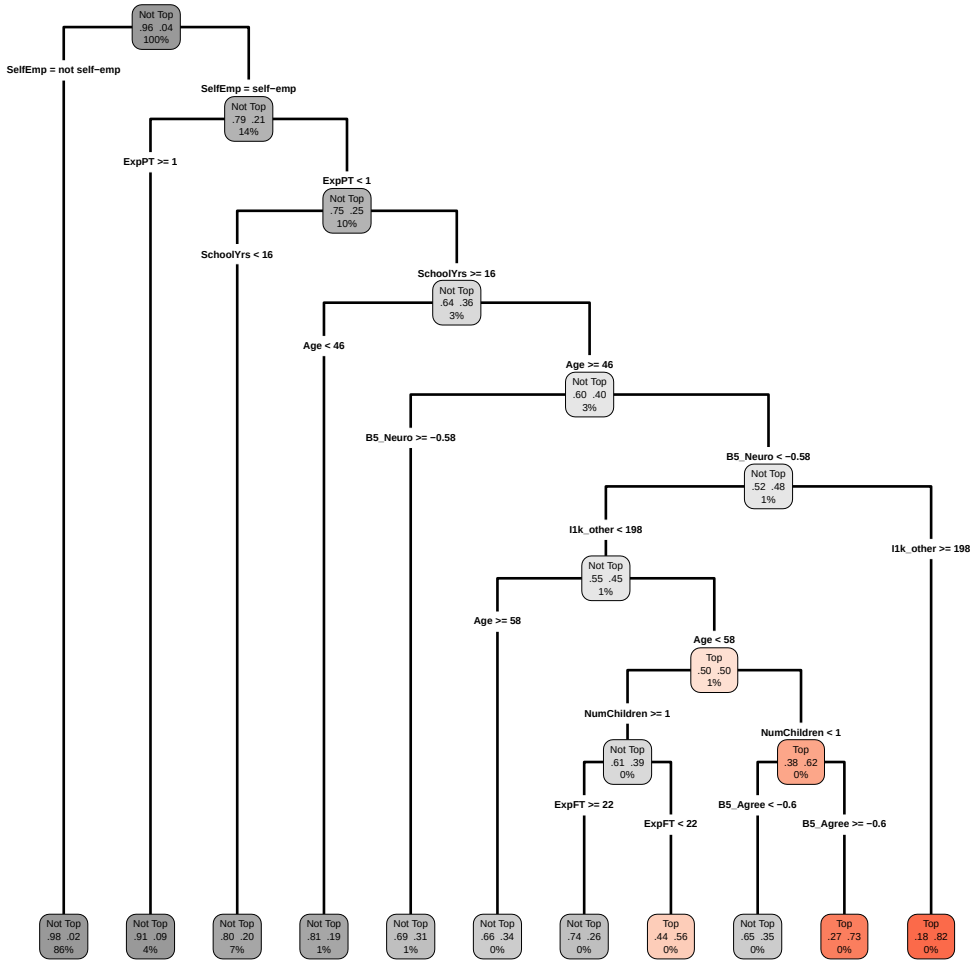
Figure 2: Single optimally pruned classification trees for top 1 percent group membership



(a) Wealth and Income



(b) Wealth



(c) Income

Notes. Optimally pruned classification trees for being in the top 1 percent of: wealth and income (panel a), wealth (panel b), and income (panel c). The pruning algorithm is explained in footnote 20. A single node reports the overall population share in line 3, the partition into not-top / top in line 2, and the dominant group for this node in line 1. As group membership in the top 1 percent is determined from weighted data, observed frequencies in the top node are not necessarily 99% and 1%. Variable definitions are given in Table 4. *Source*: SOEP+P.

The ease of interpretation of the trees follows from their *hierarchical* structure and makes them a powerful heuristic device. Consider the top 1 percent W+I group: Despite the many covariates entering the initial model for the tree, the optimally pruned tree has a very simple structure: The only top predictors selected are the self-employment indicator and firm inheritances exceeding 1.9 million euros; all other predictors have been pruned away. Of these self-employed inheritors of large fortunes 67% are in the top 1 percent group. Note that classic linear estimation practice would not uncover this joint relationship. For the top 1 percent of wealth (panel (b)), the top two predictors are again self-employment and firm inheritances although the threshold level is 0.38 million euros. For the self-employed with lower transfers, previous part-time labor market experience has a disqualifying effect. The trees for top 1 percent W and W+I share their simplicity, and the ordering of self-employment and firm inheritances. The principal difference is in the threshold level for inheritances, making the top 1 percent W+I indeed a group apart. Finally, the tree for those in the top 1 percent of income has, at first glance, a more complex structure because it has many more nodes. However, its hierarchical nature also suggests simple narratives, in particular about how people may fail to reach the top: not being self-employed (86% of the sample), or by being self-employed but having past non-full-time labor market experience (4%) or a lack of education (7%). A route to the top for the older educated self-employed is again inheritances (now exceeding 0.198 million euros).

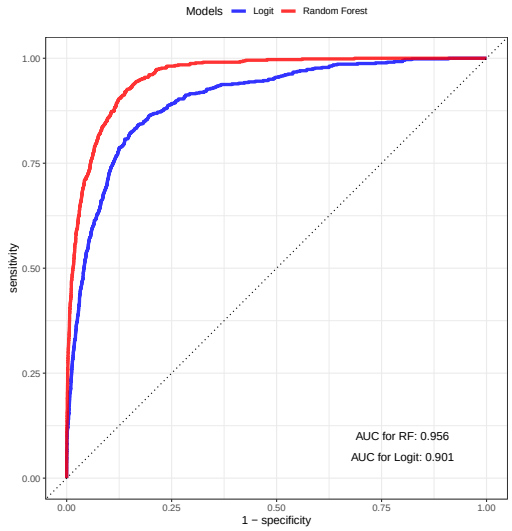
Random forests. A single classification tree is appealing because of its interpretability, but the discretization method for continuous variables may result in instability. A small change in the data could cause a large change in the estimated tree. Random forests (RFs) overcome this well-known sensitivity problem and typically improve prediction accuracy by building a large number of de-correlated deeply grown classification trees on bootstrapped training samples. Each time a split in a tree is considered, a random sample of m predictors are chosen as split candidates. This m is a parameter of the algorithm, and optimized using grid-search over cross-validation samples (using as before 10-fold stratified cross-validation) based on the usual AUC metric (see below for an explanation). The random sampling of predictors de-correlates individual trees in the forest. Averaging across this ensemble of trees, the *variance* of the estimator is reduced substantially.[21] Consequently, "random forests do not overfit" (Breiman, 2001). We

[21] To see the principal insights, recall that for $X_i \sim (\mu, \sigma^2)$ with $cor(X_i, X_j) = \rho$ and $i = 1, .., N$, the variance of the sample mean is $var(\bar{X}) = \rho\sigma^2 + \frac{1-\rho}{N}\sigma^2 \to \rho\sigma^2$ as $N \to \infty$. The bias of a RF is the same as the bias of any of the individual sampled trees, which is minimised by growing trees

discuss the empirical results based on our variable importance metrics in Section 4.2 below.

Random forests vs. logits. For completeness, we briefly demonstrate how the RF outperforms classic parametric logit models (details for the logit models and results for all rich groups are collected in Appendix F), focussing here on prediction error. For top 1% of wealth, and a given arbitrary threshold for positive classification (here 0.24), Table 5 reports a confusion matrix. The logit correctly predicts 92.93% of the true negatives and 1.74% of the true positives, while it misclassifies 5.34% of the sample. The RF correctly predicts 94.36% of the true negatives and 2.10% of the true positives, while it misclassifies 3.53% of the sample. Hence, the RF is able to both predict more true positives and negatives, but especially makes less mistakes with respect to false positives.

Figure 3: ROC curve for top 1 percent of wealth (RFs vs. logits)



Notes. AUC is the area under the ROC curve. A higher AUC implies greater predictive power. See the main text for a description of the ROC. At every classification threshold the random forest ROC curves lies above the logit ROC curve, indicating that the random forest model correctly predicts more true positive cases no matter the threshold. See Appendix F for a detailed analysis for all rich groups. *Source*: SOEP+P.

The confusion matrix is only illustrative, since it is based on an arbitrary probability threshold for binary classification. By letting the threshold value range from 0 to 1 and plotting the resulting shares of true positive and true negatives,[22] we obtain the ROC

deeply. The covariance between any two trees is reduced by the de-correlation trick, thus reducing the generalisation error of the ensemble relative to a single tree.

[22] More specifically, "sensitivity" is plotted against 1-"specificity," where sensitivity is the sample analogue of $\Pr(T = 1|T = 1)$, or "true positives," and specificity is the sample analogue of $\Pr(T = 0|T = 0)$, or "true negatives," where T denotes the binary group indicator.

Table 5: A confusion matrices for top 1 percent wealth group

		Logit		Random Forest	
		Prediction		Prediction	
		0	1	0	1
Data	0	92.93	2.87	94.36	1.43
	1	2.47	1.74	2.10	2.10

Notes. "1" indicates membership in the rich group, while "0" indicates the opposite. The arbitrary classification threshold is 0.24 (predicting "1" if the predicted probability exceeds this threshold). Observations are not weighted. As group membership in the top 1 percent is determined from weighted data, observed data frequencies are not equal to 99% and 1%. *Source*: SOEP+P.

(receiver operating characteristic) curve depicted in Figure 3. At every classification threshold the RF's ROC curves lies above the logit ROC curve, indicating that the RF correctly predicts more true positive cases irrespective of the threshold. Finally, a global threshold-invariant measure of predictive performance obtains by integrating the ROC, yield the AUC (area under the curve). A higher AUC implies greater predictive power, and the RF (AUC=.956) clearly outperforms the logit (AUC=.901). In the next section, we change the perspective from global measures of predictive performance to variable-based metrics.

4.2. Results: Identifying the best predictors using variable importance measures

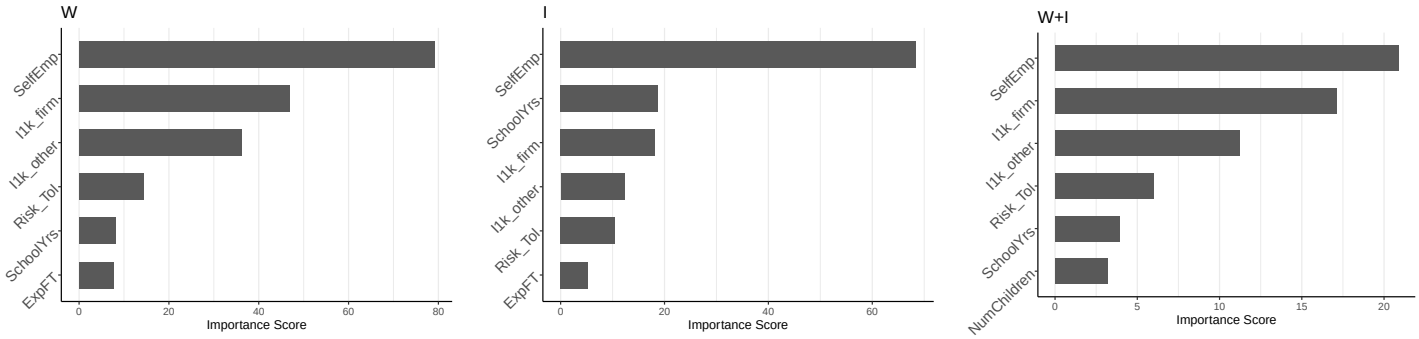
In order to identify the key predictors as well as explain and interpret the results of the RF we compute several variable importance metrics, which can be either model-specific or model-agnostic. The key empirical insight will be that these metrics rank the key predictors coherently.

Variable importance scores (VIMP). Our first metric is based on the node splitting criterion for the individual classification trees making up the RF (the Gini measure). We compute the importance score for each predictor in a tree and average across all trees. Consequently, this metric is model-specific.

Figure 4 reports the ordered variable importance scores for the top six predictors. The rapidly decreasing importance scores in each panel indicate that despite the many covariates fed into the classification model, only a small number are important predictors, and, crucially, the set of the five most important predictors shows little variation across the rich groups. Overall, the hierarchy of the predictors for the single trees broadly harmonizes with the hierarchy for the random forests. More specifically, for all the top 1 percent groups, self-employment/ entrepreneurship is the most important predictor,

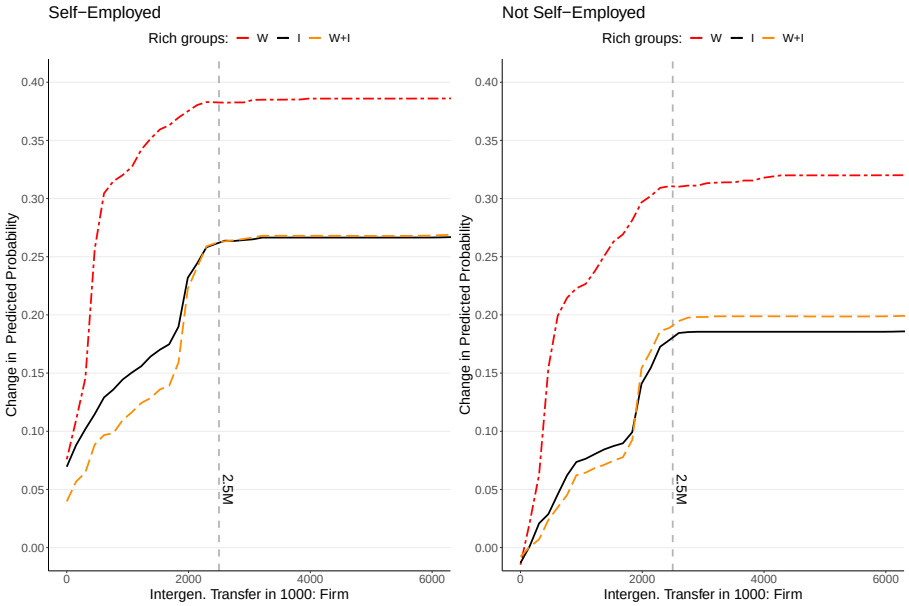
followed by firm inheritances. Ranked third for top wealth and the joint top group is other inheritances, and for the top income group it is education. Ranked fourth for the wealth and the joint group is risk tolerance and for the income group it is other inheritances. Turning to an assessment of relative predictor importance for a group, we see that, when compared to self-employment, firm inheritances play a much greater importance for the top 1 percent W+I group compared to the top 1 percent income group. We conclude that self-employment/entrepreneurship is the key predictor for reaching one of the top 1 percent groups. Further, the likelihood of reaching the top 1 percent substantially increased by firm inheritances, particularly for the W+I group. Education is comparatively less important and only seems to be significant for the top income group. In Appendix E we show importance scores for the top 10-1 percent groups. For these groups, self-employment/entrepreneurship plays an important role, but for some groups (especially the income groups) predictors such as other inheritances and education are even more important. Firm inheritances do not appear among the top three predictors for these groups. Hence, the strong connection between self-employment and firm inheritance is not found for these groups.

Figure 4: Key predictors: Variable importance scores (VIMPs) for top 1 percent group membership

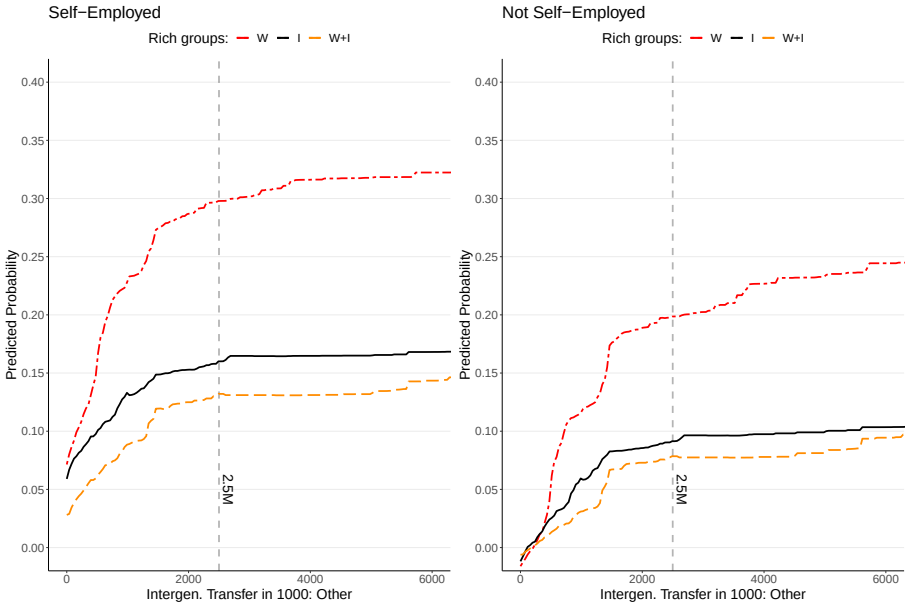


Notes. Estimations and predictions using random forest models. The importance measures are based on the Gini impurity measure, which has been corrected for the scale of the variables. Variable definitions are given in Table 4. *Source*: SOEP+P.

Figure 5: Partial dependence plots for top 1 percent: Inheritances and self-employment



(a) Firm Inheritances



(b) Other Inheritances

Notes. Partial dependence plots normalized by the base probability. Shows the change in base probability depending on the value of firm inheritances and self-employment for the prediction of being in the top 1 percent group of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

Partial dependence plots (PDPs). The partial dependence function (Friedman et al. 2001) is the basis for assessing predictor importance in a model-agnostic way by focussing

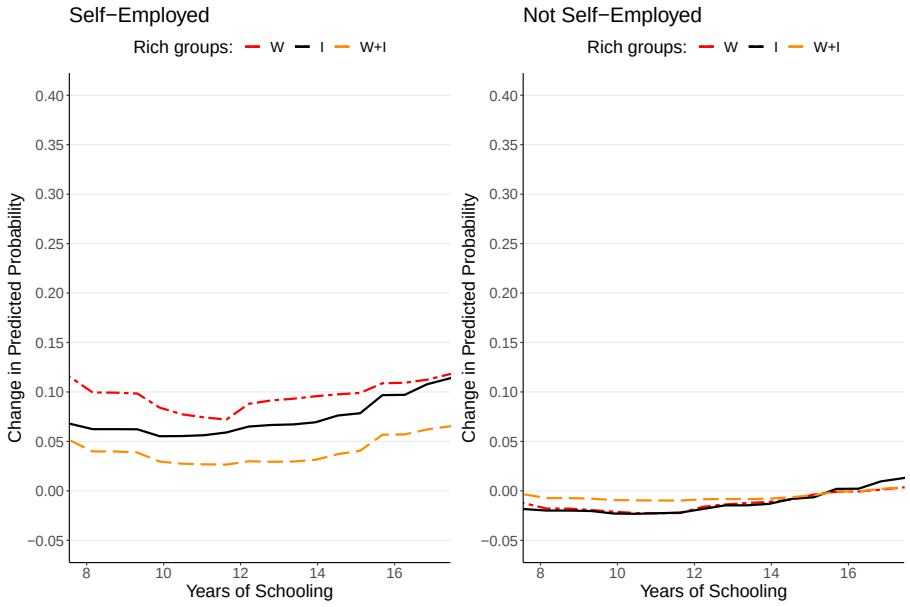
on average marginal effects on outcomes, i.e. the probability of being in a rich group.

The empirical partial dependence function with respect to predictor x_k is

$$PDP(x_k) = \frac{1}{N} \sum_{i=1}^{N} \hat{f}(x_k, X_j^i)$$

where the predicted probability of belonging to a rich group $\hat{f}(.)$ depends on the value of the predictor x_k and the other variables X_j^i whose values are not changed during the calculation.

Figure 6: Partial dependence plots for top 1 percent: Years of schooling and self-employment



Notes. Partial dependence plots normalized by the base probability. Shows the change in base probability depending on the value of years of schooling and self-employment for the prediction of being in the top 1 percent of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

Thus, $PDP(x_k)$ gives the sample average of the predicted probability of belonging to a rich group if the value of one variable is changed, while all others stay the same. To ease interpretation, we normalize the PDP by subtracting the base probability (e.g., 1% for the top 1 percent W group) from the PDP in the figures we show below, so that one can interpret them in terms of a change from the base probability.

Figure 5 shows the normalized partial dependence plot for the top 1 percent groups when calculating the PDP with respect to self-employment and a) firm inheritances or b) other inheritances. In the upper panel, we show the PDP for firm inheritances depending

on being self-employed or not. If the marginal effects are predicted conditional on self-employment (upper left graph), the predicted change in inclusion probability is positive even at zero firm inheritances, showing that entrepreneurship in itself confers a benefit with respect to the inclusion probability. The inverse of this benefit can be seen on the right side of the figure: Starting with zero firm inheritances and not being self-employed actually decreases the base inclusion probability. Whether the prediction is conditional on self-employment or not, one can see that along the distribution of firm inheritances, predicted inclusion probabilities for all three groups rise sharply until about 2.5 million euros after which the curves become essentially flat. Conditional on self-employment the change in inclusion probability rises from 7 percentage points (pp.) at zero firm inheritance to 38 pp. at 2.5 million for the top 1 percent W group, and from 7 pp to 26 pp for top 1 percent I group, and from 4 pp to 26 pp for the Top 1 Percent W+I group. Conditional on not being self-employed, the increases are much smaller: the changes in the predicted probability plateau at 31 pp. for W, 18 pp. for I, and 19 pp. for W+I. Turning to the lower panel, one can see that even when one is self-employed, receiving inheritances other firm inheritances confers far smaller benefits with respect to inclusion in a top rich group. The left lower panel shows that the change in the predicted inclusion probability rises to only 30 pp for W, 16 pp. for I, and 13 pp for I+W. Thus, receiving an inheritance of 2.5 million euros that is not a firm only confers about half the benefit of a firm inheritance of the same size with respect to the probability of being included in the joint top 1 percent. Thus, one can see that it is the *combination* of self-employment/entrepreneurship and sizable firm inheritances that determines the likelihood of being included in one of the top rich groups.

Figure 6 contrasts these previous results with normalized PDPs for education. Conditional on being self-employed or not, we show the change in predicted probabilities for the three groups along years of schooling. The trajectories, on both the left-hand and the right-hand side, are comparatively very flat. There is a slight increase on the left-hand side for the I group from about 5 pp to about 10, but generally an increase in years of schooling does not change the predicted probability nearly as much as self-employment or inheritances do. For those not in self-employment, the trajectories are even flatter. By comparison, these PDPs for education show just how important self-employment and firm inheritances are for the predicted probabilities.

Appendix E provides analogous partial dependence plots for the top 10-1 percent groups. The figures show that the respective changes in the predicted probabilities due to firm inheritances are much smaller than those for the top 1 percent groups. The gradient is more relevant with respect to education for all top 10-1 percent groups, especially the in-

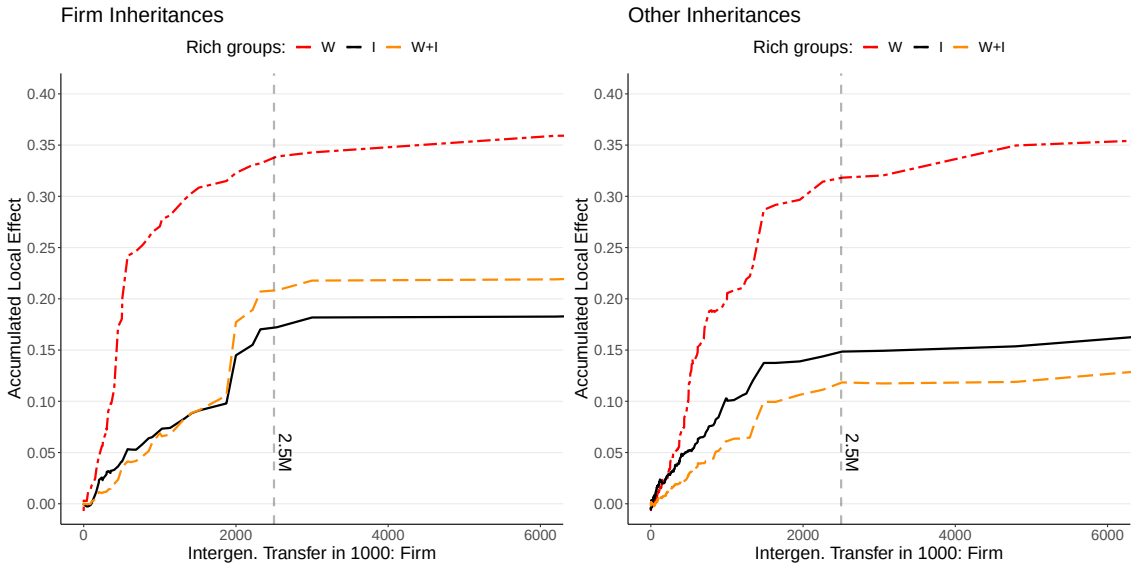
come group.

Accumulated local effects (ALEs). An alternative model-agnostic way to measure marginal effects is to use accumulated local effects. These take potential correlation between predictors into account by calculating and accumulating the local changes in predicted probabilities. ALEs are helpful because they use the values of predictors in a neighborhood of the value of the predictor being examined and can therefore avoid unrealistic variable combinations. The population ALE for predictor X_j at value x_j is defined (Apley and Zhu 2020) as the centered accumulated expected change

$$ALE(x_j) = \int_{z_{0,j}}^{x_j} E\left[\frac{\partial f(X_1, ..., X_d)}{\partial X_j} | X_j = z_j\right] dz_j - c_1$$

where, for ease of exposition, we have assumed that the probability of belonging to a rich group $f(.)$ is differentiable and predictor X_j is continuous. $z_{0,j}$ is the lower bound on X_j and c_1 is a centering constant ensuring that the ALE has mean zero with respect to the marginal distribution of X_j. The estimator is its sample analogue, where in practice we average within neighborhoods of $X_j = z_j$. For instance, considering the key predictor firm inheritances ("I1k_firm"), the ALE at an inheritance of 2.5 million for the top 1 percent wealth group is roughly .34, which means that the probability of being in the top wealth group increases by 34 pp. from the base probability.

Figure 7: Accumulated local effects for top 1 percent: Inheritances



Notes. Accumulated local effects plots of the value of inheritances ("I1k_firm" and "I1k_other") for the prediction of being in the top 1 percent of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

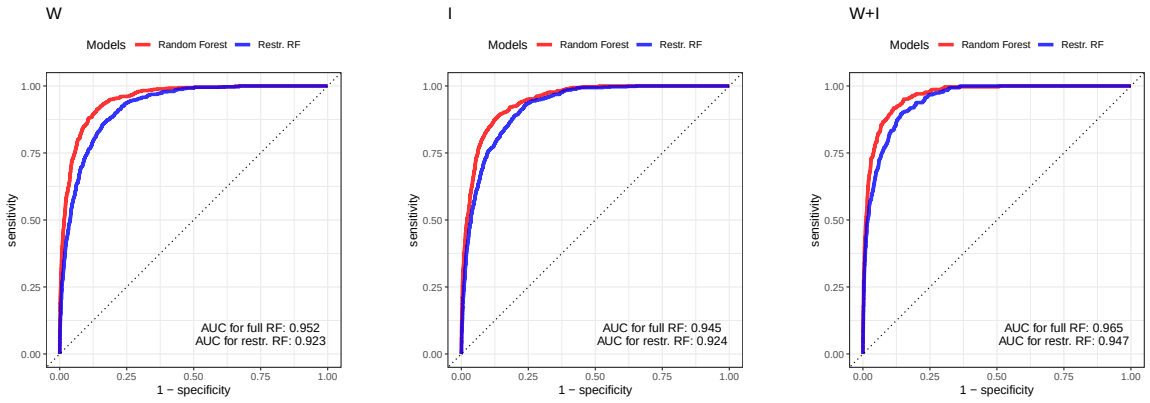
For the sake of brevity, Figure 7 presents our results only for firm and other inheritances for the top 1 percent groups. Qualitatively the graphs are very similar to the ones shown in Figure 5. However, because we do not condition on self-employment, they show slightly different levels and dynamics. For example, the ALE at 2.5 million euros firm inheritance for the top 1 percent W+I group is about 21 pp and thus slightly higher than for the top 1 percent I group, which is not the case in the analogous PDP conditioning on self-employment. This is likely due to the fact that firm inheritances and self-employment are correlated.

When we contrast firm and other inheritances, we again find the qualitative picture that emerged for the PDPs: firm inheritances confer a higher benefit in terms of the change in predicted probability. Take the W+I group as an example: The increase in the predicted probability given a 2.5 million euro firm inheritance is 21 pp, while it is about 11 pp for an non-firm inheritance of the same size. Thus, as with the PDPs, we find that other inheritances confer only about half the benefit of a firm inheritance with respect to the predicted probability of being included in the W+I group.

Appendix E contains analogous ALE plots for the top 1 percent and top 10-1 percent groups for firm and other inheritances as well as education. Again, these plots qualitatively confirm the analysis based on partial dependence plots.

Predictive performance with and without the key predictors. It is well known that some popular model-agnostic individual feature importance measures such as Shapley values can become uninformative in the presence of strong feature correlation. This problem occurs in our data for entrepreneurship and firm inheritances. Therefore, we take an alternative approach: We juxtapose the performance of two random forests which include and exclude the two most important features. We then compare the ROCs and the AUCs of this *restricted* set of trees to those of the unrestricted set of trees. Figure 8 shows the ROC curves and AUCs for all three top 1 percent groups.

Figure 8: Restricted Random Forests: ROC curves for top 1 percent groups



Notes. Shows ROC curve and AUC for restricted and unrestricted random forests. Restricted random forests omit the entrepreneurship indicator and the intergenerational transfer variables. AUC is the area under the ROC curve. *Source*: SOEP+P.

The figure shows that regardless of the rich group that is being investigated, the unrestricted, i.e. full, random forests have better predictive performance both in terms of the ROC curve and in terms of the AUC. The differences are more pronounced at low classification thresholds. The AUC measures are fairly similar, but this is to be expected as the random forest is extremely flexible, so that other predictors, like age, labor market experiences, and education, that are correlated with entrepreneurship and intergenerational transfers can compensate for the lack of our key predictors. Nevertheless, both the ROC curves and the AUCs clearly indicate that predictive performance drops markedly after restricting the set of predictors.

4.3. Discussion and Summary

Who are "the rich" in Germany? Our non-parametric analysis has revealed the multifaceted aspects of being rich, and has underscored that looking at the distribution of

income or wealth in isolation is not sufficient.

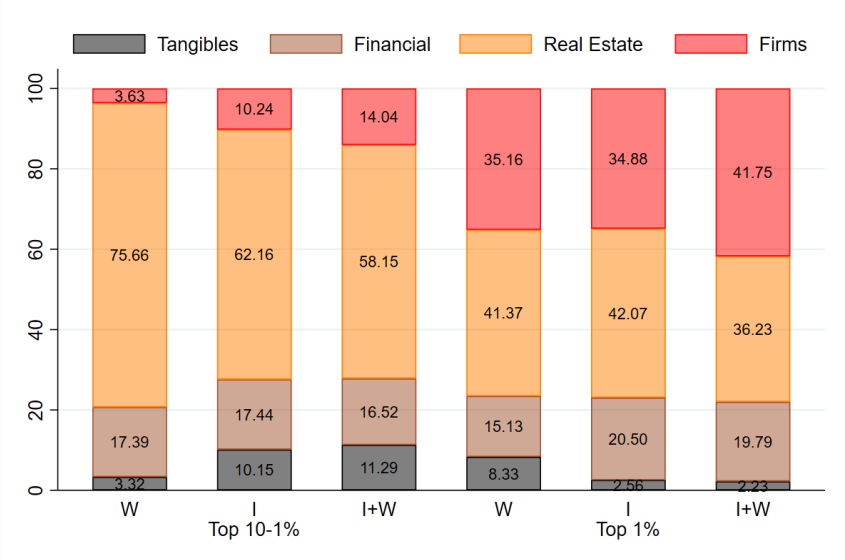
The key empirical insight from our analysis is that the large set of potential predictors for top rich group membership is reduced to a very small number of important, interacting predictors. For the top 1 percent, be it wealth, income, or both, a *combination* of self-employment/entrepreneurship and inheritance of company assets (as opposed to real estate or financial assets), lead to the highest predicted probabilities of group membership. Removing either factor leads to drastically smaller predicted probabilities, especially for the joint top 1 percent. Conversely, other covariates are not nearly as important. Our inter-rich group comparison among the top 1 percent has also highlighted the essential differences that set the members of the joint top 1 percent apart from the as a *class of intergenerational entrepreneurs*.

Entrepreneurship among the joint top 1 percent. The joint top 1 percent group stands out from the other top 1 percent groups not only with respect to the classification exercise. Aside from holding 21% of all wealth, the group appears fairly homogeneous: Members tend to be predominantly prime-aged entrepreneurs and owner-managers who have benefited from sizable inheritances, and in particular firm inheritances. Their portfolio is also markedly different from that of the marginal top 1 percent groups: 42% of their gross wealth is held in the form of closely held businesses (unlike in the other rich groups). For these reasons we see the joint top 1 percent as an entrepreneurial group that is set apart from the other groups.[23] We also note that 56% in the joint top 1 percent consider themselves to be "self-made"[24], despite the received intergenerational transfers (see Appendix Table D.3 for details). This self-perception presumably stems from the fact that they have grown their fortunes to such an extent that their wealth-to-inheritance ratio, being 0.25, appears small compared to the ratio for the marginal top 1 percent groups, which are 0.34 for for wealth and 0.38 for income.

[23] In particular, they are not the "millionaires next door" (Stanley 1996) who happened into top wealth, for example, because of advantageous regional developments in land prices (Kholodilin et al. 2018).

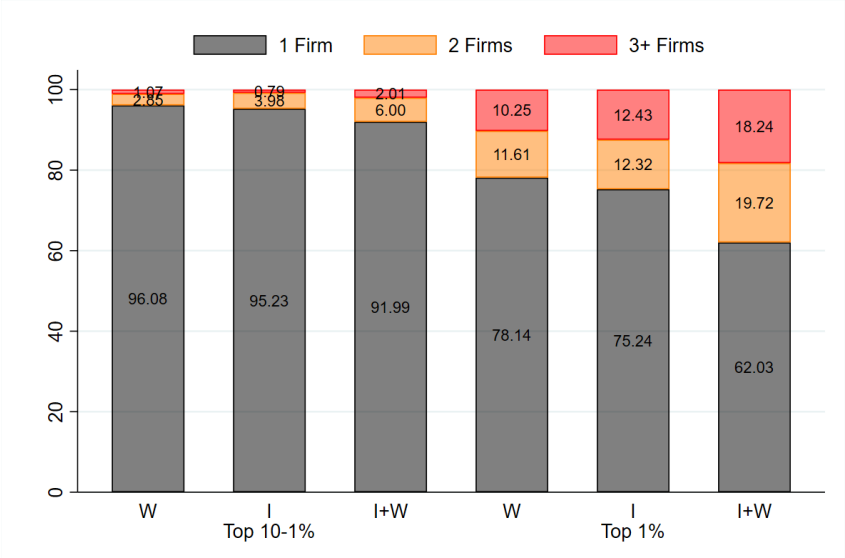
[24] Being self-made here refers to the individual considering self-employment and entrepreneurship as the main determinant of their current wealth as opposed to other sources, such as gifts or inheritances. For a detailed definition see section C.4.

Figure 9: Gross wealth shares by asset class and rich groups



Notes. Composition of the gross wealth portfolio within each of the rich groups. The shares in percent give the aggregate contribution to the wealth total within each group. Shares were computed using household survey weights. Tangibles are objects of high value such as paintings, jewelry, cars, etc. Financials are stocks, bonds, currency, insurance contracts, and private pensions. Real estate is owner-occupied and other real estate. Businesses are the value of solely or partly held private businesses, that is, closely held firms. *Source*: SOEP+P.

Figure 10: Firms in sole ownership by rich groups

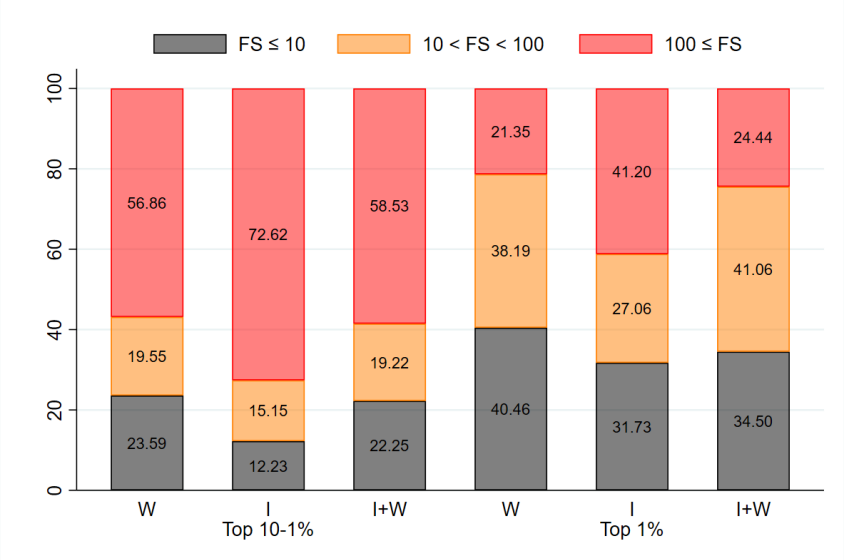


Notes. Shows shares in percent of those in a given category of number of firms owned conditional on owning at least one firm for each of the six rich groups. Shares were computed using household survey weights. *Source*: SOEP+P.

This systematic difference between the joint top 1 percent and the other rich groups is

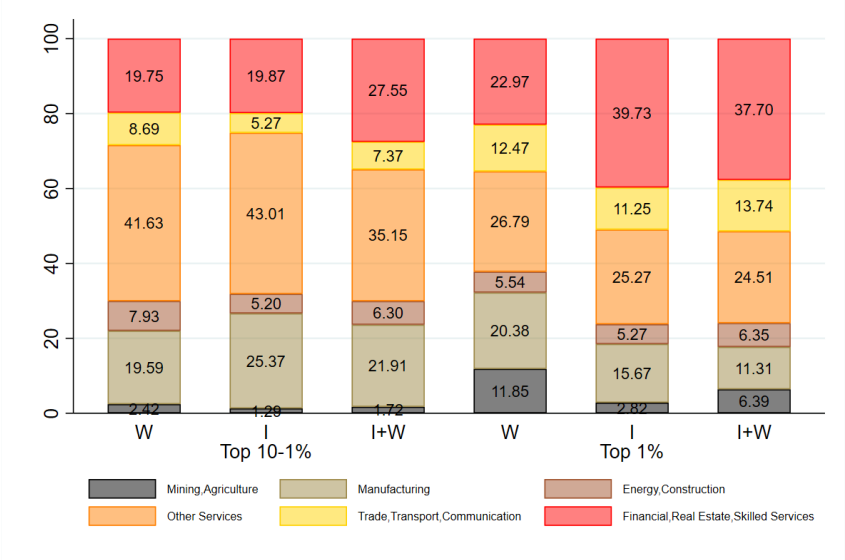
illustrated further in Figure 9, where we examine the composition of gross wealth. Firm assets constitute the predominant form of wealth in the joint top 1 percent group (42%), while the share of real estate (36%) is the lowest among all rich groups. The joint top 1 percent group exhibits a systematically different entrepreneurial focus, which likely has its origin in the very large firm inheritances they receive. We note that we find very similar asset compositions among the top 10-1 percent groups , and a real estate share of over 50%. By contrast, this share declines systematically in the top 1 percent groups (with 42% in the I, 41% in the W, and 36% in the W+I group). Figure 10 depicts the incidence of firms in sole ownership, and reveals that across all groups, the rich predominately own one firm.

Figure 11: Firm size categories by rich groups



Notes. Shares in percent of those within a firm size category for each of the six rich groups. Sample is conditional on being currently active on the labor market. Shares were computed using household survey weights. *Source*: SOEP+P.

Figure 12: Industry composition by rich groups



Notes. Shares in percent of those within a certain industry category for each of the six rich groups. Sample is conditional on being currently active on the labor market. Shares were computed using household survey weights. *Source*: SOEP+P.

Finally, in Figures 11 and 12, we examine the sizes of firms and the industries the rich are economically active in, be it by dependent or self-employment. About 38% of the joint top 1 percent work in financial and other skilled services and these firms tend to be small as more than 75% work in firms with less than 100 employees. This picture is reversed for the top 10-1 percent group members, where the majority work in large firms (over 50% are in firms with more than 100 employees); given the much lower incidence of self-employment and the small share of firms in their wealth portfolio, these members tend to be well remunerated workers in dependent employment rather than owner-managers; see Table 4. Manufacturing is more important for top 10-1 percent income members compared to the top 1 percent, and financial services less so.

Summary. Taken together, these findings suggest the following interpretation: The key predictors that help to distinguish between all the rich groups are entrepreneurship in conjunction with sizable firm inheritances. According to these variables, we can order the rich groups along an *entrepreneurial spectrum*. At one extreme of the spectrum is the top 1 percent W+I group, which mainly consists of entrepreneurs who have inherited substantial firm assets. At the other extreme of the spectrum are members of the top 10-1 percent groups who tend to be high-skilled employees in large firms. Finally, the top 1 percent of income and the top 1 percent of wealth sit in the middle of the spectrum. They

are certainly more entrepreneurial than the top 10-1 percent, but fall short of the extreme entrepreneurial focus that the joint top 1 percent exhibit.

The extreme concentration of wealth among the joint top 1 percent and the key predictors also suggest strong parallels to recent findings for the top income and top wealth in the United States. Using administrative data Smith et al. (2019) find that top income groups are not rentiers but human-capital-rich working-age entrepreneurs whose primary source of top income is private "pass-through" business profit (for tax reasons) of closely held small to mid-market firms in skill-intensive industries. More specifically, these authors estimate that up to 40% of income in the top 1 percent derives from pass-through businesses. Most top owners own just one firm, and Smith et al. (2019) consider about 2/3 of top earners as "self-made". A crucial difference with the German joint top 1 percent is that these top earners were deemed unlikely to have received large financial inheritances or inter vivos gifts. By contrast, we have shown the importance of firm inheritances in the German case. Turning to the US wealth distribution, Smith et al. (2023) find that top wealth groups predominantly hold business assets. For example, the top 0.1% hold 15% of total wealth with 60% of that share stemming from pass-through businesses and C-corporations. Garbinti et al. (2021) show that similar patterns for the rich also emerge in France. In examining the top 1 percent of wealth, they find that this group mainly holds substantial equity portfolios and that they are predominantly top capital income earners.

Importantly, despite many parallels between the rich groups in the United States and Germany, a crucial difference is the role of firm inheritances as a route to the top. This is of significant policy relevance, since such inheritances will perpetuate top wealth across generations (Kopczuk and Zwick 2020) and decrease intergenerational mobility. In the German case, firm inheritances—especially family firms—receive generous tax exemptions (Bach 2016). Hence, this form of intergenerational transmission constitutes a major force of immobility.

A central argument for giving advantageous tax treatment to (family) firm inheritances is the concern that taxing firm inheritances causes heirs to sell the firm inducing transaction costs. However, as Grossmann and Strulik (2010) illustrate, depending on the heirs' entrepreneurial abilities, there is a trade-off between these transaction costs and the efficiency costs of less capable heirs taking over the family firm. Fagereng et al. (2021) and Black et al. (2020) provide evidence that the potential genetic component of intergenerational wealth persistence—that is, the hereditary transmission of the ability to acquire wealth—is of limited importance and that environmental factors play a more important role. In particular, Black et al. (2020) show that bequests are a central determinant of the

intergenerational persistence in wealth. Thus, along the lines of Grossmann and Strulik (2010), our results not only suggest strong immobility at the top, which has implications for societal fairness judgments, they also raise the concern that this immobility leads to efficiency costs.

5. Concluding comments

In this paper, we have tackled the questions of who the rich are and what routes have led them to the top. Using state-of-the-art classification models on the new (and validated) SOEP+P dataset enables us to identify and quantify the key predictors of membership in rich groups in Germany: Entrepreneurship in combination with inheritance of closely held company assets play a crucial role in predicting top 1 percent group membership. Reflecting this, we find that the joint top 1 percent wealth and income group consists of individuals best characterized as prime-aged entrepreneurs and owner-managers who have benefited from sizeable firm inheritances and who are active in the financial and real estate sectors and skilled services. By contrast, the top 10-1 percent group members are not predominantly entrepreneurial but rather highly-skilled employees in large firms in either manufacturing or other services. Their wealth is concentrated in real estate. While the top 1 percent in Germany share many similarities with those in the United States, inheritances of company assets play a central role for the former. The implied intergenerational immobility at the top is thus partly a policy choice since inheritance tax law in Germany, as in several other European countries, exempts firm inheritances. Current debates over wealth or inheritance taxation often avoid the thorny problem of taxing closely held firm assets because of concerns about efficiency losses and, as a more technical issue, valuation of the tax base. However, if policy makers wish to seriously tackle the issue of intergenerational immobility at the top, the taxation of firm assets will have to be considered.

References

Adermon, Adrian, Mikael Lindahl, and Daniel Waldenström (2018), "Intergenerational Wealth Mobility and the Role of Inheritance: Evidence from Multiple Generations." *The Economic Journal*, 128, F482–F513, URL https://doi.org/10.1111/ecoj.12535.

Albers, Thilo N.H., Charlotte Bartels, and Moritz Schularick (2022), "The distribution of wealth in Germany, 1895-2018." *CESifo Working Paper 9739*.

Alvaredo, Facundo, Anthony B Atkinson, Thomas Blanchet, Lucas Chancel, Luis Estevez Bauluz, Matthew Fisher-Post, Ignacio Flores, Bertrand Garbinti, Jonathan Goupille-Lebret, Clara Martínez-Toledano, et al. (2021), "Distributional national accounts guidelines methods and concepts used in the world inequality database." Technical report, HAL.

Apley, Daniel W. and Jingyu Zhu (2020), "Visualizing the effects of predictor variables in black box supervised learning models." *Journal of the Royal Statistical Society Series B*, 82, 1059–1086.

Athey, Susan and Guido W Imbens (2019), "Machine learning methods that economists should know about." *Annual Review of Economics*, 11, 685–725.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez (2011), "Top incomes in the long run of history." *Journal of Economic Literature*, 49, 3–71.

Auray, Stéphane, Aurélien Eyquem, Bertrand Garbinti, and Jonathan Goupille-Lebret (2022), "Markups, taxes, and rising inequality." Technical report, CESifo.

Bach, Stefan (2016), "Erbschaftsteuer, Vermögensteuer oder Kapitaleinkommensteuer: Wie sollen hohe Vermögen stärker besteuert werden?" Technical report, DIW Berlin, German Institute for Economic Research.

Bach, Stefan, Giacomo Corneo, and Viktor Steiner (2009), "From bottom to top: the entire income distribution in Germany, 1992–2003." *Review of Income and Wealth*, 55, 303–330.

Bach, Stefan, Andreas Thiemann, and Aline Zucco (2019), "Looking for the missing rich: Tracing the top tail of the wealth distribution." *International Tax and Public Finance*, 26, 1234–1258.

Bartels, C. and D. Waldenström (2022), "Inequality and top incomes." *Handbook of Labor, Human Resources and Population Economics*.

Benhabib, Jess, Alberto Bisin, and Mi Luo (2019), "Wealth distribution and social mobility in the US: A quantitative approach." *American Economic Review*, 109, 1623–47.

Black, Sandra E., Paul J. Devereux, Fanny Landaud, and Kjell G. Salvanes (2022), "The (un)importance of inheritance." *NBER working paper*, 29693.

Black, Sandra E., Paul J. Devereux, Petter Lundborg, and Kaveh Majlesi (2020), "Poor Little Rich Kids? The Role of Nature versus Nurture in Wealth and Other Economic Outcomes and Behaviours." *The Review of Economic Studies*, 87, 1683–1725.

Boserup, Simon H., Wojciech Kopczuk, and Claus T Kreiner (2016), "The role of bequests in shaping wealth inequality: Evidence from Danish wealth records." *American Economic Review*, 106, 656–61.

Boserup, Simon Halphen, Wojciech Kopczuk, and Claus Thustrup Kreiner (2018), "Born with a silver spoon? Danish evidence on wealth inequality in childhood." *The Economic Journal*, 128, F514–F544.

Bricker, Jesse, Lisa J. Dettling, Alice Henriques, Joanne W Hsu, Lindsay Jacobs, Kevin B. Moore, Sarah Pack, John Sabelhaus, Jeffrey Thompson, and Richard A. Windle (2017), "Changes in us family finances from 2013 to 2016: Evidence from the survey of consumer finances." *Federal Reserve Bulletin*, 103, 1.

Bricker, Jesse, Alice Henriques, Jacob Krimmel, and John Sabelhaus (2016), "Measuring income and wealth at the top using administrative and survey data." *Brookings Papers on Economic Activity*, 2016, 261–331.

Brunori, Paolo and Guido Neidhöfer (2021), "The evolution of inequality of opportunity in germany: A machine learning approach." *Review of Income and Wealth*, 67, 900–927.

Bucks, Brian K, Arthur B. Kennickell, Traci L. Mach, and Kevin B. Moore (2009), "Changes in us family finances from 2004 to 2007: Evidence from the survey of consumer finances." *Federal Reserve Bulletin*, 95.

Cagetti, Marco and Mariacristina De Nardi (2006), "Entrepreneurship, frictions, and wealth." *Journal of Political Economy*, 114, 835–870.

Caliendo, Marco, Frank Fossen, and Alexander Kritikos (2010), "The impact of risk attitudes on entrepreneurial survival." *Journal of Economic Behavior & Organization*, 76, 45–63.

Caliendo, Marco, Frank M. Fossen, and Alexander S Kritikos (2009), "Risk attitudes of nascent entrepreneurs–new evidence from an experimentally validated survey." *Small Business Economics*, 32, 153–167.

Christiansen, Vidar and Matti Tuomala (2008), "On taxing capital income with income shifting." *International Tax and Public Finance*, 15, 527–545.

DeNardi, Mariacristina and Giulio Fella (2017), "Saving and wealth inequality." *Review of Economic Dynamics*, 26, 280–300.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner (2011), "Individual risk attitudes: Measurement, determinants, and behavioral consequences." *Journal of the European Economic Association*, 9, 522–550.

Drechsel-Grau, Moritz, Andreas Peichl, Kai D Schmid, Johannes F Schmieder, Hannes Walz, and Stefanie Wolter (2022), "Inequality and income dynamics in germany." *Quantitative Economics*, 13, 1593–1635.

Drees, Holger, Laurents de Haan, and Sidney Resnick (2000), "How to make a Hill plot." *Annals of Statistics*, 28, 254–274.

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch (1997), *Modelling Extremal Events*. Springer, Berlin.

Fagereng, Andreas, Magne Mogstad, and Marte Rønning (2021), "Why do wealthy parents have wealthy children?" *Journal of Political Economy*, 129, 703–756.

Fossen, Frank M. (2011), "The private equity premium puzzle revisited—new evidence on the role of heterogeneous risk attitudes." *Economica*, 78, 656–675.

Fossen, Frank M., Johannes König, and Carsten Schröder (2020), "Risk preference and entrepreneurial investment at the top of the wealth distribution." *Available at SSRN 3716271*.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001), *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2018), "Income inequality in France, 1900–2014: Evidence from distributional national accounts (DINA)." *Journal of Public Economics*, 162, 63–77.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2021), "Accounting for wealth-inequality dynamics: Methods, estimates, and simulations for France." *Journal of the European Economic Association*, 19, 620–663.

Grabka, Markus M. (2021), "SOEP-core v36-codebook for the PEQUIV file 1984-2019: CNEF variables with extended income information for the SOEP." Technical report, SOEP Survey Papers.

Grossmann, Volker and Holger Strulik (2010), "Should continued family firms face lower taxes than other estates?" *Journal of Public Economics*, 94, 87–101.

Hofert, M., I. Kojadinovic, M. Mächler, and Y. Yan (2017), *Elements of Copula Modeling with R*. Springer, Berlin.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2021), *An Introduction to Statistical Learning*. Springer series in statistics New York.

John, Oliver, Laura Naumann, and C. Soto (2008), "Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues." In *Handbook of Personality: Theory and Research, 3rd Edition*, 114–158, The Guilford Press.

Jordà, Òscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M Taylor (2019), "The rate of return on everything, 1870–2015." *The Quarterly Journal of Economics*, 134, 1225–1298.

Kholodilin, Konstantin A., Claus Michelsen, and Dirk Ulbricht (2018), "Speculative price bubbles in urban housing markets." *Empirical Economics*, 55, 1957–1983.

König, Johannes, Carsten Schröder, and E. N. Wolff (2020), "Wealth inequalities." *Handbook of Labor, Human Resources and Population*.

Kopczuk, Wojciech (2015), "What do we know about the evolution of top wealth shares in the United States?" *Journal of Economic Perspectives*, 29, 47–66.

Kopczuk, Wojciech and Eric Zwick (2020), "Business incomes at the top." *Journal of Economic Perspectives*, 34, 27–51.

Kuhn, Moritz, Moritz Schularick, and Ulrike I. Steins (2020), "Income and wealth inequality in America, 1949–2016." *Journal of Political Economy*, 128, 3469–3519.

Leckelt, Marius, Johannes König, David Richter, Mitja D. Back, and Carsten Schröder (2022), "The personality traits of self-made and inherited millionaires." *Humanities and Social Sciences Communications*, 9, 1–12.

Martinez, Isabel (2021), "Evidence from unique Swiss tax data on the composition and joint distribution of income and wealth." In *Measuring Distribution and Mobility of Income and Wealth*, 105–142, National Bureau of Economic Research, Inc.

Martínez-Toledano, Clara (2017), "Housing bubbles, offshore assets and wealth inequality in Spain (1984-2013)." *Paris School of Economics Working Paper*.

Martínez-Toledano, Clara (2020), "House price cycles, wealth inequality and portfolio reshuffling." *WID. World Working Paper*, 2.

McCrae, Robert R. and Paul T. Costa Jr (1997), "Personality trait structure as a human universal." *American Psychologist*, 52, 509.

Mullainathan, Sendhil and Jann Spiess (2017), "Machine learning: an applied econometric approach." *Journal of Economic Perspectives*, 31, 87–106.

Nekoei, Arash and David Seim (2023), "How do inheritances shape wealth inequality? theory and evidence from sweden." *The Review of Economic Studies*, 90, 463–498.

Nelsen, R.B. (2006), *An Introduction to Copulas*. Springer, Berlin.

Ozkan, Serdar, Joachim Hubmer, Sergio Salgado, and Elin Halvorsen (2023), "Why are the wealthiest so wealthy? a longitudinal empirical investigation."

Piketty, Thomas, Gilles Postel-Vinay, and Jean-Laurent Rosenthal (2014a), "Inherited vs self-made wealth: Theory & evidence from a rentier society (Paris 1872–1927)." *Explorations in Economic History*, 51, 21–40.

Piketty, Thomas and Emmanuel Saez (2003), "Income inequality in the United States, 1913–1998." *The Quarterly journal of economics*, 118, 1–41.

Piketty, Thomas, Emmanuel Saez, and Stefanie Stantcheva (2014b), "Optimal taxation of top labor incomes: A tale of three elasticities." *American Economic Journal: Economic Policy*, 6, 230–71.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018), "Distributional national accounts: methods and estimates for the United States." *The Quarterly Journal of Economics*, 133, 553–609.

Raub, Brian, Barry Johnson, and Joseph Newcomb (2010), "A comparison of wealth estimates for America's wealthiest decedents using tax data and data from the Forbes 400." In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, volume 103, 128–135, JSTOR.

Saez, Emmanuel (2002), "The desirability of commodity taxation under non-linear income taxation and heterogeneous tastes." *Journal of Public Economics*, 83, 217–230.

Saez, Emmanuel and Gabriel Zucman (2016), "Wealth inequality in the United States since 1913: Evidence from capitalized income tax data." *The Quarterly Journal of Economics*, 131, 519–578.

Saez, Emmanuel and Gabriel Zucman (2019), "Progressive wealth taxation." *Brookings Papers on Economic Activity*, 2019, 437–533.

Saez, Emmanuel and Gabriel Zucman (2020), "The rise of income and wealth inequality in America: Evidence from distributional macroeconomic accounts." *The Journal of Economic Perspectives*, 34, 3–26.

Salas-Rojo, Pedro and Juan Gabriel Rodríguez (2022), "Inheritances and wealth inequality: a machine learning approach." *The Journal of Economic Inequality*, 20, 27–51.

Schluter, Christian (2018), "Top incomes, heavy tails, and rank-size regressions." *Econometrics*, 6, 10.

Schluter, Christian (2020), "On Zipf's law and the bias of Zipf regressions." *Empirical Economics*, 1–20.

Schröder, Carsten, Charlotte Bartels, Markus M. Grabka, Johannes König, Martin Kroh, and Rainer Siegers (2020), "A novel sampling strategy for surveying high net-worth individuals—a pretest application using the socio-economic panel." *Review of Income and Wealth*, 66, 825–849.

Siegers, Rainer, Hans Walter Steinhauer, and Johannes König (2021), "SOEP-core-2019: Sampling, nonresponse, and weighting in sample P." Technical report, SOEP Survey Papers.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick (2019), "Capitalists in the Twenty-First Century." *The Quarterly Journal of Economics*, 134, 1675–1745.

Smith, Matthew, Owen Zidar, and Eric Zwick (2023), "Top Wealth in America: New Estimates under Heterogeneous Returns." *The Quarterly Journal of Economics*, 138, 515–573.

SOEP Group (2021), "Soep-core v36 – pgen: Person-related status and generated variables." Technical report, SOEP Survey Papers.

Stanley, Thomas J. (1996), *The millionaire next door: The surprising secrets of America's wealthy*. Taylor Trade Publishing.

Vermeulen, Philip (2016), "Estimating the top tail of the wealth distribution." *American Economic Review*, 106, 646–650.

Vermeulen, Philip (2018), "How fat is the top tail of the wealth distribution?" *Review of Income and Wealth*, 64, 357–387.

Wolff, Edward N. (2021), "Household wealth trends in the united states, 1962 to 2019: Median wealth rebounds... but not enough." Technical report, National Bureau of Economic Research.

(Web) Appendix

A. Statistical Appendix: Top wealth and income

We detail our statistical methods for top wealth. The case of top incomes is, of course, analogous.

A.1. The Pareto QQ plot, and the estimator for the tail index

Consider the k largest upper-order statistics of wealth $X_{n,n} \geq \cdots \geq X_{n-(k-1),n} \geq \cdots \geq X_{1,n}$ of a sample of size n from wealth distribution F.

The Pareto QQ-plot[1] has coordinates $(x, y) = (-\log(j/(n+1)),\ \log X_{n-j+1,n})_{j=1,\cdots,k}$. In Pareto-like models $1 - F(x) = x^{-\frac{1}{\gamma}} l(x)$, with $\gamma > 0$ and l slowly varying,[2] this plot becomes *ultimately* linear for a sufficiently high threshold $X_{n-k,n}$ where $k < n$. The line through the threshold point $(-\log((k+1)/(n+1)), \log X_{n-k,n})$ with slope γ is thus given by

$$y = \log X_{n-k,n} + \gamma \left[x + \log \left(\frac{k+1}{n+1} \right) \right] \qquad (1 \leq j \leq k < n).$$

The OLS estimator of the slope parameter in the Pareto QQ-plot is obtained by minimizing the least squares criterion

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

with respect to γ, which corresponds to a regression of log sizes on the log of relative ranks for sufficiently large wealth given by $X_{n-k,n}$. The resulting OLS estimator is

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

[1] Recall the probability-probability plot for distribution F given by $\{F(X_{n-j,n}), F_n(X_{n-j,n}) \sim \frac{n-j+1}{n+1}\}$ where F_n denotes the empirical distribution function. Asymptotically, its linearity follows from the Glivenko–Cantelli theorem. The quantile-quantile (QQ) Plot is simply $\{X_{n-j,n}, F^{-1}(\frac{n-j+1}{n+1})\}$, and the Pareto QQ plot follows with F being Pareto.

[2] Recall that l is said to be slowly varying at infinity if $l(tw)/l(w) \to 1$ as $w \to \infty$.

A.2. Dual regressions

Instead of regressing log sizes on log ranks, leading to the estimator given by equation (A.2), one could take a dual approach and regress log ranks on log sizes. A further variant includes the additional estimation of a regression constant, so that $\log X_{n-j+1,n}$ is regressed on a constant and $\log j$. Kratz and Resnick (1996) obtain the distributional theory for this alternative estimator and show that its asymptotic variance is $2\gamma^2/k$.

A.3. Distributional theory

Schluter (2018) develops the distributional theory for $\hat{\gamma}$ given by equation (A.2) in a setting where the slowly varying function l in $1 - F(x) = x^{-\frac{1}{\gamma}} l(x)$ exhibits second-order regular variation, that is,

$$\lim_{t\to\infty} \frac{\frac{\log U(tx) - \log U(t)}{a(t)/U(t)} - \log x}{A(t)} = H_{\gamma,\rho}(x) \tag{A.3}$$

for all $x > 0$, where $H_{\gamma>0,\rho<0}(x) = \frac{1}{\rho}(\frac{x^\rho - 1}{\rho} - \log x)$ with $\rho < 0$. U is the tail quantile function $U(x) \equiv F^{-1}(1 - 1/x)$, and a a positive norming function with the property $a(t)/U(t) \to \gamma$. The parameter ρ is the so-called second-order parameter of regular variation, and $A(t)$ is a rate function that is regularly varying with index ρ, with $A(t) \to 0$ as $t \to \infty$. As ρ falls in magnitude, the nuisance part of l decays more slowly. Schluter (2018) demonstrates then that, as $k \to \infty$ and $k/n \to 0$, this estimator is weakly consistent, and if $\sqrt{k}A(n/k) \to 0$

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

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

A.4. The choice of the threshold k

Any tail index estimator requires a choice of how many upper order statistics, 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 in the top wealth literature. 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 for estimator A.2 by optimally resolving the mean-variance trade-off by minimizing the asymptotic mean-squared

error (AMSE).

Following Beirlant et al. (1996) and Schluter (2018, 2020), we observe that the expectation of the mean-weighted theoretical squared deviation

$$\frac{1}{k}\sum_{j=1}^{k} w_{j,k} E\left(\log\left(\frac{X_{n-j+1,n}}{X_{n-k,n}}\right) - \gamma \log\left(\frac{k+1}{j}\right)\right)^2 \tag{A.5}$$

equals, to first order,

$$c_k Var(\hat{\gamma}) + d_k(\rho) b_{k,n}^2 \tag{A.6}$$

for some coefficients c_k depending only on k, and $d_k(\rho)$ depending on k and $\rho < 0$ (being the so-called parameter of second-order regular variation that governs the speed of decay of the slowly varying nuisance function l in the distribution model 1).[3] For an explicit statement of the coefficients c_k and d_k, see Schluter (2018). The procedure then consists in applying two different weighting schemes $w_{j,k}^{(i)}$ $(i = 1, 2)$ in (A.5), estimating the corresponding two mean weighted theoretical deviations using the residuals of regression (A.1), and computing a linear combination thereof such that

$$Var(\hat{\gamma}) + b_{k,n}^2$$

obtains. 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, 2020), we have set a very conservative value of $\rho = -.5$ (which implies a slow decay of the slowly varying nuisance function l).

In Appendix D.1, we provide illustrations for our wealth data. In particular, we provide plots of the AMSE as a function of k, juxtapose conventional fixed threshold choices, and compare the method to alternatives using subjective visual choices based on Hill-type plots.

A.5. Complex surveys

Survey data such as the SOEP come with sampling weights. The aforementioned theory is easily adapted to this setting if we define the weighted empirical distribution function as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} w_i 1(X_i \leq x) \tag{A.7}$$

[3] For example, for the Burr distribution with $\gamma > 0$ and $\rho < 0$, we have $1 - F_{(\gamma,\rho)}(x) = (1 + x^{-\rho/\gamma})^{1/\rho} \approx x^{-1/\gamma}[1 + (1/\rho)x^{\gamma/\rho}]$ and the tail quantile function can be expanded as $U(x) = x^{\gamma}[1 + (\gamma/\rho)x^{\rho} + o(x^{\rho})]$.

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

$$(x, y) = (-\log(\sum_{i=1}^{j} w_{(i \leq j)}/(n+1)), \log X_{n-j+1,n})_{j=1,\cdots,k},$$

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

$$\hat{\gamma} \quad = \quad \frac{\frac{1}{k}\sum_{j=1}^{k} \log\left(\frac{\sum_{i=1}^{k+1} w_{(i \leq k+1)}}{\sum_{i=1}^{j} w_{(i \leq j)}}\right) [\log X_{n-j+1,n} - \log X_{n-k,n}]}{\frac{1}{k}\sum_{j=1}^{k} \left[\log \frac{\sum_{i=1}^{k+1} w_{(i \leq k+1)}}{\sum_{i=1}^{j} w_{(i \leq j)}}\right]^2}. \tag{A.8}$$

The estimator A.2 then follows as a special case of A.8 with unitary weights $w_i \equiv 1$.

A.6. Computation of top wealth shares

Assuming that the Pareto QQ plot becomes approximately linear from the k's largest observation, $X_{n-k+1,n} \equiv w_{min}$, the complete wealth distribution F is, for $x > w_{min}$,

$$F(x) = p + (1-p)\left(1 - \left(\frac{x}{w_{min}}\right)^{-1/\gamma}\right) \tag{A.9}$$

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

$$Q(u) = \left(\frac{1-u}{1-p}\right)^{-\gamma} w_{min}, \qquad (u > p). \tag{A.10}$$

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

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

Taking into account the survey sampling weights ω_i for household i enumerated from the

poorest to the richest, we have

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

Expected wealth then is simply $E(X) = pE_{SOEP} + (1-p)\left(\frac{\alpha}{\alpha-1}\right) w_{min}$ with $\alpha = 1/\gamma$ and E_{SOEP} being the empirical mean wealth in the survey data conditional on wealth not exceeding w_{min}. The wealth share of the top $t100\%$ then is, with $1 - t = u > p$,

$$t^{1-1/\alpha} \left(\frac{\alpha}{\alpha - 1}\right) w_{min}(1-p)^{1/\alpha}/E(X). \tag{A.11}$$

The so-called inverted Pareto coefficient $E(W|W > w)/w$ with w equal to the top t quantile x_{1-t} and $1 - t = u > p$ is $\alpha/(\alpha - 1)$.
Finally, we observe that augmenting SOEP with SOEP-P changes p.

Appendix A References

Beirlant, J., Vynckier, P. and Teugels, J. L., (1996), "Tail index estimation, Pareto quantile plots, and regression diagnostics," *Journal of the American Statistical Association*, 9, 436, 1659-1667.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004), *Statistics of Extremes*, Wiley Series in Probability and Statistics, Chichester: Wiley.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997), *Modelling extremal events*, Berlin: Springer.

Kratz, M., Resnick, S. I. (1996), "The QQ-estimator and heavy tails," *Communications in Statistics. Stochastic Models*, 12, 699-724.

Schluter, C., (2018), "Top incomes, heavy tails, and rank-size regressions," *Econometrics*, 6, 10.

Schluter, C., (2020), "On Zipf's law and the bias of Zipf regressions," *Empirical Economics*, 1–20.

B. Statistical Appendix: Copulas and the dependence of wealth and income

B.1. Wealth and income: Goodness of fit of the Gumbel copula

We conjecture that the one-parameter Gumbel copula given by

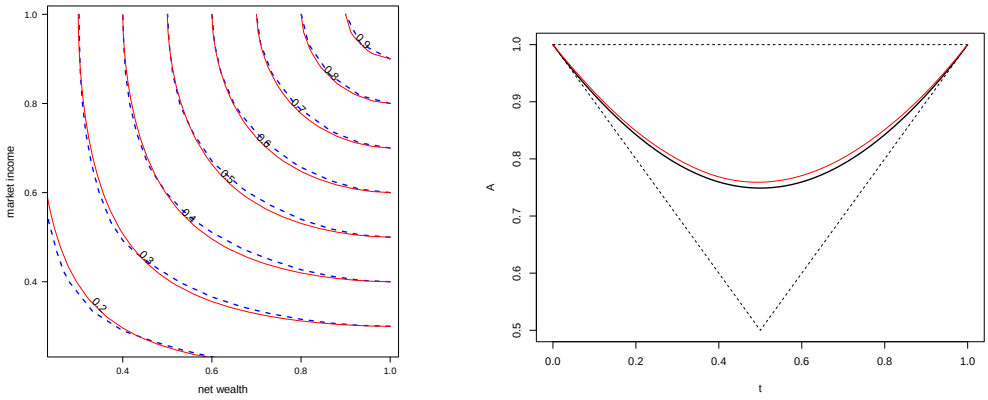
$$C_\theta(u) = \exp\Big(- (\sum_{j=1}^{2} (-\log u_j)^\theta)^{1/\theta} \Big) \qquad (u \in [0,1]^2)$$

could describe the data well and in a parsimonious fashion, since it is an extreme-value copula, the marginal distributions are of the Pareto-type, and we observe a strong dependence in the upper tail.

In order to verify this conjecture, we adopt a goodness-of-fit approach. First, we non-parametrically estimate the empirical copula, which is simply the empirical joint distribution function of the empirical ranks of wealth and income. Then we consider the Gumbel copula model, which implies a functional relationship between the rank correlation ρ and the Gumbel parameter θ. An estimate of the latter is obtained by inverting the theoretical mapping $\rho(\theta)$ and evaluating it at the empirical ρ, thus yielding a method-of-moments estimate.

We start by considering the joint CDF for wealth and market income by means of a contour plot. To this end, we depict in Panel (a) of Figure B.1 the contours of non-parametrically estimated empirical copula at the stated percentiles (red line), and the respective contours implied by the estimated Gumbel model. It is evident that the contours are closely aligned, suggesting visually that the Gumbel model fits the data well, and specifically so in the upper tails.

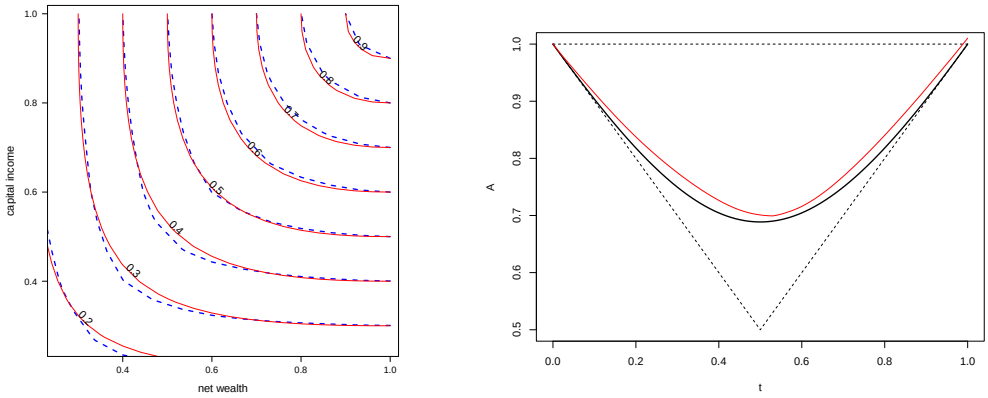
Figure B.1: Contour plot of the empirical and Gumbel copula, and Pickands tail dependence function



Notes. Income: Market Income. Panel (a): Contour plot of the joint distribution of wealth and market income, using the empirical copula (blue dashed line) and the method-of-moments fitted Gumbel copula (red dashed line). Panel (b): Pickands dependence function, using an empirical estimate (red line) and the Gumbel model implied function $A_\theta = (t^\theta + (1-t)^\theta)^{1/\theta}$ for $(t \in [0,1])$. *Source*: SOEP+P.

An alternative goodness-of-fit assessment suggested in the theoretical statistical literature is to consider the so-called Pickands dependence function A for extreme value copulas (see Pickands, 1981, or Hofert et al., 2017, for a textbook treatment). We proceed by first estimating this function non-parametrically (based on Genest and Segers, 2009), and then comparing this estimate to the function implied by the Gumbel model, which is to equal $A_\theta = (t^\theta + (1-t)^\theta)^{1/\theta}$ for $(t \in [0,1])$. This is done in Panel (b) of Figure B.1. The empirical estimate (red line) and the fitted model function (black line) are again closely aligned, thus confirming the good fit of the Gumbel model. Next, we turn to the interpretation of the Pickands dependence function. As a benchmark, no dependence would result in a horizontal line, whereas complete dependence in the depicted "v" with minimum at .5; formally, since we have only 2 dimensions, it is always true that $max\{1-t, t\} \leq A(t) \leq 1$. We have added these limits to the figure as a visual guide. It is clear that the estimated dependence function exhibits considerable dependence between wealth and market income.

Figure B.2: Capital income: Contour plot and Pickands tail dependence



Notes. As per Figure B.1. Income: Capital income. *Source*: SOEP+P.

We repeat in Figure B.2 these juxtapositions for the case of wealth and capital income. Qualitatively, similar observations obtain, leading us to conclude that the Gumbel copula also provides a good fit for the joint distribution of wealth and capital income. Quantitatively, it is of interest to observe that in the case of capital income, the larger rank correlation with wealth also manifests itself in the Pickands dependence function whose minimum is smaller than that for market income, thus indicating a larger dependence between capital income and wealth.

Table B.1: Wealth and income: Dependence analysis

wealth &	Rank cor.	Gumbel Copula	
income	ρ	θ	λ_u
MktInc	0.583	1.716	0.502
PostInc	0.613	1.791	0.527
LabInc	0.415	1.404	0.361
LabInc>0	0.551	1.644	0.476
CapInc	0.724	2.165	0.623
CapInc>0	0.648	1.887	0.556

Notes. ρ is Spearman's empirical rank correlation coefficient between wealth and income. The Gumbel copula parameter θ, the estimate of θ is obtained by inverting the theoretical mapping $\rho(\theta)$. λ_u is the upper-tail dependence measure, $\lambda_u = lim_{q\to 1} \Pr\{W > F_W^{-1}(q) | Y > F_Y^{-1}(q)\}$ which in the Gumbel case is simply $2 - 2^{1/\theta}$. See Statistical Appendix B for further details and a goodness-of-fit assessment. *Source*: SOEP+P.

B.2. Wealth and income: Shares for the joint distribution

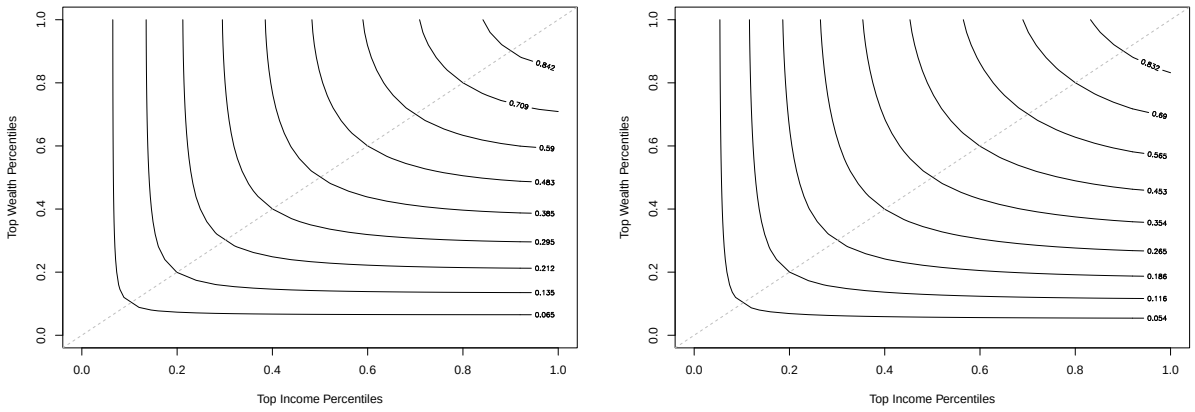
Denote the marginal distribution of wealth by F_W and that of income by F_Y. The copula C is $C(u_1, u_2) = H(F_W^{-1}(u_1), F_Y^{-1}(u_2))$ where $u_i \in [0, 1]$ and H denotes the joint distribution. Let c_d denote the copula density.

Consider the group of households that are in the top $s \times 100$ percent of the marginal wealth and income distribution. The wealth share of this group is $E\{W|W > F_W^{-1}(1-s), Y > F_Y^{-1}(1-s)\}/E(W)$ where $E(W)$ denotes average wealth. In particular,

$$E\Big\{W|W > F_W^{-1}(1-s), Y > F_Y^{-1}(1-s)\Big\} = \frac{\int_{1-s}^{1}\int_{1-s}^{1} F_W^{-1}(u_1)c_d(u_1,u_2)du_1u_2}{\int_{1-s}^{1}\int_{1-s}^{1} c_d(u_1,u_2)du_1u_2}$$

The upper wealth quantile $F_W^{-1}(u_1)$ is given by equation (A.10) for $u_1 > p$. An analogous expression holds for the income share of this group.

Figure B.3: Survival copula for capital and market income



(a) Capital Income

(b) Market Income

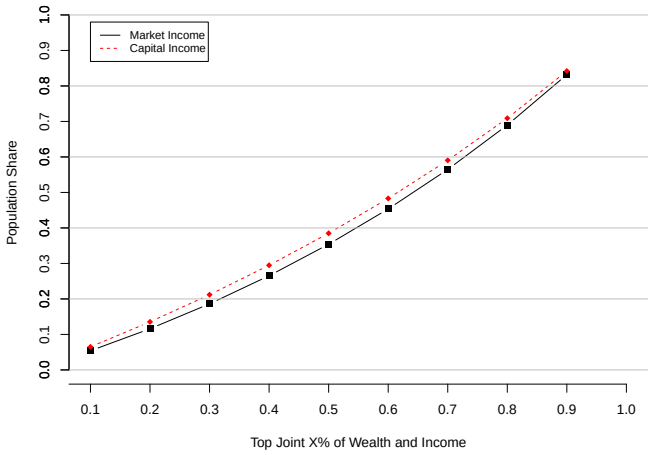
Notes. Contour plot of the survival copula of wealth and either capital or market income computed from the respective method-of-moments fitted Gumbel copula. *Source*: SOEP+P.

B.3. Ridge plots and the joint survival copula

We depict the population shares across the entire joint distribution in Figure B.4. This is a plot of the population shares for the top joint x% in distribution of wealth and either market or capital income (so for x=.1 the population share is 5.4% for market income).

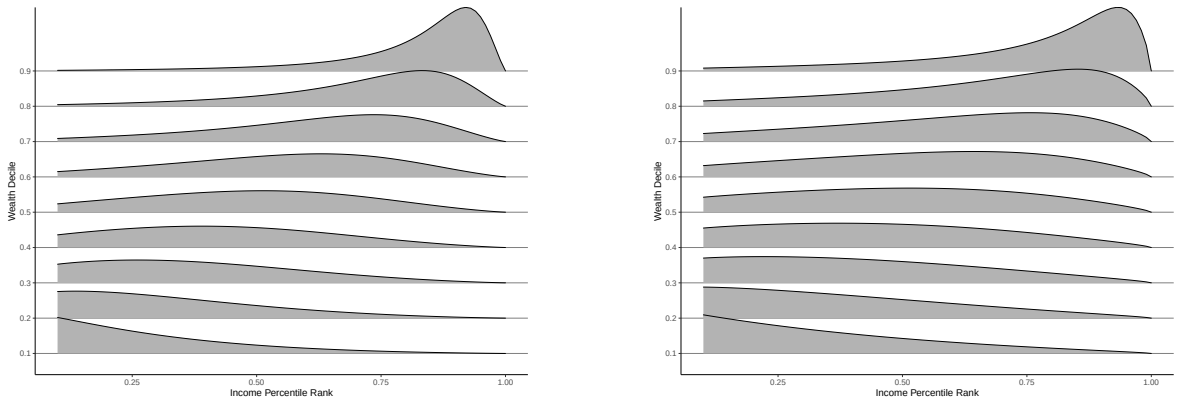
More precisely, we evaluate the joint survival copula[4] along the main diagonal. Similar to a Lorenz curve, the gap from the main diagonal indicates how dependent wealth and incomes are.

Figure B.4: Population shares at the joint top



Notes. Formally this is a plot of the joint survival copula of wealth and the respective income concept. *Source*: SOEP+P.

Figure B.5: Ridge plots for wealth and income



(a) Capital Income

(b) Market Income

Notes. We plot the copula density $c(u_w, .)$ for fixed wealth deciles u_w. *Source*: SOEP+P.

[4] The survival copula is the rank-based analogue of the survival function. Having defined our copula by $C(u_1, u_2) = H(F_W^{-1}(u_1), F_Y^{-1}(u_2))$, where H is the cdf, and letting $\bar{H}$ denote the survival function, the survival copula is $\bar{C}(u_1, u_2) = \bar{H}(F_W^{-1}(1-u_1), F_Y^{-1}(1-u_2))$ where $u \in [0,1]^2$.

A complementary visualization is the ridge plot in Figure B.5, which depicts for a selected wealth decile the copula density across income ranks.[5] The principal strength of such a plot (compared to, say, a contour plot of the copula) is to clearly reveal which area of the income-wealth space contains the largest population concentration. In particular, our ridge plot shows that for low wealth deciles, much of the mass is in the lower income tail, while for high wealth deciles, the mass is concentrated in high income deciles. Comparing capital income and market income for the .9 wealth decile shows that the former is more concentrated at the top.

Appendix B References

Genest, C. and Segers, J. (2009), " Rank-based inference for bivariate extreme-value copulas," *Annals of Statistics*, 37, 2990-3022.

Hofert, M. and Kojadinovic, I. and Mächler, M. and Yan, Y.. (2017) *Elements of Copula Modeling with R*, Springer.

Pickands, J. (1981). "Multivariate extreme value distributions," *Bulletin de l'Institut international de statistique*, 49, 859-878.

[5] That is, we plot $c(u_w, .)$ for fixed wealth decile u_w, where c is the copula density associated with copula $C(u_w, u_y)$ and $u_w, u_y \in [0, 1]$.

C. Data Appendix

C.1. The personal balance sheet

All SOEP and SOEP-P respondents completed the module "Your personal balance sheet." This module asks about respondents' portfolios in a three-stage procedure: Each respondent is asked in the first stage if (s)he holds a particular asset, in the second stage about the market value of that asset, and in the third if the individual is the only holder, and if ownership is shared, what share is held by the individual. The module differentiates twelve asset and debt positions:

1. Value of owner-occupied real estate assets
2. Value of other real estate assets
3. Value of building loan contracts
4. Value of financial assets
5. Surrender value of life insurance and private pension insurance
6. Value of company or shareholdings in companies
7. Value of tangible assets
8. Value of vehicles
9. Outstanding debt for other real estate assets
10. Outstanding debt for owner-occupied real estate assets
11. Outstanding debt in consumer loans
12. Outstanding debt in educational loans

The sum of the positions 1 through 8 gives gross wealth. Deducting positions 9 through 12 from gross wealth gives net wealth.

C.2. Income concepts

The first four income concepts we consider are recorded on the level of the household. The variables are taken from the PEQUIV dataset of the SOEP, which is an internationally harmonized dataset documented in Grabka (2021). The final income concept is individual labor income, which is also contained in the PEQUIV dataset.

Labor income Labor income is the sum of all earnings from every form of employment including from training, primary and secondary jobs, and self-employment, as well as from irregular compensation, such as bonuses, overtime, and profit-sharing.

Capital income Capital income consists of two broad income sources: financial asset income and income from real estate. Financial asset income consists of dividends, interest payments, as well as capital gains. Real estate income is net income from renting and leasing.

Market income Market income is derived from two variables: pre-government income and pension income. Pre-government income is the sum of labor income, capital income, private retirement income, and private transfers. Pension income consists of old-age and widow/widower pensions.

Post-Government income To derive post-government income, one constructs the sum of labor and capital income plus private and public transfers as well as public and private pensions. From this sum, one deducts total household taxes on these incomes to arrive at post-government income.

Individual labor income Individual labor income is also provided in the PEQUIV dataset and is definitionally equivalent to household labor income, except that it is not aggregated on the household level.

Table C.1: Households with net wealth and market income in 2019 exceeding high thresholds

	Net Wealth ≥			
	.5M€	1M€	5M€	10M€
SOEP	1493	456	27	10
SOEP+SOEP-P	2734	1306	212	66
	Market Income ≥			
	0.1M€	0.5M€	0.75M€	1M€
SOEP	1596	20	12	8
SOEP+SOEP-P	2827	122	63	37

Notes. "N" refers to the number of observations. The wealth threshold of .5M€ corresponds to about 10% of the weighted population. Market income is labor and capital incomes including pensions. The market income threshold of .1M€ corresponds to about 9% of the weighted population. *Source*: SOEP+P.

C.3. Intergenerational transfers

The SOEP questionnaires in 2001, 2017, and 2019 included questions on individual inheritances and gifts. Individuals were asked to record the year, value, type (inheritance/ gift), and asset type (real estate, securities/bonds/shares, cash/deposits, business, other) of at most three inheritances or gifts. To make inheritances and gifts received in different years comparable in the cross-section of 2019, we capitalize the inheritances and gifts recorded in the questionnaires in 2001, 2017, and 2019 using CPI-adjusted bond-rates for Germany provided by Jordà et al. (2019). Our measure of individual inheritances and gifts is the sum of all capitalized inheritances and gifts ever received by the individual.

C.4. Self-Made individuals

We define self-made individuals based on a battery of questions in the 2019 SOEP questionnaire that asks about the factors that have reduced, not influenced, or increased the amount of wealth an individual currently holds. These factors are: 1) entrepreneurship or self-employment 2) dependent employment 3) earnings from financial transactions 4) real estate 5) gifts 6) inheritances 7) marriage 8) lottery winnings. Respondents rated each of these factors on an 11-point Likert scale. This scale is split in the middle, meaning that values below 6 indicate that the factor decreased current wealth, values above 6 indicate that the factor increased wealth, while a value of 6 indicates that the factor left

current wealth unchanged. Most people will not indicate that a certain factor lowered their wealth, so responses are concentrated on values 6 or above.

We define the status of being self-made based solely on respondents' self-ratings on these scales because other "objective" factors will obfuscate the content of the self-ratings. For example, if we define self-made individuals according to respondents' self-ratings on entrepreneurship and current self-employment status, we will miss entrepreneurs who have already retired but earned a substantial amount of wealth from entrepreneurial activity.

We require that self-made individuals fulfill the following criteria: 1. They rate the importance of entrepreneurship and self-employment in increasing their individual wealth greater than 9. 2. They rate the importance of either gifts or inheritances or marriage in increasing their individual wealth less than or equal to 9.

In Leckelt et al. (2022) the measure is thoroughly examined and shown to be aligned with a multitude of objective variable correlated with self-made economic success like the size and share of business wealth or the share of one's intergenerational transfers w.r.t. current wealth.

D. Additional empirical results: The validation exercises

D.1. Threshold selection for our wealth data

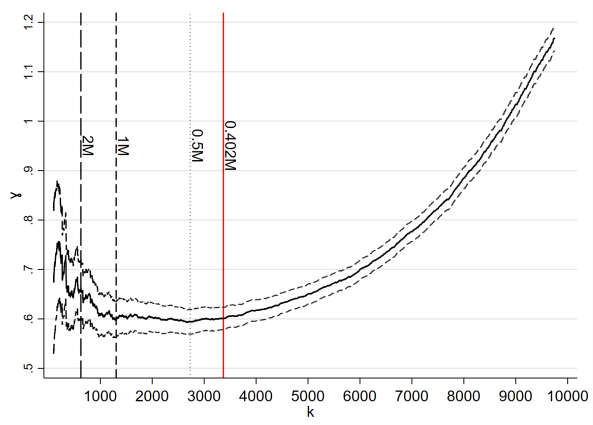
In this empirical appendix, we discuss the similarities between point estimates reported in panels B and C of the Table 1 in the main text using a Hill-type plot of the estimator as a function of the number of upper-order statistics k. For our wealth data, this plot, shown below, exhibits an extended horizontal section. Over this range, point estimates will not change significantly, while the estimated variability does, and it is then optimal to set the threshold choice at the far end of this horizontal section.

In this appendix, we also show that our threshold selection is robust against alternative data-dependent methods, as is our estimation method (see, e.g., Embrechts et al. (1997) for an extensive exposition). For instance, making a subjective visual choice based on the Hill-type plot directly would coincide with our optimal AMSE criterion. Variants such as the alt-Hill plot yield a threshold choice similar to ours. Alternative estimators yield similar results.

D.1.1. Hill-type plot for our wealth data and the rank-size regression estimator

The wealth threshold selection problem is illustrated in a so-called Hill-type plot in Appendix Figure D.1 for our wealth data in SOEP+P: The estimator, plotted as a function of the number of upper order statistics k, exhibits an extended horizontal section: While the estimate over this range will not change significantly, the estimated variability does, and it is then optimal to set the threshold choice at the far end of this section. As a result, our estimate is more precise than the alternative estimates based on the arbitrary thresholds.

Figure D.1: Hill-type plot: $\hat{\gamma}(k)$ and the optimal k^*

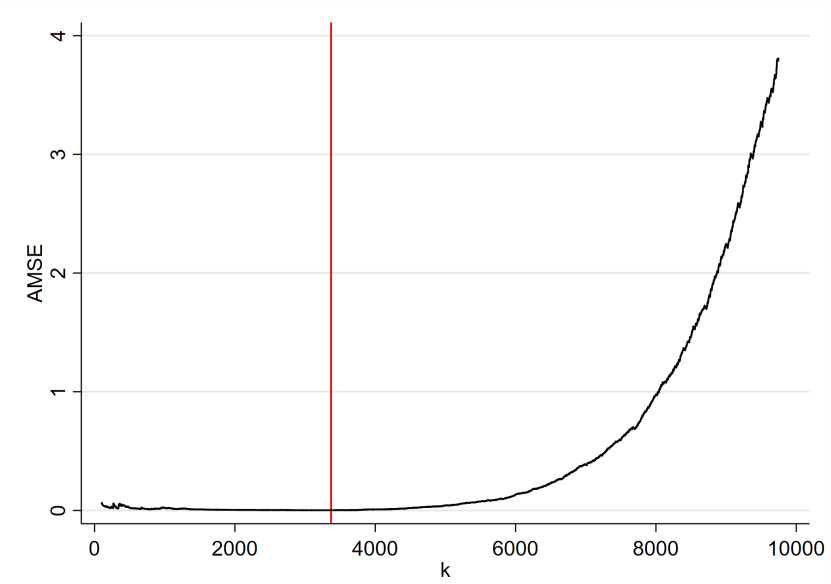


Notes. $\hat{\gamma}$ is depicted as a function of the number k of upper order statistics (solid line) and 95% confidence limits (dashed lines). Vertical lines correspond to the common wealth thresholds of 0.5M, 1M, and 2M €, and the optimal choice k^* (solid red line) is given by the minimization of the asymptotic mean-squared error (AMSE) of the estimator. For the AMSE plot, see the Appendix Section D.1.

D.1.2. Minimizing the AMSE for wealth data

We apply these methods to optimally select the number of upper order statistics for the wealth distribution based on the augmented SOEP-SOEP-P data set. In Figure D.2 we depict the AMSE as a function of the number k of upper order statistics used. The criterion is minimised at $k^* = 3{,}370$ which corresponds to a wealth threshold of 402,200 €.

Figure D.2: The asymptotic mean-squared error (AMSE)

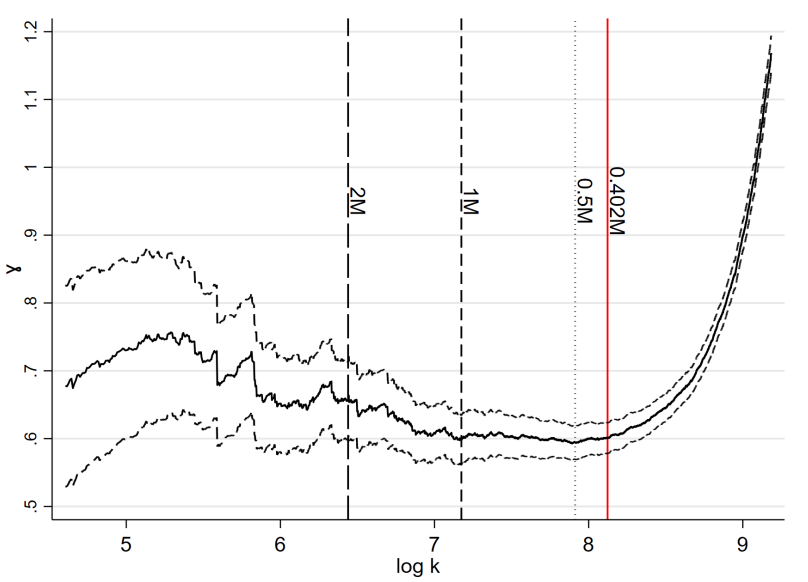


Notes. $AMSE(k)$, depicted as a function of the number k of upper-order statistics used, of the estimator $\hat{\gamma}$ given by equation (A.2), using the method detailed in Section A.4, for the augmented SOEP+P dataset. Red vertical line: indicates the optimal $k^* = argmin AMSE(k)$. *Source*: SOEP+P.

D.1.3. Alternative Hill plot for wealth data

An alternative to the subjective visual threshold choice based on the Hill-type plot is the so-called alt(ernative) Hill plot, proposed in Drees et al. (2000), which depicts $\hat{\gamma}(k)$ on $\log k$ instead instead of k. This alt-Hill plot is depicted in Figure D.3. In the current case, subjective visual methods of both the Hill plot and alt-Hill plot would result in similar choices.

Figure D.3: Alt(ernative) Hill-type plot: $\hat{\gamma}(k)$ and the optimal k^*



Notes. We depict the $\hat{\gamma}$ as a function of the number $\log k$ of upper-order statistics (solid line) and 95% confidence limits (dashed lines). Vertical lines corresponding to the common wealth thresholds of 0.5M, 1M, and 2M €, and the optimal choice k^* (solid red line), given by the minimization of the AMSE of the estimator. *Source*: SOEP+P.

D.2. Alternative tail index estimates for wealth data: The Hill estimator

Another popular estimator of the extreme value index and hence the tail index is the Hill estimator given by

$$\hat{\gamma}^{(Hill)} = \frac{1}{k}\sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n}$$

See, for instance, (Embrechts et al. 1997, chapter 6.4) for a textbook treatment, including the distributional theory based on second-order regular variation theory (similar to our Statistical Appendix, Section A.3). In particular, $var(\hat{\alpha}^{(Hill)}) = \alpha^2/k$ (their theorem 6.4.6), where of course $\hat{\alpha} \equiv 1/\hat{\gamma}$.

In order to adapt the estimator to the case of complex surveys, consider first the regular variation justification of the Hill estimator. In particular, partial integration yields $\int_t^\infty (1-F(x))/x dx = \int_t^\infty (\log x - \log t) dF(x)$. Applying Karamata's theorem to the left, it follows that $(1-F(t))^{-1}\int_t^\infty (\log x - \log t) dF(x) \to \gamma$ as $t \to \infty$. A natural estimator then employs the empirical distribution function, which in the complex survey case is

given by equation (A.7), and sets $t = X_{n-k,n}$. The resulting Hill estimator is

$$\hat{\gamma}^{(Hill)} = \frac{1}{\sum_{i=1}^{k+1} w_{(i\leq k+1)}} \sum_{j=1}^{k} w_j \big(\log X_{n-j+1,n} - \log X_{n-k,n} \big)$$

(with the implicit notation convention that $\sum_{i=1}^{j} w_{(i\leq j)}$ denotes the summation of the survey weights corresponding to the j largest upper or statistics of wealth).

We report in Table D.1 the Hill estimates of the tail index $\hat{\alpha} \equiv 1/\hat{\gamma}$, for various thresholds, and conclude that they yield broadly the same qualitative conclusions as the rank-size regression estimator. In particular, at our optimal k for the regression estimator, the point estimate of 1.67 is almost the same as our estimate of 1.66.

Table D.1: Tail index estimates of the wealth distribution using the Hill estimator

data	threshold	α	$SE(\alpha)$
A. Fixed wealth thresholds			
SOEP	500000	1.809	0.002
SOEP+SOEP-P	500000	1.728	0.001
SOEP	1000000	1.942	0.008
SOEP+SOEP-P	1000000	1.777	0.002
SOEP	2000000	1.707	0.024
SOEP+SOEP-P	2000000	1.556	0.004
B. Our optimal threshold k			
SOEP+SOEP-P	402200	1.668	0.0008

Notes. The tail index estimate is $\hat{\alpha} \equiv 1/\hat{\gamma}$, and by the delta method $var(\hat{\alpha}^{(Hill)}) = \alpha^2/k$. *Source*: SOEP+P.

D.3. Estimates and confidence intervals for top wealth shares

Table D.2: Estimates and confidence intervals for top wealth shares

		Top 10%			Top 1%		
	Thr.	CI -	Est	CI +	CI -	Est	CI +
SOEP	0.5M€	53.44	55.04	56.94	19.18	19.76	20.44
	1M€	53.53	55.02	57.19	18.08	19.41	21.35
	2M€	53.74	55.43	59.82	17.41	20.43	28.27
SOEP+P	0.5M€	56.28	57.49	58.84	22.11	22.59	23.12
	1M€	56.65	57.63	58.80	22.08	23.00	24.09
	2M€	57.25	58.14	59.30	22.79	24.40	26.50
	0.402M€	56.41	57.45	58.62	22.49	22.90	23.37

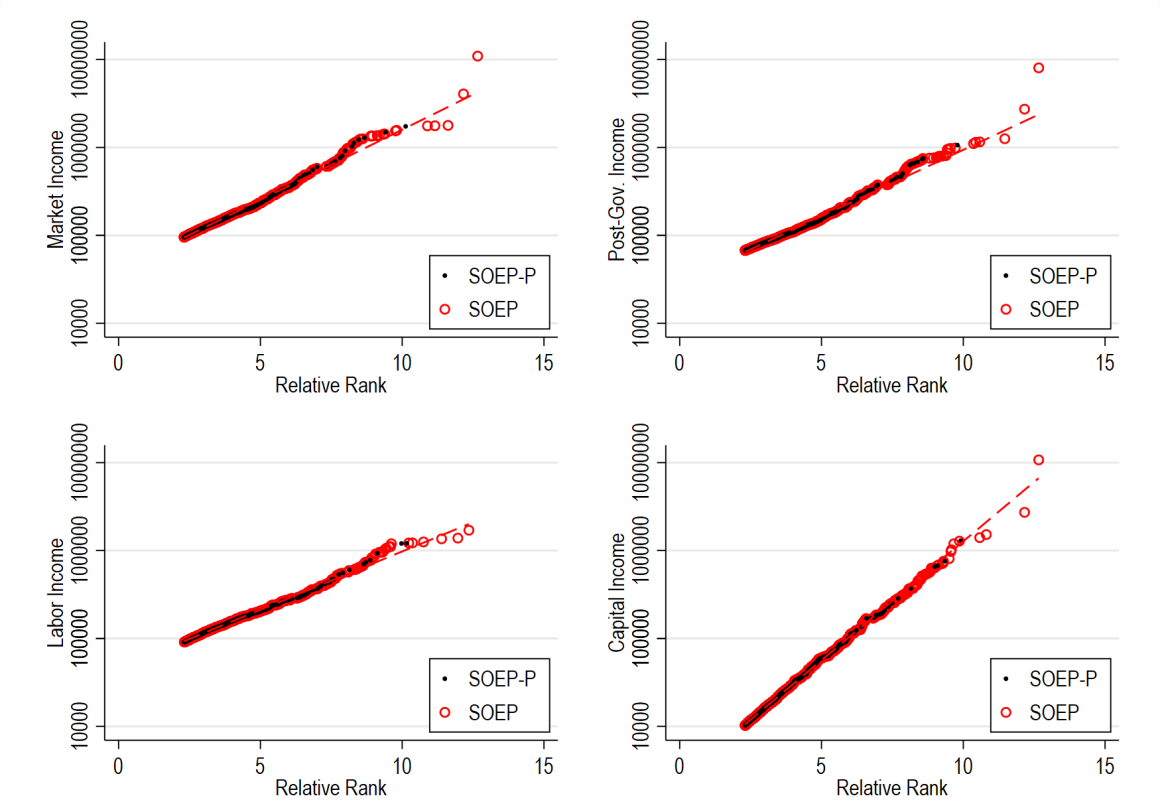
Notes. Confidence Intervals for the wealth shares are calculated by using the upper and lower 95% confidence limits of the Pareto α and following the formulae derived in Appendix A.6. *Source*: SOEP+P.

D.4. Additional results for income data

The Pareto QQ plots depicted in Figure D.4 illustrate that the inclusion of SOEP-P successfully appends extremes and fills in the upper tail of the respective income distributions.

D.4.1. Additional Pareto QQ-plots: All four income concepts

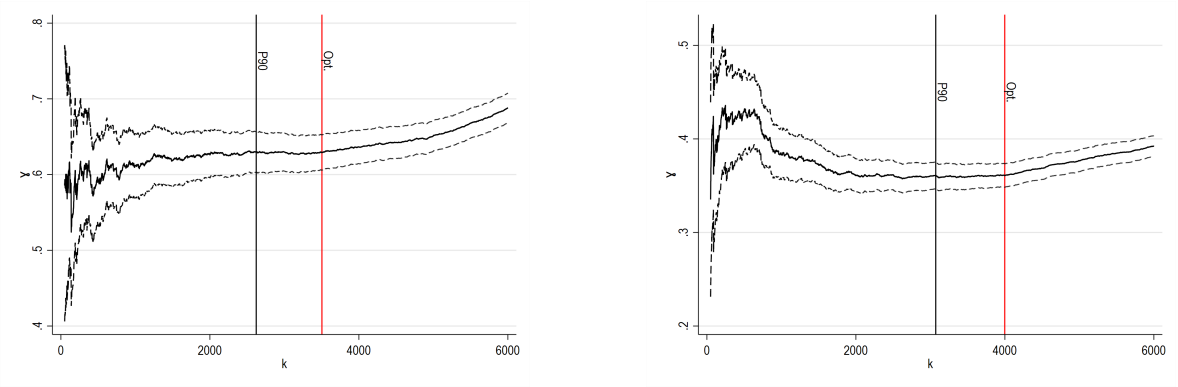
Figure D.4: Four income concepts: Pareto QQ plot



Notes. *MktInc* is household market income , *PostInc* is post-government household income, *LabInc* is household labor income, *CapInc* is household capital income. Upper-order statistics of income for the SOEP (red circles) and SOEP-P (black dots). *Source*: SOEP+P.

D.4.2. Hill-type plots for household incomes

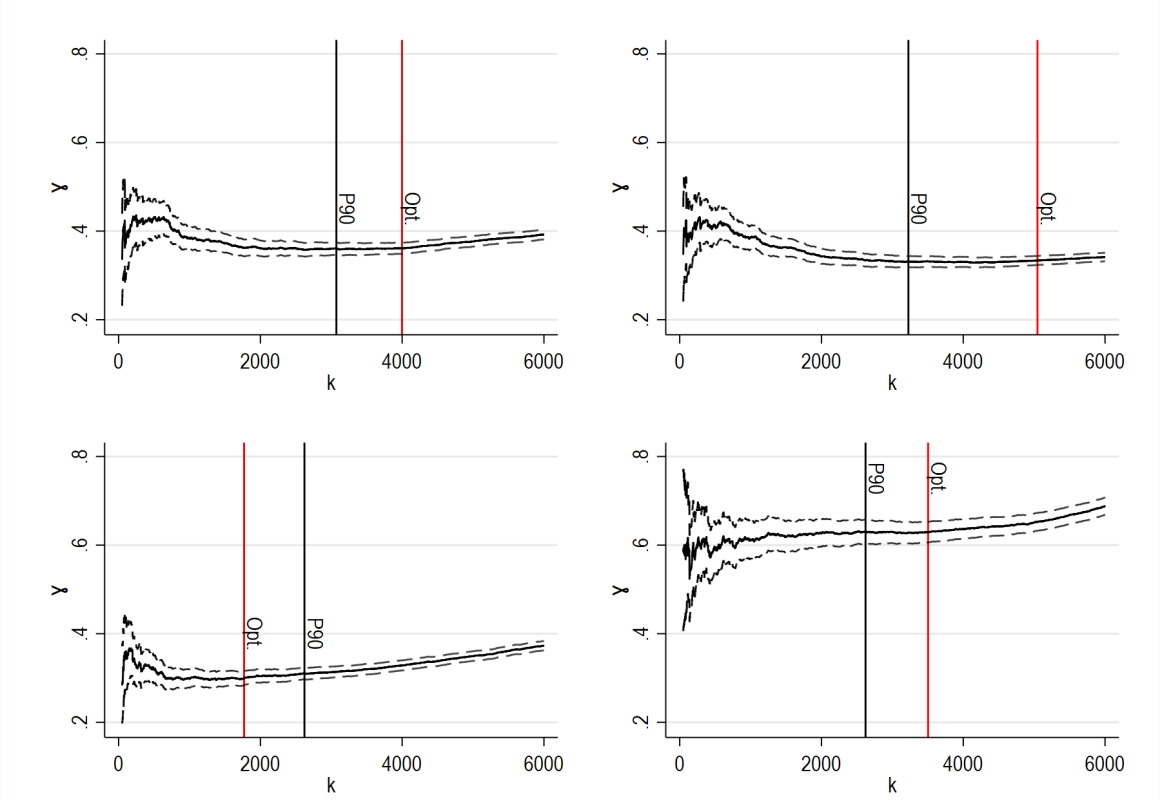
Figure D.5: Hill-type plots for capital and market income



(a) Capital Income (b) Market Income

Notes. Hill-type plots: $\hat{\gamma}$, depicted as a function of the number k of upper-order statistics, is based on the rank-size regression estimator. See Appendix, Section D.4 for the plots for the remaining two income concepts. *Source*: SOEP+P.

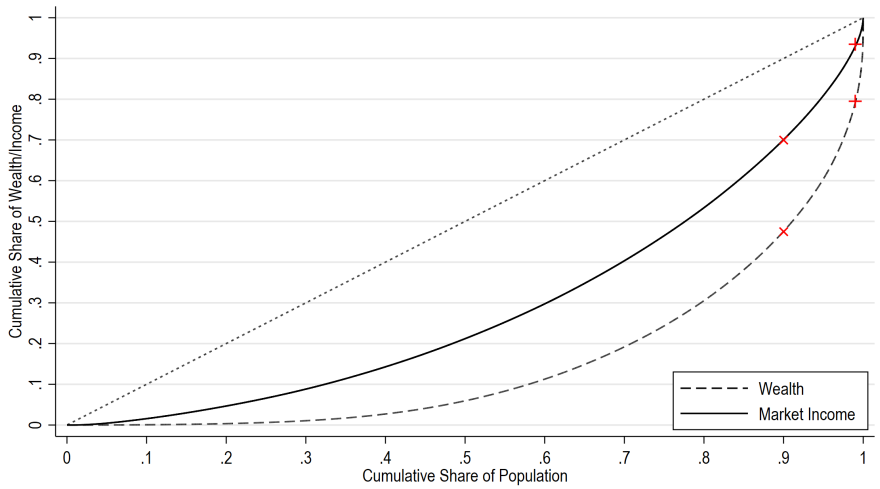
Figure D.6: Hill-type plots of household incomes for SOEP+P



Notes. MktInc is household market income, PostInc is post-government household income, LabInc is household labor income, CapInc is household capital income, which is composed of financial asset returns and income from leasing and renting. Shows rank-size-estimates of γ along different thresholds k. *Source*: SOEP+P.

D.5. The German wealth and income concentration in an international perspective

Figure D.7: Lorenz curves of the entire German wealth and (market) income distribution.



Notes. Lorenz curves for household net wealth and household market income computed from raw data. Red crosses mark the 90th percentile, red plus signs mark the 99th percentile. For precise numbers for the top 10 percent and top 1 percent shares, see Tables 1 and 2. *Source*: SOEP+P.

Having validated the SOEP+P, we pull together all preceding results for wealth and income concentrations and extend them over the entire distribution. Specifically, Figure D.7 depicts the respective Lorenz curves, that is, the plot of the cumulative income or wealth share against the cumulative population share. Hence any top share can be read off, facilitating the concentration comparison between wealth and income. It is evident that wealth in Germany is considerably more concentrated than income. For instance, the poorest 50% hold no noticeable wealth. The top 10 percent wealth share is about twice as large as the respective income share, while the top 1 percent wealth share is about 5 times as large.

As a final benchmark exercise, we ask: How do our concentration results for Germany compare to results for other countries reported in the established literature? The survey by König et al. (2020) demonstrates that such international comparisons are fraught with comparability problems: Assessment units differ, as do the types of data sources and imputation methods. Despite these caveats, an international benchmark, however imperfect, could be useful in assessing orders of magnitude and qualitative differences.

Evidence for top wealth in the United States is presented in Saez and Zucman (2020), who consider tax units and use the so-called capitalization method. They suggest a wealth

share of about 75% for the top 10 percent and 37% for the top 1 percent in the period 2015-20. Smith et al. (2023) use different rates of return and obtain somewhat lower wealth shares (65% for the top 10 percent and 32% for the top 1 percent). Using the Survey of Consumer Finances (SCF) directly, Bricker et al. (2016) report a top 1 percent share of about 33% in 2013. Turning to selective evidence for Europe, the case of France is considered in Garbinti et al. (2021) using the capitalization method combined with direct estimates from survey data. For 2014, they obtain a wealth share of 57% for the top 10% and 24% for the top 1 percent. Applying a similar methodology to Spain in 2013, Martínez-Toledano (2017) report similar wealth shares: 57% for the top 10 percent and 21% for the top 1 percent. Moving from wealth to income shares, Saez and Zucman (2020) suggests an income share of about 18% for the top 1 percent in the United States in 2019. For France, the evidence comes again from Garbinti et al. (2018), who report 32% for the top 10 percent and 10% for the top 1 percent.

To summarize this imperfect benchmarking exercise, our estimated top wealth and income shares for Germany appear quantitatively similar to several other European countries such as France and Spain. By contrast, wealth and income concentrations are larger in the United States. We note that papers appending rich lists to survey data for Germany produce top wealth shares that are more similar to those in the United States than to those in continental Europe. This benchmarking exercise also illustrates the virtues of using survey data with proper oversampling at the top: The unit of observation can be straightforwardly chosen, the variables are clearly defined, and the data preparation procedures can be publicly examined and replicated.

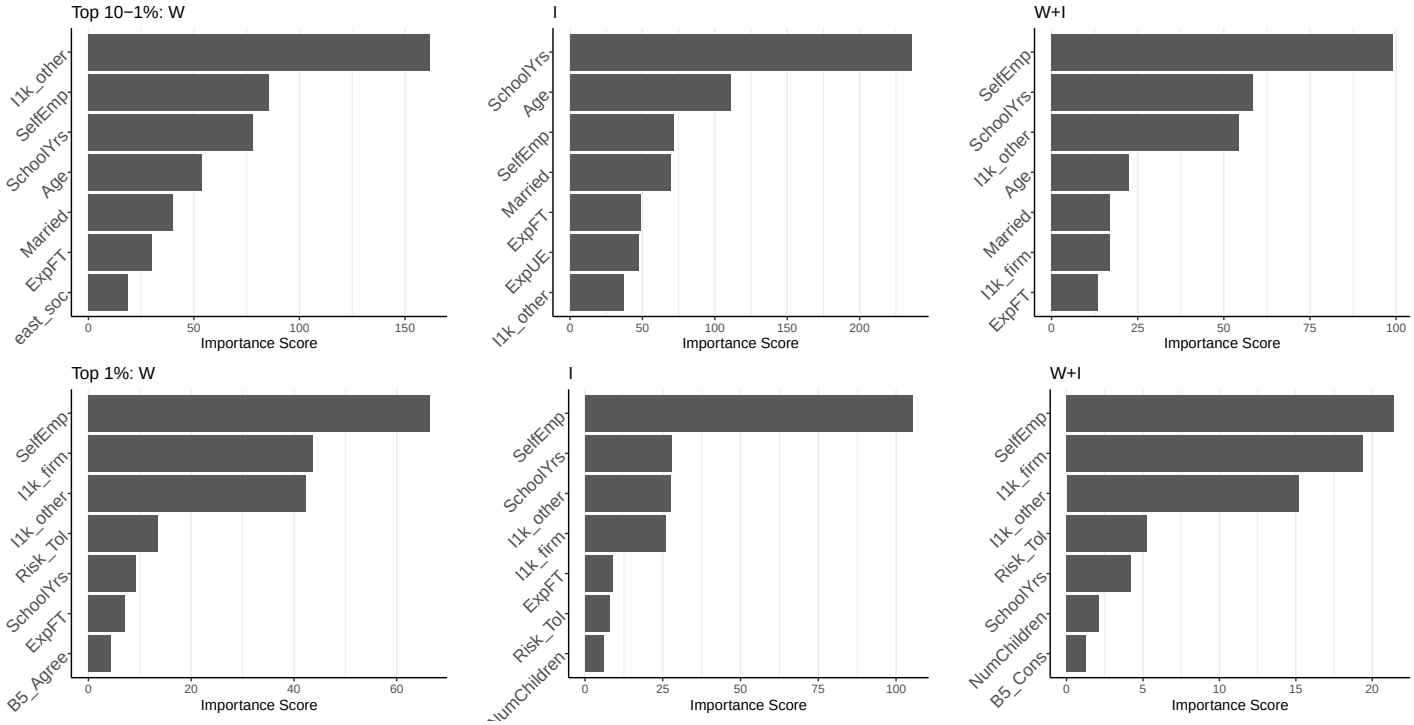
E. Additional empirical results: The classification study

E.1. Variation of the unit of analysis: Random Forest results

In the main text, we conducted our analysis at the level of the household and measured some personal characteristics for the household head. We did so in order to maintain the internal consistency of our analysis across all sections since wealth is measured at the level of the household.

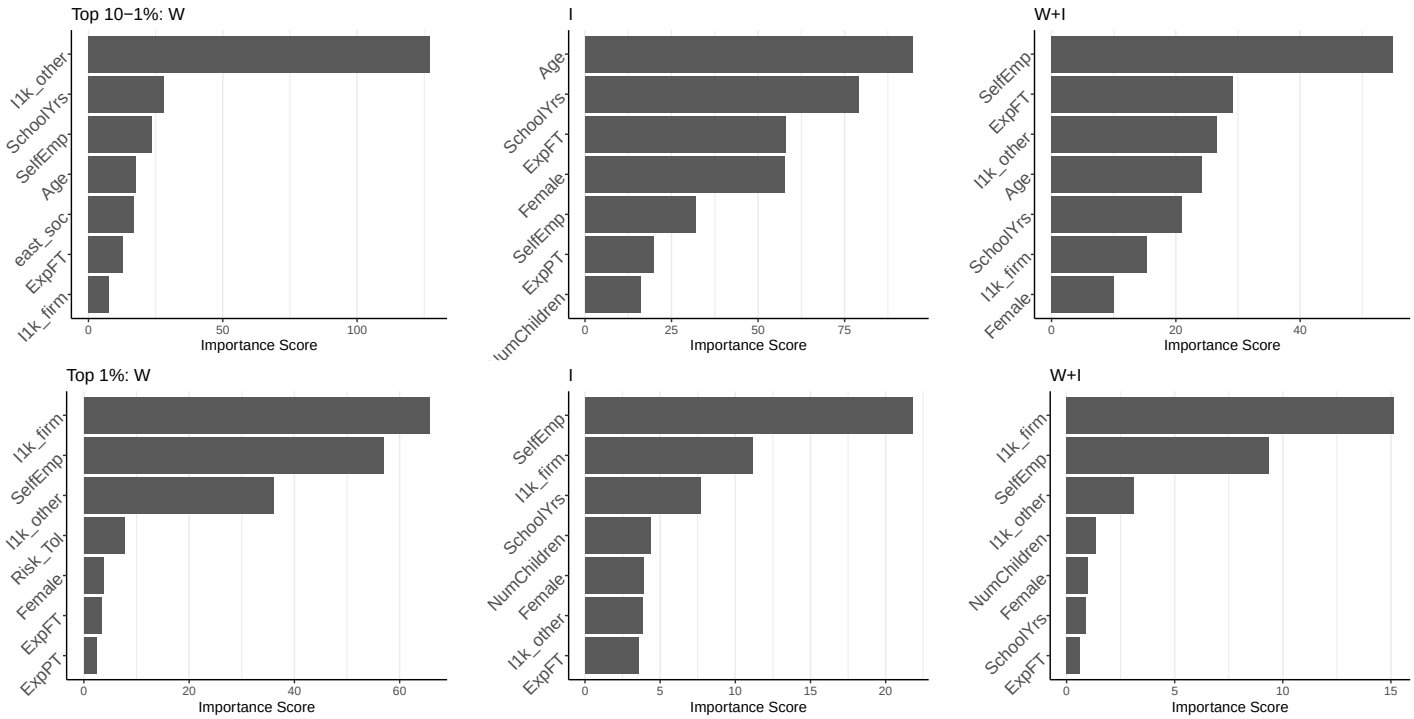
In this section, we establish the robustness of our results when the definition of the sample is changed. In the first experiment, we consider a sample of household heads and their partners. In the second experiment, we conduct the analysis at the level of the individual. For the sake of brevity, we report only the variable importance scores. This analysis parallels section 4.2 of the main text.

Figure D.8: Key predictors: Variable importance scores for top rich group membership for sample of household heads and partners



Notes. Estimations and predictions using random forest models. The importance measures are based on the Gini impurity measure, which has been corrected for the scale of the variables. Variable definitions are given in Table 4. *Source*: SOEP+P.

Figure D.9: Key predictors: Variable importance scores for top rich group membership based on individual-level group definition

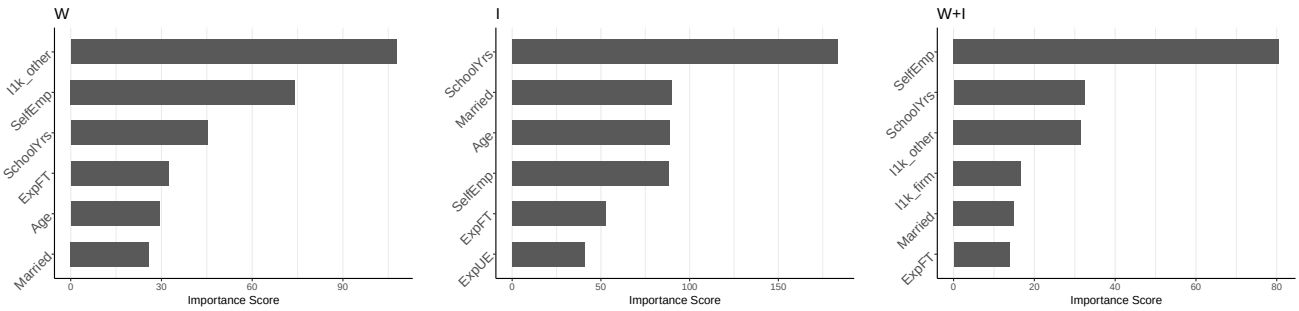


Notes. Estimations and predictions using random forest models. The importance measures are based on the Gini impurity measure, which has been corrected for the scale of the variables. Variable defintions are given in Table 4. *Source*: SOEP+P.

Both Table D.8 and Table D.9 show that for the top 1 percent groups, regardless of whether we consider individuals or household heads and partners, roughly the same set of top two predictors is selected. Self-employment/entrepreneurship and firm inheritances are the most important covariates and other inheritances as well as years of schooling come next.

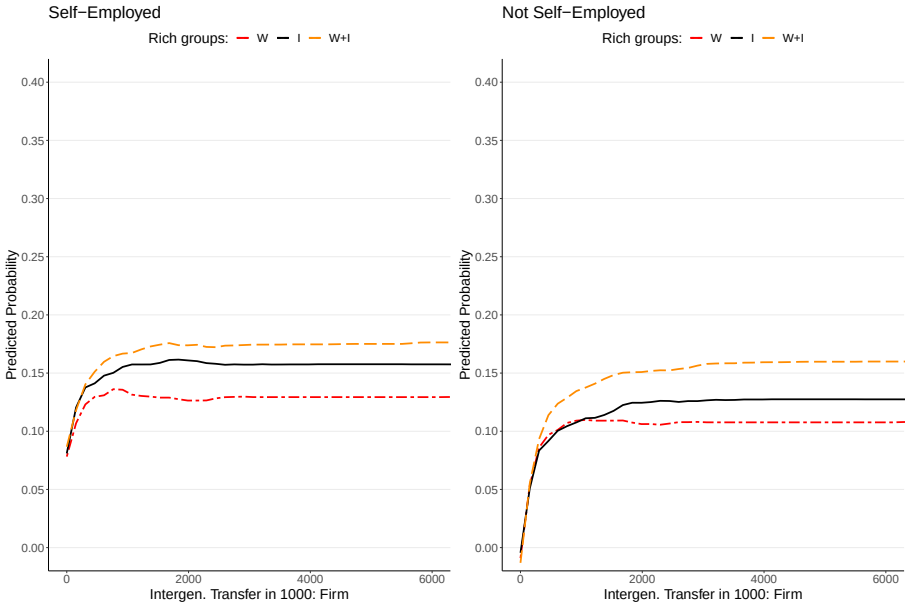
E.2. Classification results for the top 10-1 percent groups

Figure D.10: Key predictors: Variable importance scores for top 10-1 percent group membership

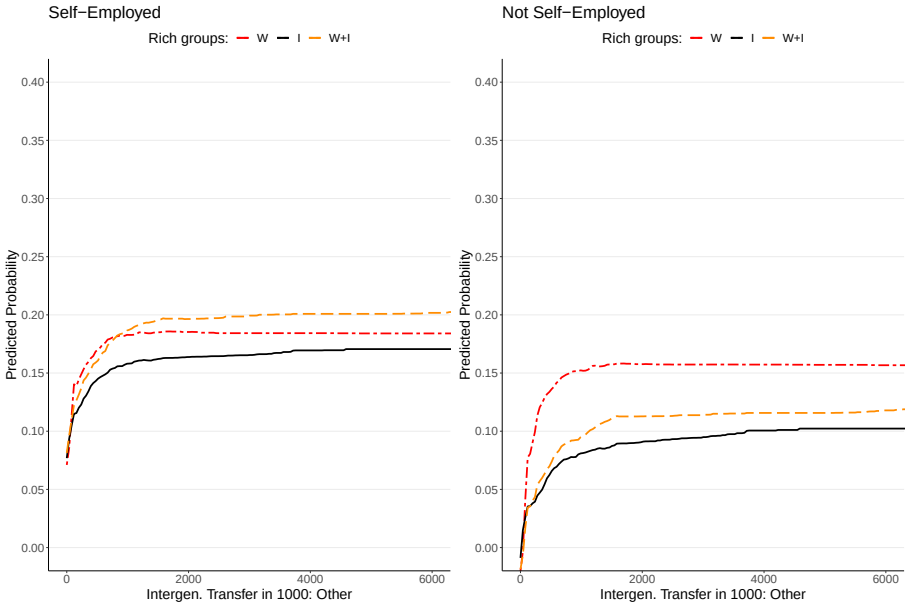


Notes. Estimations and predictions using random forest models. The importance measures are based on the Gini impurity measure, which has been corrected for the scale of the variables. Variable definitions are given in Table 4. *Source*: SOEP+P.

Figure D.11: Partial dependence plots: Inheritances and self-employment



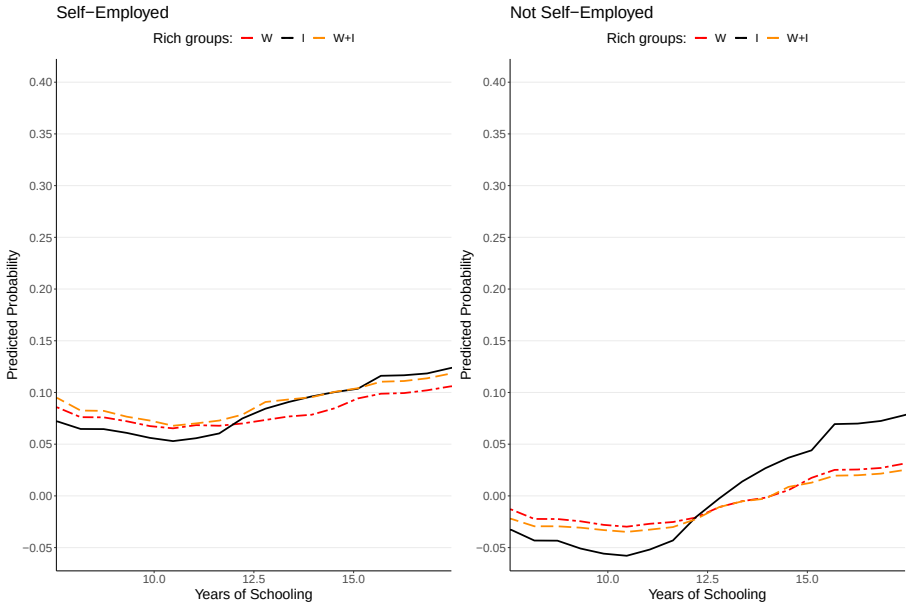
(a) Firm Inheritances



(b) Other Inheritances

Notes. Partial dependence plots of the value of firm inheritances and self-employment for the prediction of being in the Top 10-1 Percent of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

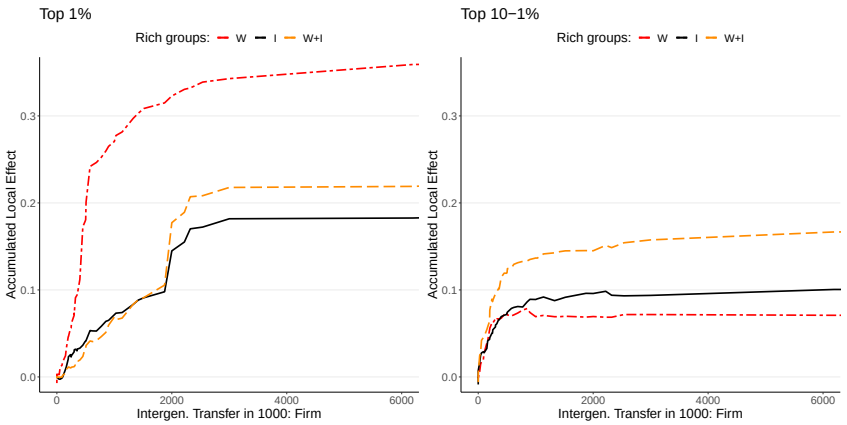
Figure D.12: Partial dependence plots: Years of schooling and self-employment



Notes. Partial dependence plots of years of schooling and self-employment for the prediction of being in the Top 10-1 Percent of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

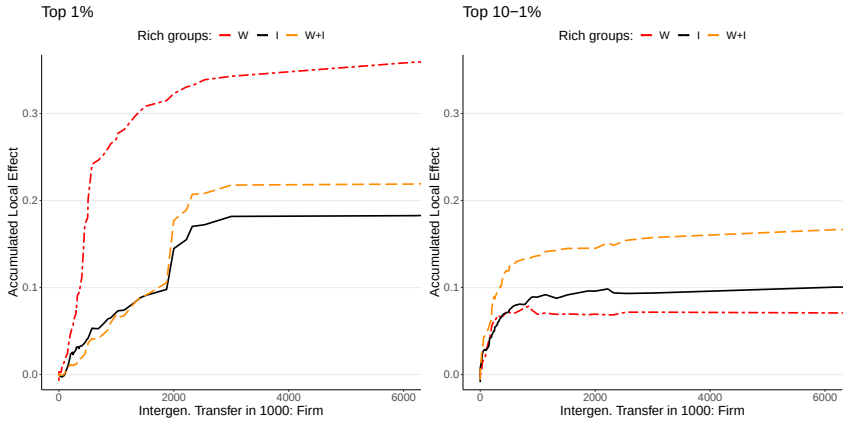
E.3. Additional accumulated local effects comparing top 1 percent and top 10-1 percent

Figure D.13: Accumulated local effect: Firm inheritances



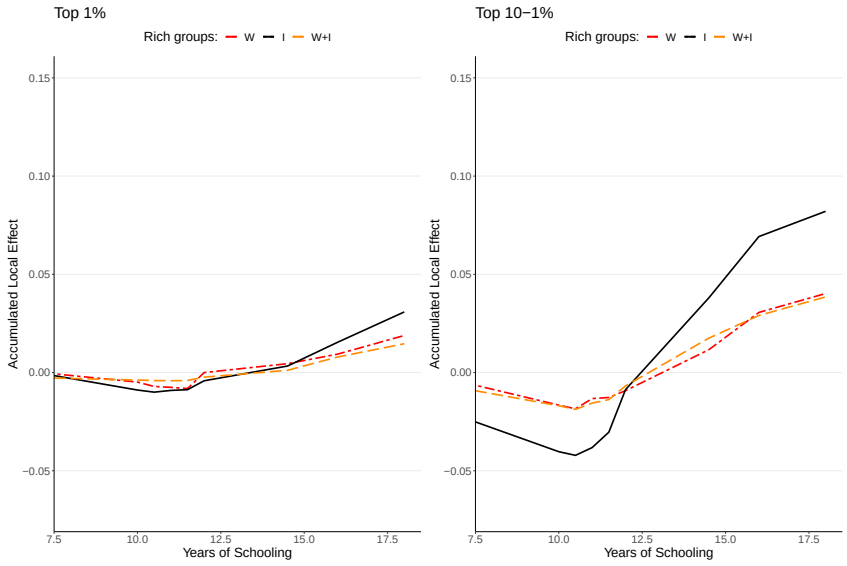
Notes. Accumulated local effects plots of firm inheritances on the prediction of being in the Top 10-1 Percent or the top 1 percent of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

Figure D.14: Accumulated local effect: Other inheritances



Notes. Accumulated local effects plots of other inheritances on the prediction of being in the Top 10-1 Percent or the top 1 percent of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

Figure D.15: Accumulated local effect: Years of schooling



Notes. Accumulated local effects plots of the years of schooling on the prediction of being in the Top 10-1 Percent or the top 1 percent of wealth, income, and wealth and income jointly. *Source*: SOEP+P.

E.4. Additional descriptive statistics

Table D.3: The rich: Descriptors

	Top 1%			Top 10-1%			Bottom 90%		
	W	I	W+I	W	I	W+I	W	I	W+I
Self-Made	0.47	0.48	0.56	0.32	0.38	0.44	0.14	0.13	0.16
Inheritance Ratio	0.29	0.36	0.24	0.38	0.36	0.34	0.56	0.54	0.53
Number of Businesses	1.58	1.75	2.19	1.13	1.09	1.19	1.09	1.11	1.10

Notes. Same as for table 4. The self-made indicator is described in C.4. The inhertiance ratio is the ratio between capitalized inheritances and current net wealth. *Source*: SOEP+P.

F. Binary classification models for routes to the top: Logits

In this section, we report the results of the logit classification models that parallels Section 4 in the main text, where we used random forests. These random forests outperform the logits, and this section provides the relevant comparative performance metrics. Recall that the objective is to predict whether or not an observation is member of a selected rich group such as the top 1 percent in wealth.

We estimate the logit models for each rich group using first the full set of regressors. The empirical model is then back-trimmed, by iteratively dropping the most insignificant regressor and re-estimating the model. The class prediction is then done on the final model. In order to discuss the importance of predictors, we follow common practice in machine learning and adopt a method-specific variable importance measure, which for logits is the absolute value of the t-statistics of a regressor. Note that as this metric is model specific, the variable importance measure cannot be cardinally compared to the variable importance measure used for random forests. The overall predictive performance metric used is the AUC, which is model-independent.

F.1. Predictive performance metrics: ROC and AUC

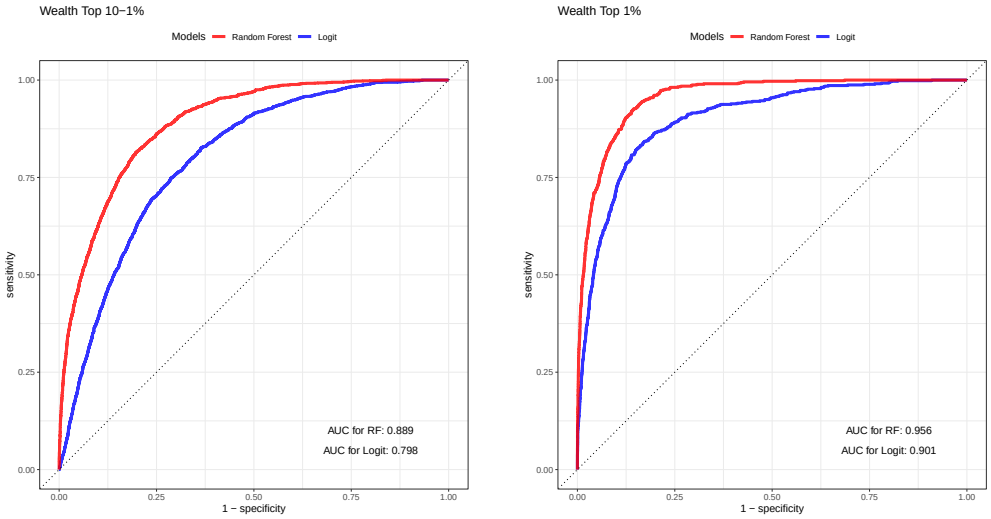
Table E.1 reports the overall AUC performance metric, the optimally chosen number of regressors, as well as the five most important predictors. Using the AUC criterion, it is evident that random forests outperform the logits. For completeness, we depict in Figure E.1 the underlying ROC curves. All ROCs for random forests dominate the ones for the corresponding logits.

Table E.1: Logit models and the five most important predictors

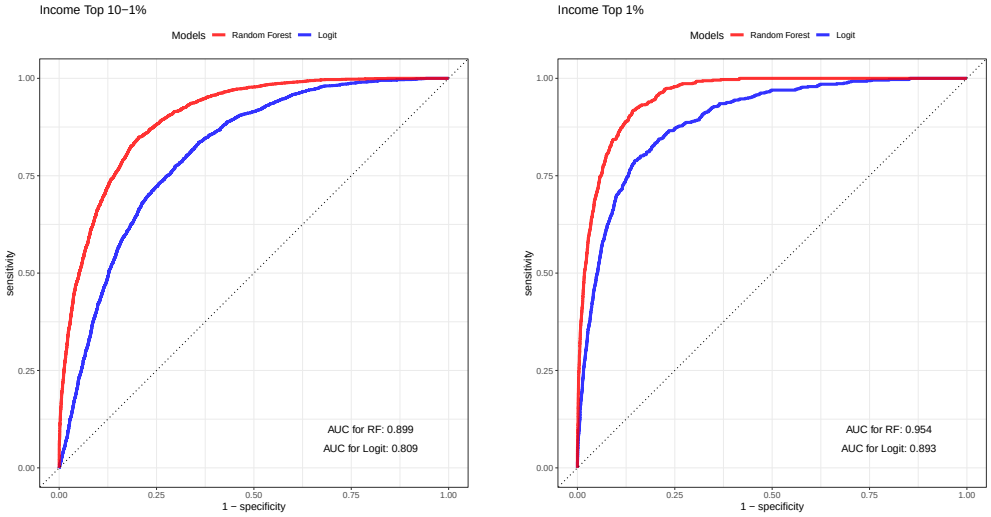
Top	Group	AUC	regressors	var 1	var 2	var 3	var 4	var 5
10-1%	W	0.80	16.00	SelfEmpself-emp	SchoolYrs	Married1	east_soc1	ExpUE
10-1%	I	0.81	14.00	SchoolYrs	Married1	SelfEmpself-emp	Age	ExpFT
10-1%	W+I	0.85	18.00	SelfEmpself-emp	SchoolYrs	Married1	ExpFT	ExpUE
1%	W	0.90	14.00	SelfEmpself-emp	I1k_firm	Risk_Tol	ExpFT	SchoolYrs
1%	I	0.89	13.00	SelfEmpself-emp	SchoolYrs	ExpFT	Risk_Tol	Married1
1%	W+I	0.92	13.00	SelfEmpself-emp	SchoolYrs	Risk_Tol	I1k_firm	ExpFT

Notes. The AUC is the Area Under the Curve measure for predictive performance. "regressor" is the optimally selected numbers of regressors using iterative back-trimming. "var 1"- "var 5" are the five top predictors for each group, ordered by the method-specific variable importance measure (the absolute value of the t-statistic).

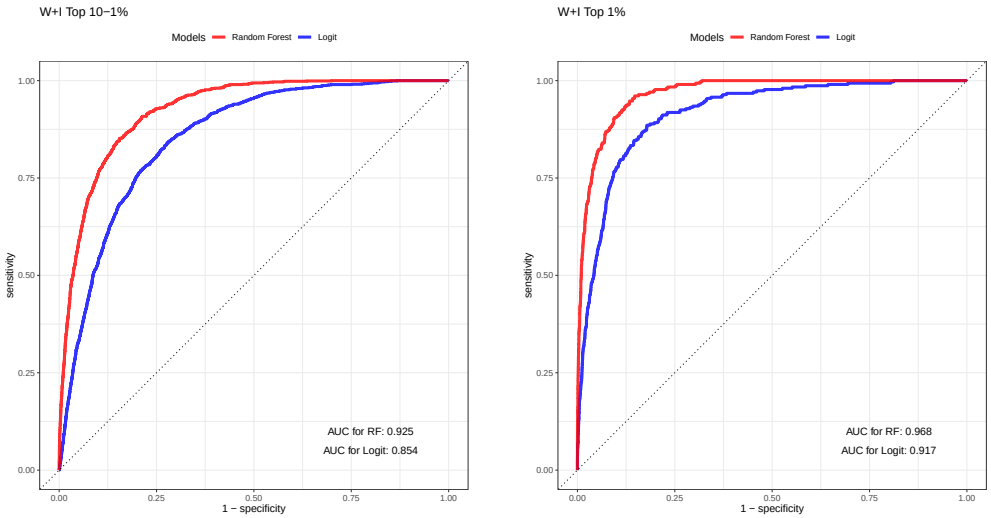
Figure E.1: Model performance: Random forests vs. Logits



(a) Wealth



(b) Income



(c) Wealth and Income

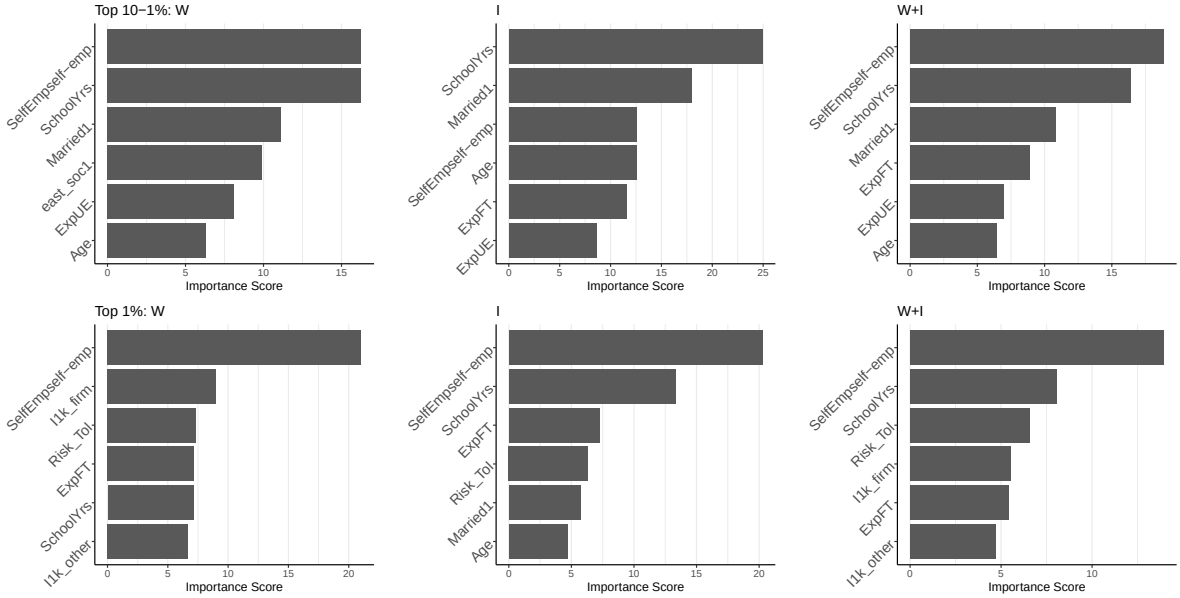
Notes. Receiver Operating Characteristic curves and Area Under the Curve (AUC) for fitted random forests and logit models. I refers to the model for the top 1 percent of income, W to the top 1 percent of wealth, and $W{+}I$ the joint top 1 percent. *Source*: SOEP+P.

F.2. Logits: Variable importance

Next, we turn to the variable importance scores for the logit models. Here, the interest is whether the suggested importance ordering differs substantially from the ordering in the random forest models. Table E.1 reports the five most important regressors.

Overall, the same ensemble tends to be selected, with minor differences in the importance ordering. For instance, self-employment is considered the top predictor except for the Top 10-1 Percent income group. Education tends to be accorded greater importance in logit models than in the random forest ones. For completeness, Figure E.2 reports the actual variable importance scores for the top six predictors.

Figure E.2: Logits: Top variable importance scores



Notes. The variable importance score for logits is given by the absolute value of the t-statistic of a regressor.

Source: SOEP+P.

F.3. The complete set of logit estimates

For completeness, we report the full set of logit coefficient estimates for all six rich group classification models. Empty spaces relate to regressors that were dropped by the back-trimming procedure.

Table E.2: Logit models: Selected models

	W Top 10-1%	I	W+I	W Top 1%	I	W+I
SelfEmpself-emp	1.028***	0.750***	1.361***	2.046***	1.911***	2.019***
	(0.063)	(0.060)	(0.072)	(0.097)	(0.094)	(0.145)
I1k_firm	−0.0001	−0.0001	−0.0001**	0.002***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.00004)	(0.0002)	(0.0002)	(0.0002)
Risk_Tol	0.047***		0.065***	0.172***	0.140***	0.217***
	(0.013)		(0.016)	(0.023)	(0.022)	(0.033)
ExpFT	0.017***	0.054***	0.058***	0.034***	0.051***	0.058***
	(0.004)	(0.005)	(0.007)	(0.005)	(0.007)	(0.011)
SchoolYrs	0.151***	0.229***	0.204***	0.115***	0.219***	0.191***
	(0.009)	(0.009)	(0.012)	(0.016)	(0.016)	(0.024)
I1k_other	0.0003***		0.0002***	0.001***	0.0003***	0.0004***
	(0.0001)		(0.0001)	(0.0002)	(0.0001)	(0.0001)
ExpUE	−0.249***	−0.326***	−0.567***	−0.835***	−0.587***	−0.812***
	(0.031)	(0.038)	(0.082)	(0.151)	(0.126)	(0.243)
NumChildren	0.086**	0.055*	0.129***	0.213***	0.141***	0.230***
	(0.033)	(0.029)	(0.039)	(0.054)	(0.051)	(0.072)
east_soc1	−0.755***	−0.458***	−0.580***	−0.430***	−0.476***	−0.623***
	(0.076)	(0.067)	(0.093)	(0.129)	(0.122)	(0.187)
B5_Extra		0.062**	0.149***	0.119**	0.075	
		(0.027)	(0.038)	(0.051)	(0.048)	
Married1	0.711***	1.146***	0.960***	0.217**	0.639***	0.403**
	(0.064)	(0.064)	(0.088)	(0.110)	(0.112)	(0.158)
B5_Agree	−0.082***	−0.094***	−0.051	−0.092*		−0.119*
	(0.027)	(0.026)	(0.035)	(0.047)		(0.063)
B5_Neuro	−0.062**	−0.107***	−0.105***	−0.103*	−0.167***	−0.108
	(0.029)	(0.027)	(0.039)	(0.053)	(0.050)	(0.073)
Female1	−0.215***		−0.271***	−0.225*		
	(0.069)		(0.089)	(0.119)		
Age	0.020***	−0.052***	−0.036***		−0.032***	−0.025**
	(0.003)	(0.004)	(0.006)		(0.007)	(0.010)
ExpPT	0.021***	0.051***	0.053***			
	(0.005)	(0.006)	(0.008)			
B5_Open	−0.068**		−0.119***			
	(0.028)		(0.038)			
B5_Cons		0.068**	0.058			
		(0.029)	(0.038)			
Constant	−6.173***	−4.263***	−6.015***	−7.433***	−7.580***	−9.093***
	(0.213)	(0.170)	(0.269)	(0.341)	(0.359)	(0.542)

Notes. Coefficients for selected logit models. *Source*: SOEP+P.