König, Johannes; Schluter, Christian; Schröder, Carsten



# Routes to the Top

Review of Income and Wealth

**Provided in Cooperation with:**
John Wiley & Sons

*Suggested Citation:* König, Johannes; Schluter, Christian; Schröder, Carsten (2025) : Routes to the Top, Review of Income and Wealth, ISSN 1475-4991, Wiley, Hoboken, NJ, Vol. 71, Iss. 2, https://doi.org/10.1111/roiw.70015

*Review of Income and Wealth*

ORIGINAL ARTICLE OPEN ACCESS

# Routes to the Top

Johannes König[1] | Christian Schluter[2,3] | Carsten Schröder[1,4]

[1]DIW Berlin, Berlin, Germany | [2]Aix-Marseille Université (Aix Marseille School of Economics), CNRS & EHESS, Marseille Cedex 01, France | [3]Department of Economics, University of Southampton, Highfield, Southampton, UK | [4]Department of Economics, Freie Universität Berlin, Berlin, Germany

**Correspondence:** Carsten Schröder (cschroeder@diw.de)

**Accepted:** 15 April 2025

**Keywords:** intergenerational transfers | predictions | rich-group classification modeling | top income | top wealth

## 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.

**JEL Classification:** D31, C38, D63

## 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; Auten and Splinter 2024). 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 policymakers 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 that 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 nonrich 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 in the wealth distribution, 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. Second, 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 of the top 1 percent in the income distribution make at least 203,000 euros, and hold a share of 7% of total income. This 7% income share 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 the individual-level administrative income data from the Sample of Integrated Labor Market Biographies.[1]

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 stands out as a group of great 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 nonlinearities 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 modeling and its clear empirical findings complement research efforts to isolate causal relationships through random variation in covariates.[2] In the analysis of the determinants of wealth, the identification of credible natural experiments is still rare,[3] and our predictive analysis contributes to this endeavor by guiding this search and revealing which sources of variation are most important. We proceed to comment first on the statistical method, before discussing the empirical results and placing them in the context of the established literature.

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 nonlinearities and variable interaction.[4] 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 modeling 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 that 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 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. 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 is the strongest predictor 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.

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, 2022; Fagereng et al. 2021; 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.

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). For Italy, Acciari et al. (2024) observe that the accumulation of wealth at the top is primarily driven by financial and business assets. However, the key difference we highlight is the importance of intergenerational transfers of company assets in Germany.

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); see also the extensive discussion in Kennickell (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 (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 largest 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; 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, 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 Appendices A–G available as Supporting Information (online) contain extensive Supporting Information.

## 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 wealth and income data used in our distributional analysis are based on survey responses to the SOEP questionnaire. Appendix C.1 details the composition of individual personal balance sheets, which cover twelve wealth items (assets as well as debt positions). Hence, valuations of assets, including those of closely held businesses or real estates, are made by interview subjects. As a consequence, wealth portfolios are measured *directly* and, except in the case of item nonresponse, not imputed. Wealth shares by asset class and rich group are also further discussed below in Section 4.3. Appendix C also defines precisely all income concepts and the calculation of inheritances. The wealth and income data are validated below in the next sections.

Next, we briefly summarize the sampling framework used in generating the SOEP-P sample. The data documentation paper for SOEP-P, (Siegers et al. 2021), provides extensive details. The point of departure is the observation that many studies in the wealth literature, examining the portfolio composition among the wealthy, document high incidences and levels of business assets (see, e.g., Bucks et al. 2009; Bricker et al. 2017; Martínez-Toledano 2020; Wolff 2021; Smith et al. 2023). For example, Wolff (2021), using the Survey of Consumer Finances for the United States, which can draw on register data for top tail sampling, suggests that 94% of those in the top tail hold business wealth. This means that only a small minority of the wealthy does not hold business assets. Our target population is thus composed of individuals with a significant share of their wealth in company shares, and we identify members of this target population using the global company database ORBIS (which contains information on the financial situation and ownership structures of more than 400 million companies worldwide). More precisely, 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 the relevant business registers. The minimum threshold for having substantial shares and thus being listed is 0.1% of all shares of a company. Note that, even when an individual's wealth is invested predominantly in another type of asset (e.g., 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, and our descriptive analysis in Section 3.2 confirms that a significant share in our new sample are not entrepreneurs. On the other hand, less successful entrepreneurs who do not make it to the "top" are members of the sampling framework for the standard SOEP.

The determination and stratification of the target population (by business wealth) required information on all of the shareholders' business wealth. The market capitalization of many of the companies these shareholders owned was not known, leading to proxying firm values by company turnover. Note, however, that an exactly estimated value of business wealth according to ORBIS is not needed, an ordinal measure being sufficient, since wealth portfolios, as explained above, are directly elicited from the SOEP interview subjects.

From all these shareholders, a probabilistic sample of individuals was drawn, stratified according to the value of their shareholdings.[5] 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 computation of sampling weights is detailed in Siegers et al. (2021). The sampled individuals—and their household members—were then surveyed using the standard SOEP questionnaire. SOEP-P is therefore 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. 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.[6] 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.[7]

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 Appendix A.1 for a formal exposition). Figure 1a 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 0.55. As is evident in Figure 1a, 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.

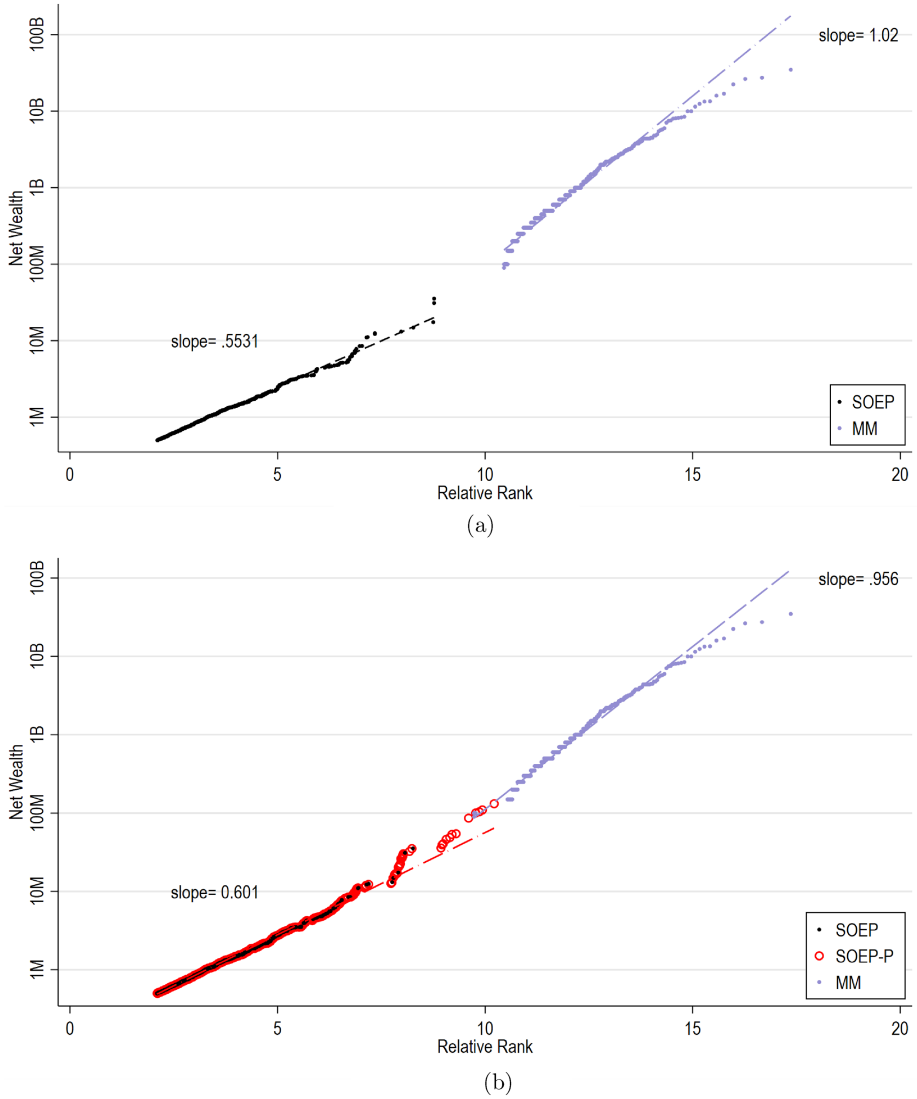


**FIGURE 1** | Pareto QQ-plots for top wealth: Filling in and adding tail observation. (a) Without SOEP-P. (b) With SOEP-P. Notes. Panel (a): Upper wealth order statistics of the German Socio-Economic Panel (SOEP, black) in 2019, and the national rich list (Manager Magazin, MM, purple). Panel (b) includes 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.

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

## 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 semiparametric 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,[8] the distributional theory of which has recently been developed in Schluter (2018).

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 Appendix A.6 for details).[9]

We consider as benchmarks some estimates reported in the literature.[10] 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 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 $\alpha$ estimate is 1.8 (and $1/1.8 = 0.56$ as reported in Figure 1A), implying a top 1 percent wealth share of about 20%. Increasing the wealth threshold to 2 million euros decreases the estimate of $\alpha$ (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, 0.097 when the threshold is 1 million euros and 0.167 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).

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 $\alpha$ 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 the 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 0.042 (instead of 0.167), 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, that is, the asymptotic mean-squared error of the estimator is minimized. The statistical theory, developed in Schluter (2018, 2020), is summarized in Appendix A.4.[11] The optimal wealth threshold is estimated to be about 0.4 million euros, leading to an estimate of 1.66 for $\alpha$ 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., 0.042 for a 2 million euros threshold). The left and right 95% confidence limits are 1.60 and 1.72, respectively.[12]

**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 |

*Note:* 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 Appendix A.2. $k$ denotes the number of upper-order statistics corresponding to the fixed threshold. Appendix A.6 gives details for the computation of the wealth shares, Appendix Table D.2 reports the confidence limits for the wealth shares. *Source*: SOEP+P.

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 datasets 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.[13] 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 underrepresented in leading surveys,[14] and that an appropriate model is given by Equation (1). Data augmentation using national rich lists cannot remedy this problem, however, since their focus is on wealth and not on 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 postgovernment income. See Appendix C for detailed income definitions.

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 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 0.056 to 0.049. 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 pregovernment income distribution is heavier than that for postgovernment 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, postgovernment, 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.

**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 |

*Note:* Income concepts: *MktInc* is household market income, *PostInc* is postgovernment 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 Appendix A). *Source*: SOEP+P.

Our estimated top income share for household market income and the top 1 percent are 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, rare exceptions are Martinez (2021) using Swiss tax data, Garbinti et al. (2021) using French data, Ozkan et al. (2023) using Norwegian data, Acciari et al. (2024) using Italian inheritance tax data), and Bricker et al. (2020) and Fisher et al. (2022) using the US Surveys of Consumer Finances. 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 nonparametrically using (Spearman's) rank correlations

for the four income concepts. Throughout, the marginal distribution of wealth is denoted by $F_W$ and that of generic income is denoted 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 0.72, and between wealth and market income is 0.58. It is slightly lower for labor income at 0.41 unconditionally and at 0.55 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.[15] It turns out, as detailed in Appendix B and our extensive goodness-of-fit analysis, that the one-parameter Gumbel copula, say $C_\theta$, describes the dependence structure across the entire distribution very well in the German case.[16] 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 Appendix B.2 for the detailed computation). Table 3 reports these wealth and income shares.

**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 ($1\% = 100 \times (.1)^2\%$). 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.

**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 |

*Note:* 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 Appendix B.2. *Source*: SOEP+P.

**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 at 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: sociodemographics, 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,[17] 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) 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). Levine and Rubinstein (2017) examine traits of US entrepreneurs and conclude that a particular mixture of "smart" and "illicit" characteristics explains sorting into entrepreneurship. Risk tolerance has been shown to matter for entrepreneurial survival (Caliendo et al. 2010) and entrepreneurial investment (Fossen et al. 2024) 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).[18] 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.2 show.[19]

#### 3.2.1 | Intra- and Intergroup 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 nonrich 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.

**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 times 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 set 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[20] that are designed to

**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 |

*Note:* Weighted means with groups defined by their household-level position in the wealth and income distributions. Sample is 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. For the top 0.5% group, see Appendix H. *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.

fit complex data relationships, in particular nonlinearities and covariate interactions, without simply overfitting, and perform well on not-yet-seen data. 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. Since our analysis examines static correlations, we do not, of course, seek to infer dynamic patterns or relationships. In order to *interpret* the role of the top predictors, we use model-agnostic and model-specific importance metrics. Presented in Section 4.2 below, these metrics paint a coherent picture of the top predictors for rich group membership. Section 4.3 summarizes 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 the 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 are 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 sample is computed using the Gini impurity measure.[21] 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.[22] All trees presented in Figure 2 are optimally pruned in this way. 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. 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.

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.

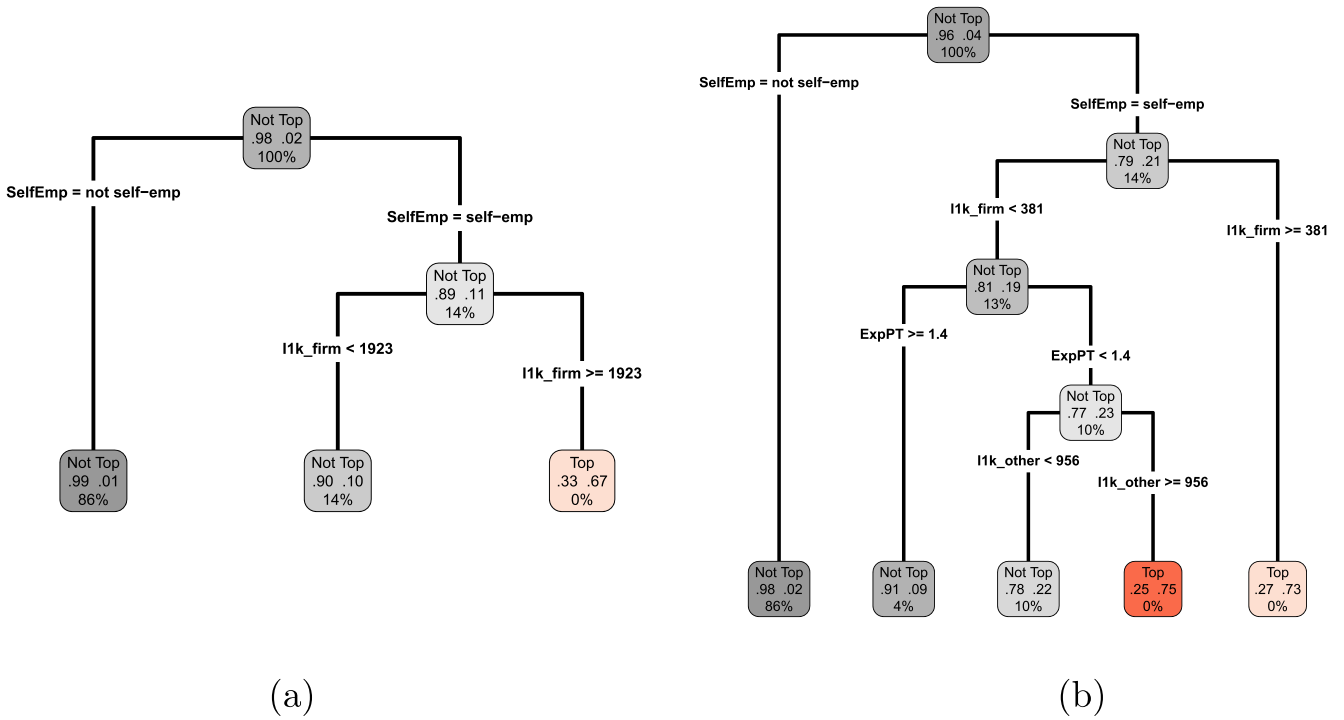


**FIGURE 2** | Single optimally pruned classification trees for top 1 percent group membership. (a) Wealth and Income. (b) Wealth *Notes.* Optimally pruned classification trees for being in the top 1 percent of: wealth and income (panel a), wealth (panel b). For income, see Appendix F. The pruning algorithm is explained in footnote 22. A single node reports 1) the dominant group for this node in line 1, 2) the partition into not-top/top in line 2, and 3) the overall population share in line 3. 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%. Along the branches of the tree, we can read the splitting criterion that was used to create the child nodes, for example, self-employment status being true or false. Variable definitions are given in Table 4. *Source*: SOEP+P.

**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 decorrelated deeply grown classification trees on bootstrapped training samples. Each time a split in a tree is considered, a random sample of $m$ predictors is chosen as split candidates. This $m$ is a parameter of the algorithm and is 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 decorrelates individual trees in the forest. Averaging across this ensemble of trees, the *variance* of the estimator is reduced substantially.[23] Consequently, "random forests do not overfit" (Breiman, 2001). We 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 G), focusing 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.

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,[24] we obtain the ROC (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 = 0.956) clearly outperforms the logit (AUC = 0.901). In the next section, we change the perspective from global measures of predictive performance to variable-based metrics.

**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 |

*Note:* "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.

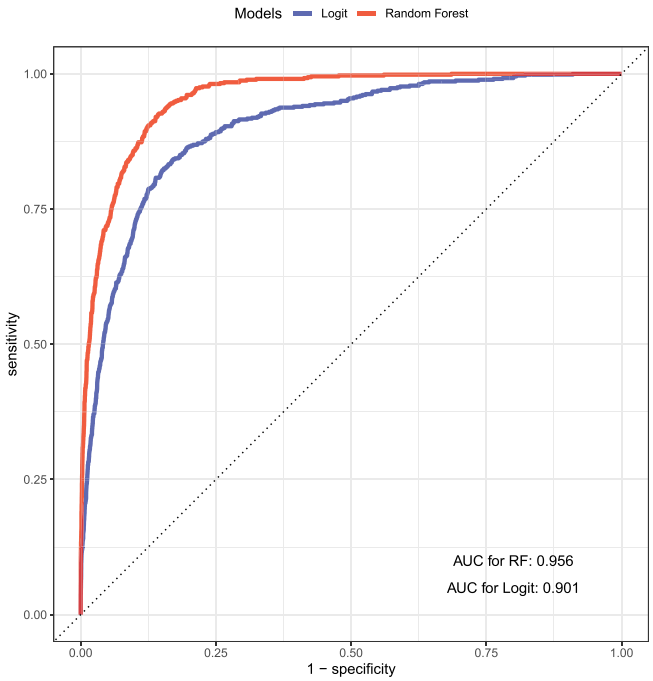


**FIGURE 3** | ROC curve for top 1 percent of wealth (RFs vs. logits). *Notes.* AUC is the area under the ROC curve. A higher AUC implies greater predictive power. See the main text for a description of the ROC. At every classification threshold the random forest ROC curves lies above the logit ROC curve, indicating that the random forest model correctly predicts more true positive cases no matter the threshold. See Appendix G for a detailed analysis for all rich groups. *Source*: SOEP+P.

## 4.2 | Results: Identifying the Best Predictors Using Variable Importance Measures

The dual task of interpreting the Random Forest results and identifying the key predictors is achieved by computing variable importance metrics. The Variable Importance (VIMP) score is directly based on the node splitting criterion for the individual classification trees making up the RF (see footnote 21), and averaged across all trees, yielding thus a model-specific measure. In Appendix E.1 we consider complementary but model-agnostic importance measures (such as partial dependence plots (PDPs) and accumulated local effects (ALEs)), and compare the predictive performance of the RF with and without the key predictors. There we confirm that all these metrics rank the key predictors coherently. For the sake of brevity, we can thus concentrate in this section on the VIMP results, and more specifically, on the ranking of predictors for each rich group.

Figure 4 reports the ordered VIMP 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 RF.

More specifically, for all the top 1 percent groups, self-employment/entrepreneurship is the most important

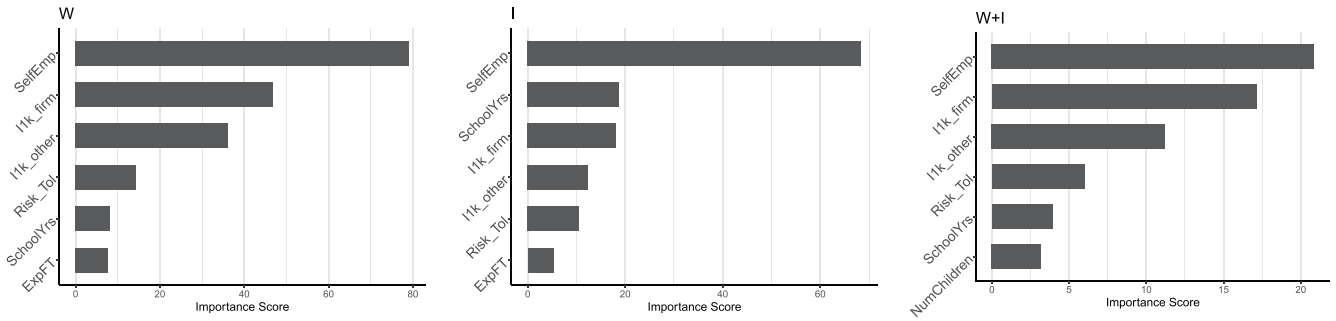


**FIGURE 4** | Key predictors: Variable importance scores (VIMPs) for top 1 percent group membership. *Notes.* Estimations and predictions using random forest models. The importance measures are based on the Gini impurity measure, which has been corrected for the scale of the variables. For the top 0.5% group, see Appendix H. Variable definitions are given in Table 4. *Source*: SOEP+P.

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 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.3, 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.

Our alternative variable importance metrics, presented in Appendix E.1, confirm these insights. There, we consider partial dependence plots (PDPs), which focus on average marginal effects on outcomes by changing the value of one variable at a time. We show 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. By contrast, PDPs of educational attainment are comparatively very flat. Accumulated local effects (ALEs) also focus on marginal effects, but take into account the potential correlation between predictors (and thus have the benefit of avoiding unrealistic variable combinations). We show that the plotted ALEs are qualitatively very similar to the PDPs. In Appendix E.3, we repeat this analysis for the top 10-1%.

### 4.3 | Discussion and Summary

Who are "the rich" in Germany? Our nonparametric 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) leads 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 interrich 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 *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.[25] We also note that 56% in the joint top 1 percent consider themselves to be "self-made"[26], despite the received intergenerational transfers (see Appendix Table C.2 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 wealth and 0.38 for income.

This systematic difference between the joint top 1 percent and the other rich groups is illustrated further in Figure 5, 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 6 depicts the incidence of firms in sole ownership, and reveals that across all groups, the rich predominately own one firm.

Finally, in Figures 7 and 8, 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 fewer 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

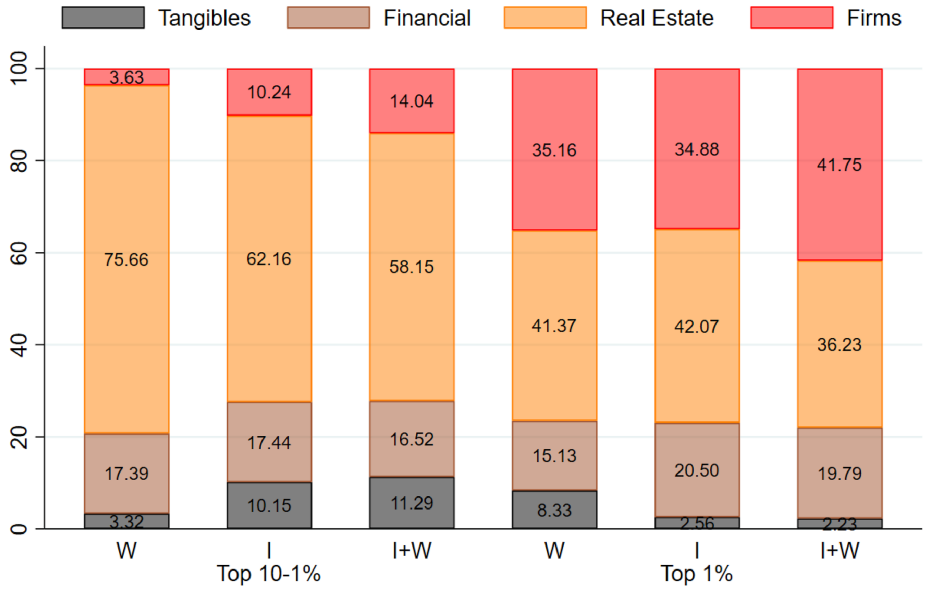


**FIGURE 5** | Gross wealth shares by asset class and rich groups. *Notes.* Composition of the gross wealth portfolio within each of the rich groups. The shares in percent give the aggregate contribution to the wealth total within each group. Shares were computed using household survey weights. Tangibles are objects of high value, such as paintings, jewelry, and cars. 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.

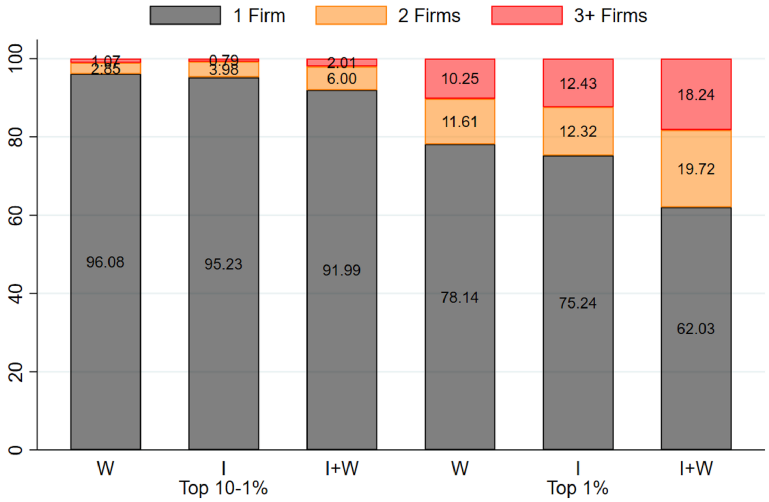


**FIGURE 6** | Firms in sole ownership by rich groups. *Notes.* Shows shares in percent of those in a given category of a number of firms owned, conditional on owning at least one firm for each of the six rich groups. Shares were computed using household survey weights. *Source*: SOEP+P.

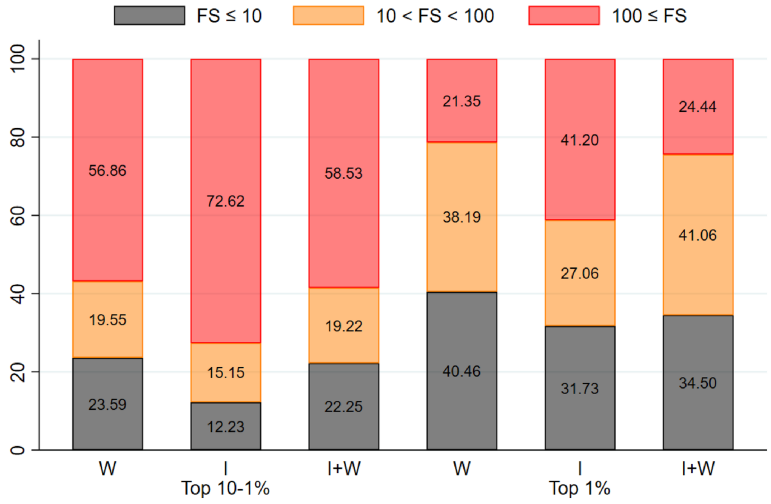


**FIGURE 7** | Firm size categories by rich groups. *Notes.* Shares in percent of those within a firm size category for each of the six rich groups. A sample is conditional on being currently active on the labor market. Shares were computed using household survey weights. *Source*: SOEP+P.

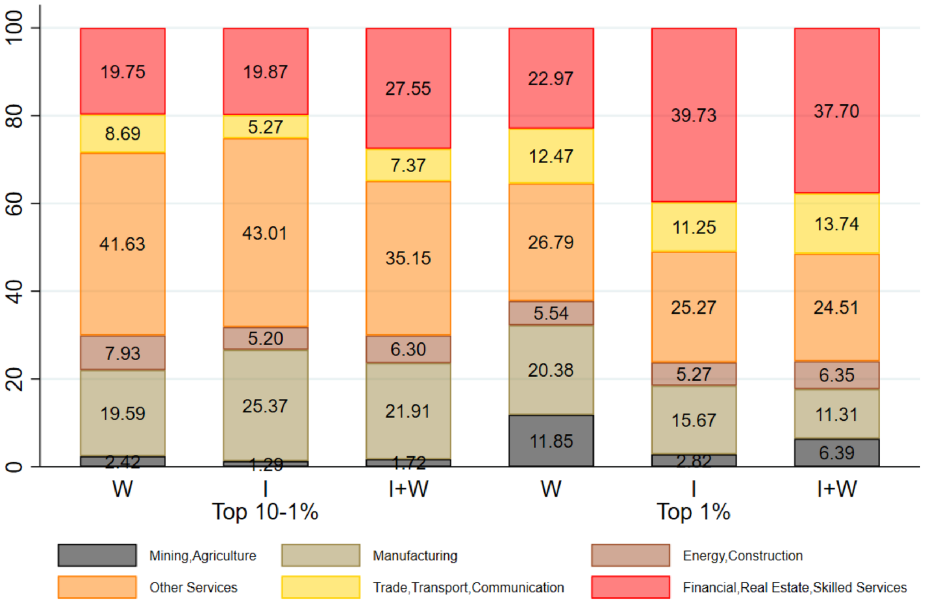


**FIGURE 8** | Industry composition by rich groups. *Notes.* Shares in percent of those within a certain industry category for each of the six rich groups. A sample is conditional on being currently active on the labor market. Shares were computed using household survey weights. *Source*: SOEP+P.

with sizable firm inheritances. According to these variables, we can order the rich groups along an *entrepreneurial spectrum*. At one end 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 end 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 midmarket firms in skill-intensive industries. More specifically, these authors estimate that up to 40% of income in the top 1 percent derive 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, but 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.

**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. Open Access funding enabled and organized by Projekt DEAL.

**Endnotes**

[1] In Appendix D.5, we depict the entire wealth and income distribution using Lorenz curves, and place our concentration measures for Germany in an international context by comparing them to results for other countries reported in the established literature. The level of wealth concentration in Germany appears to be in line with recent evidence, other European countries, but is lower than in the United States.

[2] 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, specifically the use of ML techniques to this end, has 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.

[3] An exception is Nekoei and Seim (2023), who study the impact of the random timing of the receipt of inheritances. A different literature examines lottery wins as natural experiments of exogenous wealth transfers (e.g., Hankins et al. (2011)), while noting that lottery participation tends to decrease with income.

[4] 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).

[5] SOEP-P's sampling frame used the obtained estimates of shareholdings for stratification were as follows: 4/7ths of the gross sample within

the target population come from the highest tercile of estimated values, 2/7ths from the middle tercile, and 1/7ths from the lowest tercile. Across the terciles, the reported business assets of the SOEP-P respondents differ considerably: The tercile-specific median (mean) of business assets is about 0.83 million Euro (2.6 million Euro) in the highest tercile, about 0.42 million Euro (0.87 million Euro) in the second, and 0.29 million Euro (0.82 million Euro) in the bottom tercile.

[6] 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.

[7] 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.

[8] 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 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.

[9] We have made these methods available via a Stata ado-package called `beyondpareto`. See the vignette posted at https://christianschluter.github.io/beyondpareto.

[10] 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).

[11] 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.

[12] In 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.

[13] 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.

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

[15] 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 $\rho$ depends only on the underlying copula and can be interpreted as a moment of the copula. See, for example, Nelsen (2006) for an extensive textbook treatment.

[16] 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 $\theta$ of the Gumbel copula is obtained by inverting the theoretical mapping of the rank correlation $\rho$ and $\theta$ and evaluating it at the empirical $\rho$, thus yielding a method-of-moments estimate. Appendix Table B.1 reports the estimates of $\theta$.

[17] 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.

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

[19] 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 rerun 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.

[20] See, for example, James et al. (2021) for a textbook treatment of statistical learning.

[21] 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.

[22] 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 0.5. 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 the 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.

[23] To see the principal insights, recall that for $X_i \sim (\mu, \sigma^2)$ with $cor(X_i, X_j) = \rho$ and $i = 1, \ldots, N$, the variance of the sample mean is $var(\overline{X}) = \rho\sigma^2 + \frac{1-\rho}{N}\sigma^2 \to \rho\sigma^2$ as $N \to \infty$. The bias of a RF is the same as the bias of any of the individual sampled trees, which is minimized by growing trees deeply. The covariance between any two trees is reduced by the decorrelation trick, thus reducing the generalization error of the ensemble relative to a single tree.

[24] 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.

[25] 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).

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

## References


Acciari, P., F. Alvaredo, and S. Morelli. 2024. "The Concentration of Personal Wealth in Italy 1995–2016." *Journal of the European Economic Association* 22: 1228–1274.

Adermon, A., M. Lindahl, and D. Waldenström. 2018. "Intergenerational Wealth Mobility and the Role of Inheritance: Evidence From Multiple Generations." *Economic Journal* 128, no. 612: F482–F513.

Albers, T. N., C. Bartels, and M. Schularick. 2022. *The Distribution of Wealth in Germany*, 1895–2018. CESifo Working Paper 9739. CESifo.

Alvaredo, F., A. B. Atkinson, T. Blanchet, et al. 2021. *Distributional National Accounts Guidelines Methods and Concepts Used in the World Inequality Database*. Technical report. HAL.

Athey, S., and G. W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11: 685–725.

Atkinson, A. B., T. Piketty, and E. Saez. 2011. "Top Incomes in the Long Run of History." *Journal of Economic Literature* 49, no. 1: 3–71.

Auray, S., A. Eyquem, B. Garbinti, and J. Goupille-Lebret. 2022. *Markups, Taxes, and Rising Inequality*. Technical report. CESifo.

Auten, G., and D. Splinter. 2024. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." *Journal of Political Economy* 132, no. 7: 2179–2227.

Bach, S. 2016. *Erbschaftsteuer, Vermögensteuer oder Kapitaleinkommensteuer: Wie sollen hohe Vermögen stärker besteuert werden?* Technical report. German Institute for Economic Research.

Bach, S., G. Corneo, and V. Steiner. 2009. "From Bottom to Top: The Entire Income Distribution in Germany, 1992–2003." *Review of Income and Wealth* 55, no. 2: 303–330.

Bach, S., A. Thiemann, and A. Zucco. 2019. "Looking for the Missing Rich: Tracing the Top Tail of the Wealth Distribution." *International Tax and Public Finance* 26, no. 6: 1234–1258. https://doi.org/10.1007/s10797-019-09578-1.

Bartels, C., and D. Waldenström. 2022. *Inequality and Top Incomes*. Handbook of Labor, Human Resources and Population Economics.

Black, S. E., P. J. Devereux, F. Landaud, and K. G. Salvanes. 2022. "The (un)Importance of Inheritance." *Journal of the European Economic Association*.

Black, S. E., P. J. Devereux, P. Lundborg, and K. Majlesi. 2020. "Poor Little Rich Kids? The Role of Nature Versus Nurture in Wealth and Other Economic Outcomes and Behaviours." *Review of Economic Studies* 87, no. 4: 1683–1725.

Boserup, S. H., W. Kopczuk, and C. T. Kreiner. 2016. "The Role of Bequests in Shaping Wealth Inequality: Evidence From Danish Wealth Records." *American Economic Review* 106, no. 5: 656–661.

Boserup, S. H., W. Kopczuk, and C. T. Kreiner. 2018. "Born With a Silver Spoon? Danish Evidence on Wealth Inequality in Childhood." *Economic Journal* 128, no. 612: F514–F544.

Bricker, J., L. J. Dettling, A. Henriques, et al. 2017. "Changes in Us Family Finances From 2013 to 2016: Evidence From the Survey of Consumer Finances." *Federal Reserve Bulletin* 103, no. 1: 1–42.

Bricker, J., S. Goodman, A. H. Volz, and K. B. Moore. 2020. *Wealth and Income Concentration in the Scf*, 1989–2019. Technical report. Board of Governors of the Federal Reserve System.

Bricker, J., A. Henriques, J. Krimmel, and J. Sabelhaus. 2016. "Measuring Income and Wealth at the Top Using Administrative and Survey Data." *Brookings Papers on Economic Activity* 2016, no. 1: 261–331.

Brunori, P., and G. Neidhöfer. 2021. "The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach." *Review of Income and Wealth* 67, no. 4: 900–927.

Bucks, B. K., A. B. Kennickell, T. L. Mach, and K. B. Moore. 2009. "Changes in Us Family Finances From 2004 to 2007: Evidence From the Survey of Consumer Finances." *Federal Reserve Bulletin* 95: A1-A56.

Cagetti, M., and M. De Nardi. 2006. "Entrepreneurship, Frictions, and Wealth." *Journal of Political Economy* 114, no. 5: 835–870.

Caliendo, M., F. Fossen, and A. Kritikos. 2010. "The Impact of Risk Attitudes on Entrepreneurial Survival." *Journal of Economic Behavior & Organization* 76, no. 1: 45–63.

Caliendo, M., F. M. Fossen, and A. S. Kritikos. 2009. "Risk Attitudes of Nascent Entrepreneurs–New Evidence From an Experimentally Validated Survey." *Small Business Economics* 32, no. 2: 153–167.

Christiansen, V., and M. Tuomala. 2008. "On Taxing Capital Income With Income Shifting." *International Tax and Public Finance* 15, no. 4: 527–545. https://doi.org/10.1007/s10797-008-9076-x.

DeNardi, M., and G. Fella. 2017. "Saving and Wealth Inequality." *Review of Economic Dynamics* 26: 280–300. https://doi.org/10.1016/j.red.2017.06.002.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9, no. 3: 522–550.

Drechsel-Grau, M., A. Peichl, K. D. Schmid, J. F. Schmieder, H. Walz, and S. Wolter. 2022. "Inequality and Income Dynamics in Germany." *Quantitative Economics* 13, no. 4: 1593–1635.

Fagereng, A., M. Mogstad, and M. Rønning. 2021. "Why Do Wealthy Parents Have Wealthy Children?" *Journal of Political Economy* 129, no. 3: 703–756.

Fisher, J. D., D. S. Johnson, T. M. Smeeding, and J. P. Thompson. 2022. "Inequality in 3-d: Income, Consumption, and Wealth." *Review of Income and Wealth* 68, no. 1: 16–42.

Fossen, F. M., J. König, and C. Schröder. 2024. "Risk Preference and Entrepreneurial Investment at the top of the Wealth Distribution." *Empirical Economics* 66, no. 2: 735–761.

Garbinti, B., J. Goupille-Lebret, and T. Piketty. 2021. "Accounting for Wealth-Inequality Dynamics: Methods, Estimates, and Simulations for France." *Journal of the European Economic Association* 19, no. 1: 620–663.

Grossmann, V., and H. Strulik. 2010. "Should Continued Family Firms Face Lower Taxes Than Other Estates?" *Journal of Public Economics* 94, no. 1: 87–101. https://doi.org/10.1016/j.jpubeco.2009.10.005.

Hankins, S., M. Hoekstra, and P. M. Skiba. 2011. "The Ticket to Easy Street? The Financial Consequences of Winning the Lottery." *Review of Economics and Statistics* 93: 961–969.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer.

Jordà, Ò., K. Knoll, D. Kuvshinov, M. Schularick, and A. M. Taylor. 2019. "The Rate of Return on Everything, 1870–2015." *Quarterly Journal of Economics* 134, no. 3: 1225–1298.

Kennickell, A. 2019. "The Tail That Wags: Differences in Effective Right Tail Coverage and Estimates of Wealth Inequality." *Journal of Economic Inequality* 17: 443–459. https://doi.org/10.1007/s10888-019-09424-8.

Kholodilin, K. A., C. Michelsen, and D. Ulbricht. 2018. "Speculative Price Bubbles in Urban Housing Markets." *Empirical Economics* 55, no. 4: 1957–1983.

König, J., C. Schröder, and E. N. Wolff. 2020. *Wealth Inequalities*. Handbook of Labor, Human Resources and Population.

Kopczuk, W. 2015. "What Do We Know About the Evolution of Top Wealth Shares in the United States?" *Journal of Economic Perspectives* 29, no. 1: 47–66.

Kopczuk, W., and E. Zwick. 2020. "Business Incomes at the Top." *Journal of Economic Perspectives* 34, no. 4: 27–51.

Kuhn, M., M. Schularick, and U. I. Steins. 2020. "Income and Wealth Inequality in America, 1949–2016." *Journal of Political Economy* 128, no. 9: 3469–3519.

Leckelt, M., J. König, D. Richter, M. D. Back, and C. Schröder. 2022. "The Personality Traits of Self-Made and Inherited Millionaires." *Humanities and Social Sciences Communications* 9, no. 1: 1–12. https://doi.org/10.1057/s41599-022-01099-3.

Levine, R., and Y. Rubinstein. 2017. "Smart and Illicit: Who Becomes an Entrepreneur and Do They Earn More?" *Quarterly Journal of Economics* 132, no. 2: 963–1018.

Martinez, I. 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, Incorporation.

Martínez-Toledano, C. 2020. *House Price Cycles, Wealth Inequality and Portfolio Reshuffling*. WID World Working Paper, 2.

McCrae, R. R., and P. T. Costa Jr. 1997. "Personality Trait Structure as a Human Universal." *American Psychologist* 52, no. 5: 509–516. https://doi.org/10.1037//0003-066x.52.5.509.

Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31, no. 2: 87–106.

Nekoei, A., and D. Seim. 2023. "How Do Inheritances Shape Wealth Inequality? Theory and Evidence From Sweden." *Review of Economic Studies* 90, no. 1: 463–498.

Nelsen, R. 2006. *An Introduction to Copulas*. Springer.

Ozkan, S., J. Hubmer, S. Salgado, and E. Halvorsen. 2023. Why are the Wealthiest so Wealthy? A Longitudinal Empirical Investigation.

Piketty, T., G. Postel-Vinay, and J.-L. Rosenthal. 2014a. "Inherited vs Self-Made Wealth: Theory & Evidence From a Rentier Society (Paris 1872–1927)." *Explorations in Economic History* 51: 21–40.

Piketty, T., and E. Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118, no. 1: 1–41.

Piketty, T., E. Saez, and S. Stantcheva. 2014b. "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities." *American Economic Journal: Economic Policy* 6, no. 1: 230–271. https://doi.org/10.1257/pol.6.1.230.

Piketty, T., E. Saez, and G. Zucman. 2018. "Distributional National Accounts: Methods and Estimates for the United States." *Quarterly Journal of Economics* 133, no. 2: 553–609.

Raub, B., B. Johnson, and J. 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*, vol. 103, 128–135. JSTOR.

Saez, E. 2002. "The Desirability of Commodity Taxation Under Non-linear Income Taxation and Heterogeneous Tastes." *Journal of Public Economics* 83, no. 2: 217–230.

Saez, E., and G. Zucman. 2016. "Wealth Inequality in the United States Since 1913: Evidence From Capitalized Income Tax Data." *Quarterly Journal of Economics* 131, no. 2: 519–578.

Saez, E., and G. Zucman. 2019. "Progressive Wealth Taxation." *Brookings Papers on Economic Activity* 2019, no. 2: 437–533.

Salas-Rojo, P., and J. G. Rodríguez. 2022. "Inheritances and Wealth Inequality: A Machine Learning Approach." *Journal of Economic Inequality* 20, no. 1: 27–51.

Schluter, C. 2018. "Top Incomes, Heavy Tails, and Rank-Size Regressions." *Econometrics* 6, no. 1: 10.

Schluter, C. 2020. "On Zipf's Law and the Bias of Zipf Regressions." *Empirical Economics* 61: 1–20.

Schröder, C., C. Bartels, M. M. Grabka, J. König, M. Kroh, and R. 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, no. 4: 825–849. https://doi.org/10.1111/roiw.12452.

Siegers, R., H. W. Steinhauer, and J. König. 2021. SOEP-Core-2019: Sampling, Nonresponse, and Weighting in Sample P Technical report, SOEP Survey Papers.

Smith, M., D. Yagan, O. Zidar, and E. Zwick. 2019. "Capitalists in the Twenty-First Century." *Quarterly Journal of Economics* 134, no. 4: 1675–1745.

Smith, M., O. Zidar, and E. Zwick. 2023. "Top Wealth in America: New Estimates Under Heterogeneous Returns." *Quarterly Journal of Economics* 138, no. 1: 515–573. https://doi.org/10.1093/qje/qjac033.

SOEP Group. 2021. *SOEP-Core v36 – Pgen: Person-Related Status and Generated Variables* Technical report, SOEP Survey Papers. DIW-SOEP.

Stanley, T. J. 1996. *The Millionaire Next Door: The Surprising Secrets of America's Wealthy*. Taylor Trade Publishing.

Vermeulen, P. 2016. "Estimating the Top Tail of the Wealth Distribution." *American Economic Review* 106, no. 5: 646–650.

Vermeulen, P. 2018. "How Fat Is the Top Tail of the Wealth Distribution?" *Review of Income and Wealth* 64, no. 2: 357–387.

Wolff, E. 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.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Online Appendix.** Supporting Information.