Beyond Average Social Mobility: The Distribution of Economic Opportunity

Moritz Hörl*

This version: October 2026

Abstract

Studies of heterogeneity in social mobility typically compare a few predefined groups. This paper instead develops the Conditional Mobility Forest, a nonparametric estimator of each child's expected income rank at the 25th parental percentile. In the PSID, this measure of upward mobility varies by about 20 percentile ranks across children. A Shapley-value decomposition identifies race as the most important characteristic, followed by childhood health and parental age at birth. Parental education, homeownership, and household structure, despite strong raw correlations, contribute comparatively little once correlated characteristics are accounted for. These results can sharpen the targeting of social mobility policies.

Keywords: intergenerational income mobility; machine learning; SHAP

JEL codes: J62; D63; C14

*Vienna University of Economics and Business (WU). Email: moritz.hoerl@wu.ac.at. I am grateful to Franziska Disslbacher, Anna Schwarz, Per Engzell, Jan Stuhler, Xavier Ramos, Astrid Pennerstorfer, Thomas Rusch, Paul Hufe, Alejandro Puertas and Joël Terschuur for their valuable comments and suggestions. I also thank participants at the internal INEQ research seminar, the Winter School on Inequality and Social Welfare Theory in Alba di Canazei, and the Internal Economic Research Seminar Series of the Department of Economics at WU Vienna for insightful feedback. All remaining errors are my own.

1 Introduction

The ability of children from low-income families to move up the income distribution as adults is central to the ideal of equality of opportunity. Yet upward mobility from the bottom varies substantially across places, social groups, and birth cohorts (Chetty et al. 2014b; Chetty et al. 2020; Jácome et al. 2025). In the United States, for example, Davis and Mazumder (2026) show that African-American children are markedly less likely than White children to escape low-income origins.

Who faces limited prospects for escaping the bottom of the income distribution is not only a question of fairness but one with tangible political and social consequences. Persistent low upward mobility implies that disadvantage is transmitted across generations, increasing the likelihood that specific groups remain economically and socially excluded over long periods. Such persistent disadvantage has been linked to higher incarceration rates (Chetty et al. 2020), and to greater support for populist or extremist political movements (Kurer and Van Staalduinen 2022). Despite these stakes, most empirical work characterizes upward mobility either through national summary statistics or through estimates computed for a handful of broad subgroups. Such measures are informative, but they average over much of the heterogeneity that determines whose opportunities are constrained. A society that looks mobile on average, or mobile for certain large subgroups, may nonetheless contain sharply unequal upward mobility patterns across finer segments of the population.

This paper develops a new framework to uncover and visualize these hidden disparities. I introduce the Conditional Mobility Forest (CMF), a nonparametric estimator designed to capture heterogeneity in upward mobility across many dimensions of family background. The estimator focuses on heterogeneity in the Conditional Expected Rank at the 25th parental percentile (CER_{25})—a widely used measure of upward mobility from the bottom of the income distribution (Deutscher and Mazumder 2023). Building on the generalized random forest framework of Athey et al. (2019), the CMF is tailored to detect variation in expected child outcomes across a rich set of parental, household, and contextual characteristics.[1] This approach addresses fundamental challenges that have constrained empirical work on measuring heterogeneity in intergenerational mobility. Conventional approaches rely either on ex-ante subgroup definitions based on observed characteristics or on regression specifications that introduce prespecified interaction terms to capture differences in upward mobility patterns across groups. Both strategies require researchers to determine in advance which dimensions and combinations thereof are most relevant for characterizing heterogeneity in upward mobility. However, the determinants of upward mobility are

[1] While social mobility can be assessed along multiple dimensions—such as education, occupation, or wealth—this paper focuses on intergenerational persistence in income, reflecting the dimension most extensively examined in the empirical literature (Deutscher and Mazumder 2023).

potentially numerous, interconnected, and not necessarily known a priori, meaning that prespecified subgroup definitions may overlook combinations of characteristics associated with large differences in upward mobility. Additionally, increasingly fine partitions of the population generate smaller cells, reducing statistical precision and increasing the risk of unstable estimates (Chetty et al. 2025). Interaction-based approaches face an additional challenge: they require researchers to impose a parametric structure on how characteristics jointly shape upward mobility, which may be poorly suited to complex nonlinear and high-dimensional patterns of heterogeneity and can lead to overfitting as the number of interactions expands (Chang et al. 2025).

The Conditional Mobility Forest (CMF) addresses these limitations by using data-adaptive recursive partitioning to uncover, in a fully nonparametric manner, which combinations of characteristics generate variation in upward mobility. The method requires no ex-ante subgroup specification and imposes no functional-form assumptions on how characteristics interact. Instead, it produces a high-resolution map of heterogeneity in upward mobility, revealing the dimensions along which upward mobility differs most. Unlike machine learning methods developed for estimating heterogeneous causal treatment effects, the CMF is explicitly designed to exploit the correlational structure of the data. This is appropriate as upward mobility is a descriptive quantity rather than a causal parameter. In this respect, the CMF complements and extends the empirical literature documenting heterogeneity in upward mobility across population subgroups (Chang et al. 2025; Chetty et al. 2020; Jácome et al. 2025).

Empirically, I apply the Conditional Mobility Forest to longitudinal data from the Panel Study of Income Dynamics (PSID) to estimate individualized conditional expected ranks at the 25th percentile of the parental income distribution. The resulting estimates reveal substantial heterogeneity in upward mobility and identify the background characteristics most strongly associated with these differences. I assess the validity of the estimates in several ways. First, I demonstrate that the results are stable across sample splits and robust to alternative hyperparameter choices. Second, I show that the CER_{25} estimates are well calibrated by comparing estimated upward mobility with the realized outcomes of individuals whose parents are located at the 25th percentile of the parental income distribution. As an initial descriptive exercise, I examine the relationship between estimated CER_{25} values and observable characteristics included in the CMF. Race emerges as the strongest correlate of estimated upward mobility by a substantial margin. Identifying as White is associated with considerably higher upward mobility, whereas identifying as African American is associated with markedly lower upward mobility. Beyond race, growing up in a two-parent household, having parents who own their home, higher parental education, parental employment in managerial or professional occupations, older parental age at birth, and more frequent shared meals are all positively associated with upward mobility. Conversely, growing up

with many siblings or in the U.S. South is associated with lower prospects for upward mobility.

To move beyond univariate correlations, I complement the analysis with SHapley Additive exPlanations (SHAP). SHAP decomposes each individual's upward mobility estimate, relative to the sample mean, into additive contributions from the underlying background variables. Unlike correlational analysis, which assesses each characteristic in isolation, this yields a measure of variable importance in which all included characteristics are considered jointly. Because Shapley values reflect a variable's average contribution given the other variables, a characteristic that adds little predictive value beyond correlated background variables receives a correspondingly small contribution. By this metric, race is again by far the most important dimension of heterogeneity in upward mobility, followed by childhood health, parental age at birth, birth year, and family meal frequency. In contrast, household structure and apartment ownership receive little weight once the other correlated childhood characteristics are accounted for, indicating that their univariate associations largely reflect shared variation with other variables. These findings are consistent with previous research. Racial disparities in upward mobility persist conditional on parental income and neighbourhood characteristics (Davis and Mazumder 2018; Chetty et al. 2020; Jácome et al. 2025). Early-life health is an important channel through which disadvantage is transmitted across generations (Case and Paxson 2006; Currie 2009), while family structure explains relatively little once broader socioeconomic circumstances are accounted for (Bloome 2017). Whereas these studies typically examine individual dimensions separately or within a limited set of prespecified subgroups, the SHAP decomposition evaluates their relative contributions jointly within a single nonparametric framework. To my knowledge, this yields the first joint ranking of the importance of background characteristics for upward mobility that holds parental income fixed and lets the data determine how characteristics interact. Because the ranking is necessarily restricted to characteristics observed in the PSID, it should be read as an approximation. It describes relative importance among observed variables, some of which may partly proxy for unobserved factors such as neighbourhood quality or discrimination. Even so, it gives policymakers a systematic basis for identifying which children face the most constrained mobility prospects and should therefore be prioritized in targeted interventions.

By integrating recent advances in machine learning with the empirical literature on heterogeneity in social mobility, this paper contributes to two distinct strands of research. First, it contributes to the literature on measuring social mobility and documenting heterogeneity across population subgroups (Chetty et al. 2014b; Chetty et al. 2020; Nybom and Stuhler 2025). Rather than relying on prespecified subgroup classifications or restrictive parametric specifications, the proposed approach provides a flexible, data-driven framework for uncovering heterogeneity in upward mobility across a high-dimensional covariate space. Second, the paper contributes to the methodological literature on

estimating heterogeneity in social mobility. Measuring how upward mobility varies across the population is difficult precisely because the set of potentially relevant background characteristics is large. The paper adapts the core ideas of machine learning methods for heterogeneous effect estimation (Athey et al. 2019; Athey and Wager 2019; Chernozhukov et al. 2025) to the analysis of upward mobility. Although CER_{25} is a descriptive measure rather than a causal parameter, the underlying principles of adaptive, nonparametric heterogeneity analysis carry over directly. The closest related contribution is Chang et al. (2025), who develop a flexible nonparametric framework for studying heterogeneity in social mobility without imposing strong functional-form assumptions. Their approach, however, necessarily considers only a limited number of covariates simultaneously and therefore retains the central limitation of conventional subgroup analysis, since researchers must specify ex-ante the dimensions along which social mobility is allowed to vary. By contrast, the Conditional Mobility Forest searches over a rich set of covariates and their interactions without requiring prior specification of the relevant subgroup structure. Related work in machine learning has employed flexible prediction algorithms to forecast children's income ranks from a broad set of background characteristics, including parental income (Blundell and Risa 2018). These studies are concerned with predictive performance for children's outcomes rather than with characterizing heterogeneity in social mobility itself. The objective of the present paper is instead to estimate individualized measures of expected upward mobility and to identify the variables associated with systematic differences in those upward mobility prospects.

The remainder of the paper is organized as follows. Section 2 develops the conceptual framework, motivates the focus on CER_{25}, and discusses the limitations of conventional approaches to studying heterogeneity in social mobility. Section 3 introduces the Conditional Mobility Forest estimator and situates it within the broader methodological literature. Section 4 describes the PSID data and the construction of the empirical variables. Section 5 presents the empirical results, including the validation exercises and the comparison with conventional subgroup analysis. Section 6 concludes by discussing the implications of the findings and directions for future research.

2 Theoretical Framework

I extend the classic social mobility framework, following Chetty et al. (2020) and Becker and Tomes (1979), to allow for individual-level heterogeneity in upward mobility, measured as the expected income rank of a child whose parents are at the 25th percentile of the national income distribution. The model assumes that income is rank-normalized within cohorts.[2]

[2]Social mobility can be estimated using various methods. The rank-rank specification is particularly useful for analyzing heterogeneity in social mobility rates. Unlike other approaches such as the intergenerational elasticity (IGE), the rank-rank approach allows for identifying variation in social mobility conditional on an individual's position in the

2.1 Model Setup

I consider the standard rank-rank specification commonly used to estimate income mobility. Let $y_c \in [0, 100]$ denote the national income percentile rank of child c, and let $y_p \in [0, 100]$ represent the national percentile rank of parental income. This leads to the following baseline model:

$$y_c = \alpha + \beta y_p + \varepsilon \tag{1}$$

Here,

- α captures **absolute mobility**—the expected income rank of child c if the parent were at the bottom of the income distribution. According to Chetty et al. (2020), differences in α reflect disparities in economic outcomes conditional on parental income rank. Corak (2020) interprets this parameter as a measure of absolute mobility devoid of family background.
- β captures **relative mobility**—how strongly the child's income rank is associated with that of the parent. A higher value of β indicates stronger persistence (i.e., less mobility), while lower values imply greater social mobility.

My primary object of interest is the conditional expected rank of children born to families at the 25th rank in the income distribution (CER_{25}). Formally, I define:

$$\mu_c = \mathbb{E}[y_c \mid y_p = 25, X_c],$$

where X_c denotes a vector of observable characteristics (e.g., gender, race, parental education, region). The quantity μ_c thus captures the expected economic position of child c, conditional on coming from a low-income background (parents at the 25th percentile). In the baseline model, CER_{25} corresponds to the predicted value of child rank when parental rank is set to 25. Alternatively, CER_{25} can be estimated nonparametrically by binning the data into parental income centiles and computing conditional means (Deutscher and Mazumder 2023).

The Conditional Expected Rank at the 25th percentile of the parental income distribution (CER_{25}) has emerged as a particularly salient metric in the study of social mobility, frequently interpreted as a rough approximation of the poverty threshold (Deutscher and Mazumder 2023). As such, CER_{25} serves as an intuitive and policy-relevant summary measure of mobility potential for economically disadvantaged groups. Importantly, CER_{25} reflects a combination of absolute and relative mobility components: it incorporates the intercept of the rank–rank relationship (α), which captures positional outcomes net of family background (Corak 2020), as well as the slope (β), which

national income distribution (Mazumder 2016).

measures the degree of intergenerational persistence. This is an important advantage as traditionally, studies of heterogeneity in social mobility have focused on differences in the slope coefficient, β, across subgroups (Chang et al. 2025). However, as emphasized by Cholli and Durlauf (2022), β only captures the degree of intergenerational persistence—it reflects how closely children's outcomes are linked to those of their parents—but offers no information about where children ultimately end up in the income distribution. Consequently, upward and downward mobility are treated symmetrically, which is not justified from a normative perspective. For example, it may be that a subgroup has a low β coefficient, indicating high social mobility, even if most of its members end up in the lower end of the income distribution. Conversely, affluent families may pass much of their advantage, reflected in high β coefficients and low social mobility rates for this subgroup. In both cases, the slope coefficient fails to convey whether a group faces upward mobility constraints or benefits from strong persistence. Thus, reliance on β alone offers limited insight into which subgroups suffer from constrained economic opportunities.

Alternatively, Cholli and Durlauf (2022) propose to examine heterogeneity in the intercept α, which captures subgroup-level differences in absolute mobility. In the rank-rank specification (1), the intercept represents the expected income rank of a child whose parents are at the bottom of the income distribution. Consequently, it is a measure of positional advantage after removing all factors associated with parental income. Variation in α across subgroups reflects differences in outcomes that cannot be attributed to parental income or, more broadly, to family background—since parental rank is highly correlated with other family-level characteristics. This distinction has important implications. For instance, a difference in α between African-American and White individuals suggests that the source of the disparity lies outside the family's socioeconomic status, potentially reflecting structural factors such as racial discrimination. Yet, despite its conceptual relevance, the intercept has received limited attention in the applied literature. Davis and Mazumder (2018) show that while β does not substantially differ by race, the α values between African-American and White people differ significantly. However, as emphasized by Chetty et al. (2020), it is the joint combination of α and β —rather than absolute mobility alone—that determines where individuals end up in the income distribution. It is therefore important to analyze heterogeneity in the joint distribution of both parameters to identify barriers to upward mobility and to inform the design of targeted policies aimed at promoting equality of opportunity. Motivated by this insight, the present paper focuses on heterogeneity in the CER_{25}, a measure that integrates both the intercept and slope of the rank–rank relationship.

2.2 Limitations of Traditional Approaches for Analyzing Heterogeneity in Social Mobility

Traditionally, researchers have examined heterogeneity in social mobility indicators like the CER_{25} by estimating the rank-rank relationship within subsamples defined by characteristics such as race or parental education, or by extending the baseline specification as follows:

$$y_c = \alpha + \beta y_p + \gamma D + \delta(y_p \times D) + \varepsilon \tag{2}$$

In this specification, D denotes an indicator for subgroup membership (e.g., gender, race, or region), and the interaction term $y_p \times D$ allows social mobility to vary across these subgroups. However, the specification in (2) has become increasingly criticized in the recent literature. As argued by Chang et al. (2025), this specification is very restrictive because it does not allow for interactions between the included observable characteristics in determining a child's income level. However, including interaction terms in the rank-rank specification effectively approximates the interaction term through a second-order Taylor series, which is neither justified by economic theory, nor a particularly accurate description of the real-world processes underlying the intergenerational transmission of income. For instance, threshold effects may be present: while barriers to upward mobility may be non-existent for African-American children from low-income families, children from high-income African-American families could face significantly greater discrimination as they move further up the social ladder. Consequently, the specification in equation (2) is incapable of capturing nonlinear or higher-order relationships between parental and child income, which may be important to approximate the true nature of intergenerational processes.

A further limitation is that researchers must specify in advance the characteristics which they want to include in specification (2) or according to which they want to form subgroups to examine heterogeneity in social mobility. However, it cannot be known a priori which variables are most important for characterizing heterogeneity in social mobility. Consequently, empirical analyses risk overlooking important bottlenecks. Characteristics such as race, parental education, and wealth may interact in complex, nonlinear ways, creating bottlenecks impossible to analyze with traditional methods. Systematically searching for such higher-order interactions with parametric models is infeasible as including a large number of interaction terms in specification (2) would lead to overfitting. Similarly, splitting the sample into multiple subgroups is not advisable, as the relevant sources of heterogeneity in social mobility—and their interactions—cannot be known ex-ante. In addition, dividing the data into smaller subsamples increases the variance of the estimated mobility values, making subgroup comparisons less reliable. As argued by Chetty et al. (2025), estimating the CER_{25} non-parametrically when taking various characteristics into account is often impossible

as sample size decreases in smaller subgroups, increasing estimation uncertainty.

To address these limitations, this paper introduces the Conditional Mobility Forest—a fully nonparametric, data-driven estimator of heterogeneity in upward mobility, measured as the CER_{25}. The method allows for the systematic identification of heterogeneity across a rich set of individual and family characteristics, accommodating complex, potentially nonlinear interactions without requiring prior specification of functional forms or subgroup definitions. Unlike conventional parametric models, which impose restrictive assumptions and rapidly become infeasible in high-dimensional settings, the CMF leverages recursive partitioning to flexibly uncover sources of heterogeneity in upward mobility. This approach enables the detection of latent bottlenecks in economic opportunity that may be obscured by model misspecification or omitted interaction terms, allowing the data—rather than researcher decisions—to guide the discovery of upward mobility heterogeneity.

3 Empirical Strategy and Estimation Method

3.1 Conceptual Framework

My objective is to estimate heterogeneous patterns of upward mobility, defined as the expected child income rank at the 25th parental income rank. Let $Y_p \in [0, 100]$ denote the parental income rank, and $Y_c \in [0, 100]$ the corresponding child income rank. Let X_c represent a vector of observed characteristics, defined in Section 4, that may explain heterogeneity in upward mobility. The quantity of interest is then the conditional expectation:

$$\mu_c = \mathbb{E}[y_c \mid y_p = 25, X_c],$$

which captures the expected economic outcome of child c if their parent were at the 25th percentile of the income distribution. The goal is to estimate heterogeneity in μ_c flexibly across the covariate space X_c, without assuming a specific functional form or prespecifying subgroup comparisons. The underlying idea is simple: the Conditional Mobility Forest (CMF) uses the data to identify which children are most comparable to one another in terms of their upward mobility prospects. Rather than imposing a functional form, the algorithm recursively partitions the data along background characteristics to create groups within which the relationship between parental and child income is relatively homogeneous.

In this way, the CMF sequentially divides the sample into subgroups whose estimated CER_{25} values differ as sharply as possible. At each splitting point, the algorithm evaluates candidate partitions

across different variables and threshold values, computing the corresponding CER_{25} estimates for each potential split. It then selects the partition that maximizes the difference in CER_{25} between the resulting subgroups. This procedure is repeated recursively until subsample sizes fall below a specified threshold or a predefined maximum splitting depth is reached. This contrasts with standard random forests, which choose splits solely to improve prediction of an outcome variable. The CMF instead chooses splits that maximize differences in CER_{25} across groups, enabling it to uncover fine-grained heterogeneity in upward mobility. The resulting subgroups across individual trees are then used to construct a similarity-based weighting matrix for each child. Using these weights, the CMF estimates a personalized rank–rank regression for every individual, yielding an individualized estimate of CER_{25} accounting for the background characteristics of an individual.

3.2 Estimation with a Conditional Mobility Forest

The Conditional Mobility Forest is inspired by the Generalized Random Forest (GRF) framework of Athey et al. (2019), but is tailored specifically to the goal of estimating heterogeneity in upward mobility. Whereas GRFs are typically used to study variation in slopes or treatment effects, my method focuses instead on heterogeneity in the expected child rank at the fixed parental income rank of 25.

The algorithm builds a forest of honest regression trees that recursively partition the covariate space to uncover heterogeneity in CER_{25}.[3] The core components of the estimation procedure are summarized below, with a full technical description in Appendix A.

1. **Honest tree construction:** Each tree is built on a 70% subsample of the data (without replacement), which is further split into a *splitting sample* and an *estimation sample*. Splits are chosen only on the splitting sample, while the estimation sample is used exclusively to assign observations to leaves and define forest weights for each out-of-bag unit. This honesty design prevents overfitting and ensures unbiased neighbourhood construction.

2. **Recursive partitioning:** Trees are grown by recursively splitting on covariates X_c to maximize heterogeneity in the estimated child rank at $Y_p = 25$. Each split is evaluated using a penalized score function based on the squared difference in estimated mobility outcomes at the target point, adjusted for sample size and balance. To avoid extrapolation and ensure meaningful local estimation, I impose a common support condition: a split is only accepted if $Y_p = 25$ lies within the 5th–95th percentile range of both resulting child nodes.

[3] An honest tree uses separate subsamples for determining splits and for determining weights in the terminal nodes. This prevents overfitting and reduces bias by ensuring that the data used to select splits is not reused for outcome estimation. For more details see Athey et al. (2019).

3. **Sparse out-of-bag weighting:** For each child c, I construct a sparse, individualized kernel using only trees where c was completely out-of-bag. In each such tree, I identify estimation-sample units that share a terminal leaf with c and assign them equal kernel weights within that leaf. This kernel captures which observations are comparable to c. The individual survey weights enter at the local-regression stage (step 4), so the equal weighting here is over honest neighbors. These honest neighbourhoods are aggregated across trees and normalized by the number of trees in which c was out-of-bag and had valid neighbors. This strict honesty principle ensures that estimation weights are never influenced by the observation itself.

4. **Local intercept estimation:** For each child c, I estimate CER_{25} by running a weighted local regression of Y_c on Y_p, where each neighbor's weight is the product of the forest kernel and its PSID longitudinal survey weight. The regression is centered at $Y_p = 25$. Survey weighting also enters split selection (Appendix A), so the estimates are representative of the national population rather than of the PSID sample.

This procedure delivers fully nonparametric, individual estimates of the conditional expected rank at the 25th parental percentile (CER_{25}), without imposing restrictive functional-form assumptions or relying on predefined subgroup classifications.

The Conditional Mobility Forest (CMF) builds on core principles from the Generalized Random Forest (GRF) framework, including honest sample splitting, recursive partitioning, and adaptive kernel weighting (Athey et al. 2019). However, the algorithm is tailored specifically to the goal of discovering heterogeneity in upward mobility. This focus motivates several key departures from the standard GRF approach. First, while GRFs typically target slope heterogeneity—such as causal treatment effects—the CMF is designed to uncover variation in the estimated CER_{25} of a local rank–rank regression. Second, I evaluate candidate splits based on the squared difference in estimated upward mobility at the target point $y_p = 25$, not via pseudo-outcomes or gradients. This ensures that the tree-growing process is directly aligned with the parameter of interest. Third, to prevent extrapolation and enhance interpretability, I impose a common support condition: candidate splits are only accepted if $y_p = 25$ lies within a central range of the parent rank distribution in both resulting child nodes. This helps to avoid misleading comparisons driven by very rich parents and ensures that estimates at the 25th percentile reflect actual patterns in the data, rather than being distorted by extreme cases. Finally, while GRFs support formal pointwise inference through a delta-method variance estimator (Athey and Wager 2019), I do not report confidence intervals for individual CER_{25} estimates. Because the final step of the CMF solves a local estimating equation of the same form as in GRF, pointwise inference could in principle be obtained along these lines. The baseline configuration, however, does not satisfy all of these conditions. It imposes no minimum share of observations per child node and grows trees on 70 percent subsamples, whereas the GRF

variance estimator requires trees grown in groups on half-samples. In addition, the common-support restriction may limit how far leaves shrink, and the use of survey weights falls outside the standard theory. Establishing valid pointwise inference under these modifications is beyond the scope of this paper. Instead, I assess the reliability of the estimates through three empirical validation checks, described in detail in Section 5.2 and Online Appendix C.

4 Data

I draw on data from the Panel Study of Income Dynamics (PSID), a longitudinal survey of U.S. families that began in 1968 and remains one of the most comprehensive sources for studying social mobility. The PSID follows individuals and their families over time, thus providing an enormously rich dataset to study upward mobility patterns. While administrative data often constrain researchers to focus on structural variables— such as education or family structure— the PSID enables me to incorporate a broader set of behavioral variables, including family dynamics and health conditions. This allows for a more detailed analysis of heterogeneity in upward mobility. Online Appendix B.1 provides a detailed description of the sample construction and variable definitions, following the approach of Kenedi and Sirugue (2023). Here, I briefly summarize the key sampling restrictions and variable definitions.

4.1 Sample Definitions

The analysis focuses on individuals born between 1963 and 1988 who can be reliably observed in both childhood—through linkage to parental records—and adulthood. To be included in the sample, individuals must be observed at least once as a child within a family unit and subsequently as either a reference person or spouse/partner between ages 30 and 50.[4] Following Chetty et al. (2014b), children are assigned to the family unit in which they are first observed, and this assignment remains fixed regardless of any subsequent changes in household composition. I further require that all individuals and at least one parent report income at least once between the ages of 30 and 50. The resulting sample includes 5,077 individuals. Online Appendix B.1 provides a detailed account of the sample construction process. To maximize sample size and enhance statistical power, I utilize the full PSID dataset and apply longitudinal individual weights to preserve representativeness.

[4]The last survey wave from the PSID used in this analysis is from 2019.

4.2 Variable Definitions

In this subsection, I briefly define the variables used in the primary analysis. Details of all variables and their operationalization are provided in Online Appendix B.2. All monetary variables are measured in 2019 dollars. To address missing data in variables, I implement machine learning–based imputation procedures where appropriate, ensuring robust and nonparametric recovery of underlying relationships (see Online Appendix B.3 for details). The resulting dataset includes a rich set of variables beyond household income, enabling detailed exploration of heterogeneity in upward mobility. Further details on the variable selection process are provided in Online Appendix B.2.

1. *Parental Income Rank:* Average family household income (ages 30–50), adjusted for inflation and ranked within national income distributions. Household income rather than individual income is used as this provides a better picture of the economic situation of an individual.
2. *Child Income Rank:* Defined analogously to parental income rank.
3. *Gender:* Binary indicator for individual gender.
4. *Parental Education:* Highest level of schooling among parents, grouped as low, middle, or high.
5. *Parental Occupation:* Occupational group of the father (or mother if father missing), categorized using standard Census-based groupings.
6. *Child Race:* Most recently reported race category of an individual.
7. *Childhood Region:* Most common reported region of residence between ages 0 and 18.
8. *Shared Family Meals:* Average weekly frequency of shared main meals (parent-reported).
9. *Apartment Ownership:* Binary indicator for whether the household owned their home between ages 30–50.
10. *Parental Work-Limiting Condition:* Binary indicator for any long-term condition limiting the parent's ability to work.
11. *Child Health Status (Pre-18):* Average reported health status below age 18.
12. *Parental Age at Birth:* Average age of parents at the child's birth.
13. *Number of Siblings:* Total number of siblings observed in the family unit.

14. *Parental Household Structure:* Number of parents present in the household.

15. *Birth Cohort:* Year in which an individual was born.

4.3 Summary Statistics

Online Appendix Tables B.3 and B.4 report summary statistics for children and parents. Average household income for children, measured as the mean between ages 30 and 50 to mitigate life-cycle bias, is $81,888. For parents, the corresponding figure is $79,396. Parental income is constructed using an average of 14.5 maternal and 11.5 paternal observations, while child income is based on an average of 5.1 observations. Despite the smaller number, child income is still averaged over five years, substantially reducing the risk of measurement error. Most households consist of two parents, with an average of 1.79 parents present, and the mean number of siblings is 1.31. The frequency of shared family meals ranges from 0 to 7 per week, with a sample average of 3.82. The average parental age at birth is 26.57, with values ranging from 13 to 59 years. Educational attainment has increased across generations, with children attaining higher levels of education than their parents. The majority of parents report no work-limiting health conditions. Parental occupations are diverse, and nearly 87% of parents report having owned an apartment at some point between ages 30 and 50.

5 Empirical Results

This section examines heterogeneity in upward mobility in three steps. I first document the distribution of individual-level CER_{25} estimates produced by the Conditional Mobility Forest and provide evidence from a range of validation exercises that these estimates capture systematic heterogeneity rather than statistical noise. After these validation steps I turn to interpreting the results, relating the estimates to background characteristics through univariate associations and SHAP analysis.

5.1 Individual Level Distribution

I start by documenting the empirical distribution of upward mobility as captured by the individual-level CER_{25} estimates. Figure 1 displays the distribution of CER_{25} values obtained from the Conditional Mobility Forest. Each bar represents the number of individuals whose estimated CER_{25} falls within a given interval. These estimates should not be interpreted as the realized upward mobility of any specific individual—whose actual parental and child income ranks may deviate substantially from the reference point—but rather as the expected level of the CER_{25} for an individual with similar background characteristics whose parents occupy the 25th percentile of the

income distribution. Formally, each estimated CER_{25} value represents the outcome of an individualized, weighted rank–rank regression, where the weights are learned by the CMF algorithm. In this sense, the individual CER_{25} estimates summarize the expected economic prospects of individuals who share comparable circumstances, capturing how their background characteristics relate to opportunities for upward mobility, conditional on having parents at the 25th percentile of the income distribution.

Figure 1 reveals substantial dispersion in individual-level upward mobility estimates, which range from approximately rank 29 to rank 50—a spread of about 20 percentile ranks. To assess whether this dispersion reflects meaningful heterogeneity rather than estimation noise, I follow the validation strategy of Athey and Wager (2019). I split the sample at the median of the estimated upward mobility values $\hat{\alpha}(x_c)$ and, within each half, re-estimate the CER_{25} using a weighted regression of child income rank on parental income rank. In the high-mobility group, the estimated CER_{25} is $\hat{\alpha} = 45.376$ (95% CI: [43.600, 47.152]), compared with $\hat{\alpha} = 37.085$ (95% CI: [35.913, 38.257]) in the low-mobility group. The difference of 8.291 rank points (95% CI: [6.164, 10.419]) is statistically significant and provides evidence that the CMF captures systematic heterogeneity.

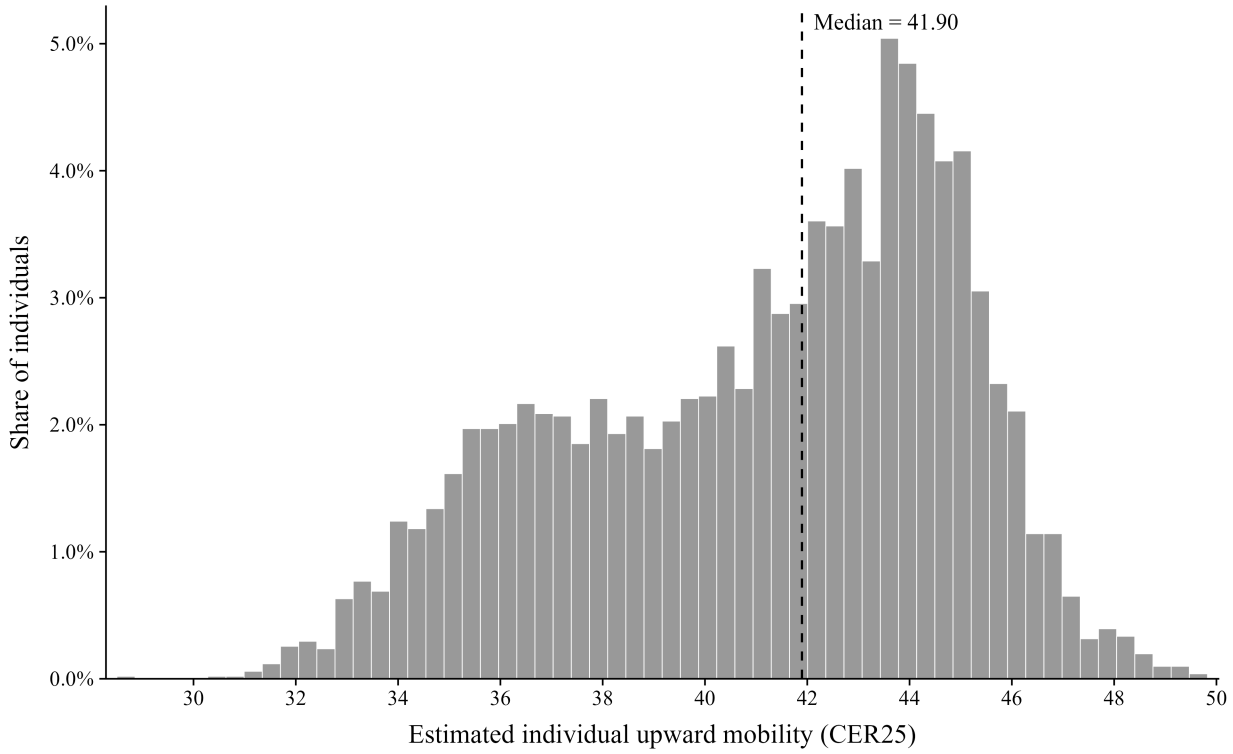


Figure 1: Distribution of Estimated Individual-Level Upward Mobility

Notes: This figure displays the empirical distribution of estimated upward mobility at the individual level obtained from the Conditional Mobility Forest — measured as the estimated child income rank conditional on parental income at the 25th percentile (CER_{25}).

5.2 Calibration and Stability of Individual CER_{25} Estimates

I conduct three validation checks to assess the reliability of the individual-level CER_{25} estimates. These exercises examine whether the estimated upward mobility values are well calibrated to observed outcomes, robust to alternative hyperparameter choices, and stable under resampling. Online Appendix C reports the full results.

First, I assess calibration by comparing the CMF estimates with realized outcomes among children whose parents are located near the 25th percentile of the parental income distribution. For this group, child outcomes at the evaluation point of CER_{25} are directly observed, which provides a transparent check of the estimates against realized mobility. I consider bands of increasing width around the 25th percentile, ranging from ± 1 to ± 10 percentile ranks and containing between 153 and 1,064 individuals. Within each band, I compare the average estimated CER_{25} with two measures of realized outcomes. The first is the average observed child income rank within the band. The second is the observed child rank evaluated at exactly the 25th percentile, obtained from a local linear regression of child rank on parental rank within the band. The second measure adjusts for the fact that the average parental rank within a band need not coincide with the 25th percentile. The estimates closely track realized outcomes. The average estimated CER_{25} is stable across bands at 41.3 to 41.4 rank points, while the observed outcomes range from 40.2 to 42.3 rank points. The average absolute difference between the estimated CER_{25} and the observed outcome is 0.70 rank points for the within-band mean and 0.54 rank points for the local linear value at the 25th percentile. The largest difference in any band is 1.2 rank points. This close agreement indicates that the CMF estimates are well calibrated.

Second, I examine sensitivity to alternative hyperparameter choices. I re-estimate the CMF under different tree depths, numbers of candidate split variables, split thresholds, subsampling rates, minimum leaf sizes, split penalties, and heterogeneity thresholds, as well as under a different random seed. Across all specifications, the median predicted CER_{25} remains between 41.8 and 42.7 rank points, compared with 42.7 in the baseline, and individual-level predictions remain strongly rank-correlated with the baseline estimates (Spearman correlations between 0.92 and 1.00). For all specifications except one, the median absolute deviation from the baseline predictions is below one rank point, and the divergence is of similar magnitude to that obtained by merely changing the random seed (Spearman correlation of 0.96). Increasing the maximum tree depth leaves the estimates unchanged, indicating that the depth constraint does not bind. The ranking of individuals by estimated upward mobility and the presence of meaningful heterogeneity are therefore not driven by a particular hyperparameter choice.

Third, I evaluate sensitivity to sampling variation. I re-estimate the CMF on 100 random subsamples

containing 90 percent of the observations and compare the resulting individual-level estimates with the full-sample estimates. The estimates remain strongly rank-correlated with the full-sample estimates (mean Spearman correlation 0.94; 5th percentile 0.93), and the mean absolute deviation is 0.95 rank points, about one quarter of the cross-sectional standard deviation of the estimates. Most of this variation stems from the randomness of the forest rather than from the omitted observations: re-estimating on the full sample with only a different random seed yields a Spearman correlation of 0.96 and a mean absolute deviation of 0.78 rank points. Membership in narrowly defined tails is less precisely determined than the overall ranking. Across resamples, roughly two-thirds of individuals in the bottom five percent of the full-sample distribution also fall in the bottom five percent of the resampled distribution (Jaccard index 0.52, compared with 0.03 under random assignment). Together with the calibration and hyperparameter checks, these results indicate that the ranking of individuals by estimated upward mobility and the presence of substantial heterogeneity are robust.

5.3 Benchmarking the CMF Against Conventional Subgroup Analysis

A natural benchmark for the Conditional Mobility Forest (CMF) is an exhaustive search over conventional predefined subgroups. If the heterogeneity identified by the CMF simply reflected arbitrary partitioning of the data, a sufficiently rich search over predefined subgroups should recover more extreme low-mobility groups. Conversely, if the CMF captures systematic heterogeneity, it should reach the same empirical frontier while identifying the relevant population in a more flexible way.

To construct this benchmark, I enumerate all one- to four-way interactions among observed categorical characteristics and estimate a separate weighted rank-rank regression within each subgroup.[5] To avoid selecting extremely small cells with imprecise estimates, I restrict attention to subgroups containing at least 200 observations. Candidate subgroups are ranked according to the upper bound of their 95 percent confidence interval for CER_{25}. First, individuals are ranked by their estimated individual-level CER_{25} values, and those in the bottom five percent of the distribution are assigned to the CMF low-mobility group. Second, a conventional weighted rank–rank regression is estimated within this group, yielding a CER_{25} estimate based on the same estimator used for the predefined subgroups. Table 1 compares the CMF bottom five percent mobility group with the most extreme reliable predefined subgroup identified from the search. The best predefined subgroup consists of children with low parental education who are African-American and has an estimated CER_{25} of 25.01. The corresponding estimate for the CMF low-mobility group is 24.41. Because

[5]The search is restricted to categorical characteristics because the literature documents the largest differences in upward mobility along these dimensions, such as race and gender (Jácome et al. 2025). Moreover, incorporating continuous covariates would require discretizing them, substantially increasing the number of candidate cells relative to the available sample size and thereby limiting the reliability of the resulting estimates.

the two groups overlap substantially (117 individuals belong to both groups), standard independent-sample inference would underestimate uncertainty. I therefore use a stacked sandwich estimator that accounts explicitly for the covariance between the two CER_{25} estimates. The estimated difference is -0.60 rank points with a standard error of approximately 2.81, a 95 percent confidence interval of $[-6.10, 4.89]$, and a p-value of 0.83. Thus, the CMF low-mobility group and the best reliably estimated predefined subgroup cannot be statistically distinguished in terms of expected upward mobility.

Table 1: CMF Low-Mobility Group versus Best Predefined Subgroup

Group	CER_{25}	SE	N	Upper 95% CI
CMF bottom 5%	24.41	1.23	254	26.83
Best predefined subgroup	25.01	1.44	302	27.84
Difference (CMF − predefined)	-0.60			
95% confidence interval		$[-6.10,\ 4.89]$		
p-value		0.83		

Notes: The predefined subgroup is the most reliable subgroup identified by an exhaustive search over all one- to four-way interactions among observed categorical characteristics, subject to a minimum cell size of 200 observations. Candidate subgroups are ranked according to the upper bound of the 95 percent confidence interval for CER_{25}. The best predefined subgroup corresponds to children with low parental education who are African-American. Because the two groups overlap (117 observations belong to both groups), inference accounts for the covariance between the two CER_{25} estimates using a stacked sandwich estimator.

The comparison in Table 1 highlights one advantage of the CMF: it reaches the same reliable frontier as an exhaustive predefined search without requiring the relevant interaction structure to be specified in advance. The CMF identifies a similarly extreme low-mobility group without being informed which characteristics should define it.

A second limitation of predefined subgroup analysis is that the frontier saturates as interaction depth increases. Table 2 reports the results of exhaustive searches over progressively higher-order interactions. Although the number of candidate subgroups grows rapidly—from 31 one-way cells to 390 two-way cells, 2,387 three-way cells, and 7,565 four-way cells—the most extreme reliably estimable subgroup does not improve beyond the two-way interaction frontier. The best upper confidence bound declines from 35.71 for one-way interactions to 27.84 for two-way interactions, but remains similar for three- and four-way interactions (29.24 and 27.92, respectively). This saturation reflects a fundamental limitation of predefined subgroup search. Additional interaction dimensions can only identify more extreme groups by repeatedly subdividing the sample. Eventually, cells become too small for precise estimation, and deeper interactions no longer improve

the reliable frontier. The CMF reaches the same frontier adaptively, without requiring the researcher to enumerate thousands of possible interaction structures.

Table 2: Exhaustive Search over Higher-Order Predefined Subgroups

Order	Variable sets	Candidate cells	Reliable cells	Best CER_{25} upper bound	Best subgroup
1-way	8	31	24	35.71	African-American
2-way	28	390	159	27.84	Low education × African-American
3-way	56	2,387	372	29.24	–
4-way	70	7,565	424	27.92	–
CMF bottom 5%	–	–	–	26.83	Data-driven

Notes: The table reports the most extreme reliably estimable subgroup obtained from exhaustive searches over predefined interactions. Reliable subgroups are required to contain at least 200 observations. Candidate groups are ranked by the upper bound of the 95 percent confidence interval for CER_{25}. The CMF low group is reported for comparison.

The predefined search also faces an inferential challenge that is abstracted from in the comparison above. Selecting the most extreme cell among thousands of candidates constitutes a specification search, such that the nominal confidence interval of the selected subgroup understates the true uncertainty unless the selection process is explicitly accounted for. By construction, the CMF avoids this issue because honest sample splitting separates the data used to discover heterogeneity from the data used to estimate subgroup-specific upward mobility. This further highlights the advantages of a data-driven approach, which not only offers greater flexibility in identifying relevant dimensions of heterogeneity but also allows the data to guide the identification of these dimensions.

5.4 Drivers of Heterogeneity in Social Mobility

What accounts for the heterogeneity in children's expected ranks at the 25th percentile of parental income? While the previous section provided descriptive evidence on the overall variation in upward mobility, this section investigates which factors are most strongly associated with that variation. As a first step, I examine the correlations between individual upward mobility estimates and the covariates included in the estimation procedure. Naturally, these associations should not be interpreted causally, but serve as a descriptive starting point.

In a second step, I employ SHapley Additive exPlanations (SHAP) analysis, following S. M. Lundberg et al. (2020), to decompose each individual's upward mobility value, relative to the sample-average estimate, into additive contributions of the input variables. SHAP applies Shapley values from cooperative game theory (Shapley 1953). Each variable is assigned its marginal contribution to the estimate, averaged over all possible combinations of the remaining variables.

This provides a theoretically grounded and practically appealing framework for assessing variable importance at the individual level (Molnar 2025). SHAP values satisfy several key axioms that make them well suited for explaining complex, nonlinear estimation models in a transparent and interpretable manner (S. Lundberg and Lee 2017). Specifically, they satisfy local accuracy (the prediction for an individual equals the average prediction plus the sum of the contributions of all variables), consistency (if a variable becomes more influential in the model, its attributed contribution cannot decrease), and the null-player property (a variable that does not affect the model's predictions receives zero attribution).

5.4.1 Univariate Associations with Upward Mobility Outcomes

As a first descriptive exercise, I examine the unconditional association between individual upward mobility estimates, and each observed background characteristic. Figure 2 reports the twenty largest Pearson correlations together with their 95% confidence intervals.

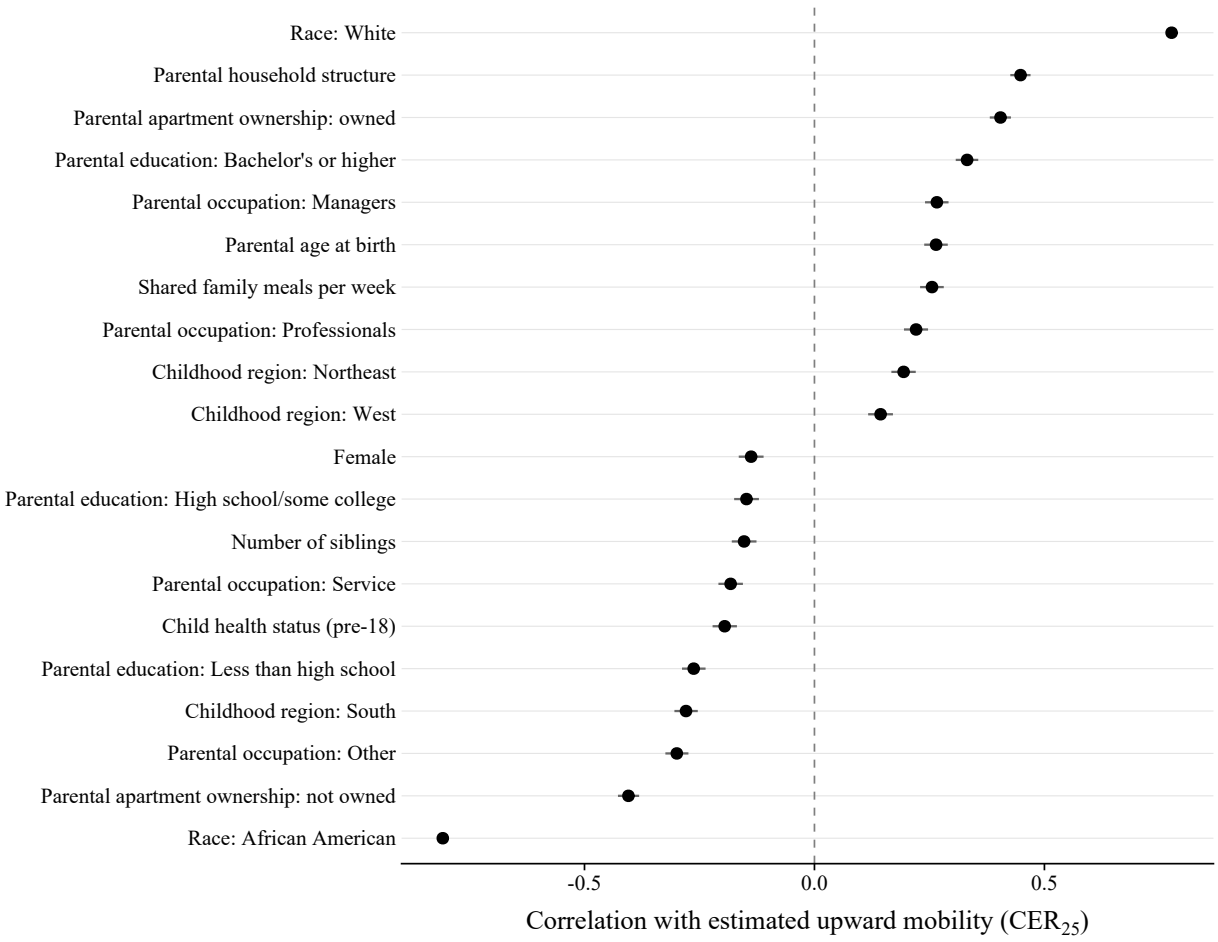


Figure 2: Top 20 Correlates of Estimated Upward Mobility (CER_{25})

Notes: Pearson correlations between each observed covariate and the estimated upward mobility values, shown with 95% confidence intervals. The figure displays the twenty strongest univariate associations.

Several broad patterns emerge. First, race is by far the strongest correlate of estimated upward mobility. Identifying as White exhibits the largest positive association with upward mobility, while identifying as African-American exhibits the largest negative association. These correlations substantially exceed those of all other covariates, highlighting the central role of racial disparities in the distribution of upward mobility outcomes. The finding is consistent with a large literature documenting persistent racial gaps in upward mobility. Using administrative tax records, Chetty et al. (2020) show that African-American children experience substantially lower rates of upward mobility even when raised in comparable neighbourhoods, while Jácome et al. (2025) document that these racial differences have persisted throughout much of the twentieth century. More recently, Chang et al. (2025) likewise identify race as one of the principal background characteristics associated with low upward mobility.

Second, the remaining positive correlates predominantly reflect dimensions of parental

socioeconomic status. Growing up in a two-parent household and living in a household that owns its apartment are among the strongest positive associations after race. Higher parental education and employment in managerial or professional occupations are likewise positively correlated with estimated upward mobility, as are older parental age at birth and more frequent shared family meals. Together, these variables point toward the importance of family resources, human capital, and the broader home environment. The positive association with parental education is consistent with evidence that parental education complements financial resources in promoting children's long-run economic outcomes (Chang et al. 2025), while the relationship with parental age mirrors the same study's finding that children born to older parents experience substantially higher rates of upward mobility. Frequent shared family meals may capture broader dimensions of family stability and parental investment. Geographic differences also appear in the expected direction. Childhood residence in the Northeast and West is positively associated with upward mobility, whereas childhood residence in the South is negatively correlated. These regional differences closely resemble the geographic heterogeneity in upward mobility documented by Chetty et al. (2014a), who identify substantially lower rates of upward mobility throughout much of the southeastern United States.

Among the negative correlates, parental non-ownership of the household's apartment represents the largest association after race, followed by low parental education, unknown parental occupation, and residence in the South. Lower self-reported childhood health also appears among the strongest negative correlates. Although modest in magnitude relative to race, this association is noteworthy given extensive evidence that childhood health affects later educational attainment, labour market success, and ultimately economic outcomes (Case and Paxson 2006; Currie 2009). Gender exhibits comparatively weak unconditional associations. Being male is associated with slightly higher upward mobility. The relatively small magnitude of these coefficients is consistent with recent evidence suggesting only modest gender differences in upward mobility (Jácome et al. 2025; Chang et al. 2025).

Taken together, these descriptive patterns closely mirror the determinants of upward mobility identified in the existing literature. At the same time, these results remain purely descriptive. Each correlation captures the association between a single characteristic and estimated upward mobility without accounting for the fact that many background characteristics are themselves strongly correlated. For example, race, parental education, household structure, and region jointly characterize childhood disadvantage and therefore cannot be interpreted independently. The SHAP decomposition presented in Section 5.4.2 addresses precisely this limitation by quantifying each variable's contribution to upward mobility estimates while accounting for the joint dependence among variables. Although this approach provides a more comprehensive picture of how different

characteristics contribute to observed patterns of upward mobility, the results remain descriptive and cannot be interpreted as causal effects.

5.4.2 Explaining Heterogeneity in Upward Mobility Using SHAP

Having established the strongest correlates of estimated upward mobility, I next examine the relative importance of background characteristics after accounting for the joint distribution of observed covariates. To do so, I use SHapley Additive exPlanations (SHAP), an additive variable-attribution framework (S. M. Lundberg et al. 2020). Whereas the correlations reported in the previous section describe unconditional associations, SHAP decomposes each individual estimate into the contributions of the observed covariates relative to the average estimate while fully accounting for nonlinearities and interaction effects captured by the Conditional Mobility Forest (CMF). These importance measures should be interpreted descriptively rather than causally. A causal interpretation would require that all relevant sources of heterogeneity be observed and adequately accounted for, an assumption that cannot generally be guaranteed in observational data. Nevertheless, unlike unconditional associations, the SHAP measures account for the joint distribution of the included covariates and the nonlinear and interactive relationships captured by the model. They therefore provide a more complete picture of which variables are associated with differences in estimated upward mobility than approaches that examine each variable separately, without accounting for its relationship with other characteristics.

Exact SHAP algorithms are not available for customized estimators such as the Conditional Mobility Forest (CMF), and computing Shapley values directly from the CMF for the full sample is computationally prohibitive. I therefore estimate a gradient boosting decision tree (GBDT) surrogate on the same covariate set to reproduce the individual-level CER_{25} estimates generated by the CMF, and compute SHAP values for the surrogate using the exact TreeSHAP algorithm.[6] The surrogate reproduces the CMF estimates closely. Across the 5,077 observations in the analysis sample, it accounts for 96.6 percent of the variation in the CMF estimates ($R^2 = 0.966$). The root mean squared prediction error is about 0.56 percentile-rank points, compared with a cross-sectional standard deviation of the CMF estimates of 3.05 rank points.

Prediction fidelity alone does not guarantee that the surrogate attributes the estimates to covariates in the same way as the CMF. I therefore validate the attributions directly. For a random validation subsample of 100 individuals, I compute Monte Carlo Shapley values from the CMF itself and compare them with the surrogate's TreeSHAP values. The two decompositions agree closely. The ranking of variables by mean absolute contribution has a Spearman correlation of $\rho = 0.97$ with the

[6]The surrogate is implemented with the `LightGBM` package in R. Using surrogate models to interpret complex prediction algorithms is standard practice in interpretable machine learning (Molnar 2025).

CMF ranking. The individual-level contributions have a Pearson correlation of $r = 0.97$, close to the ceiling of about 0.99 implied by Monte Carlo noise in the CMF reference. Applying the identical Monte Carlo estimator to both models isolates discrepancies due to surrogate misfit. Under this comparison, the correlation of individual-level contributions rises to $r = 0.99$, while the rank correlation is $\rho = 0.96$. All approaches identify race as by far the most important variable, with differences in rankings arising only among variables with very similar contribution values. The surrogate therefore replicates both the CMF estimates and the attribution of these estimates to observed characteristics through Shapley values. Details of the validation procedure are reported in Online Appendix D.2.

5.4.3 The Relative Importance of Background Variables

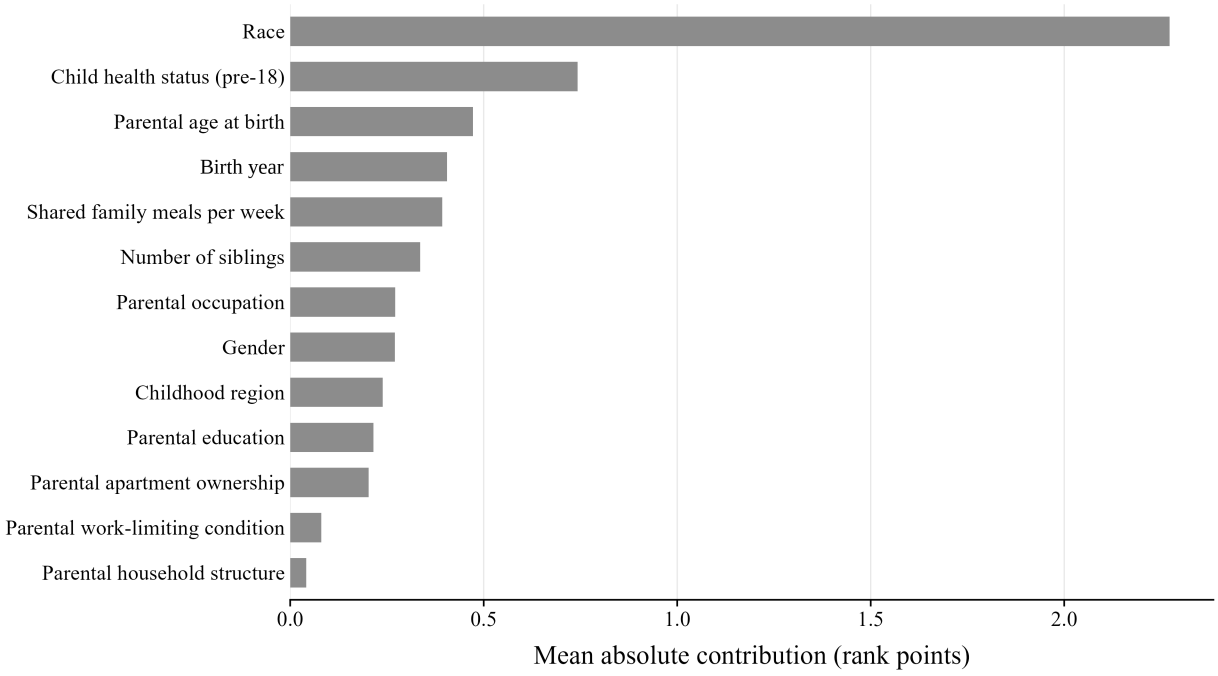


Figure 3: Variable importance for upward mobility (CER_{25})

Notes: Mean absolute SHAP contribution of each background variable to upward mobility. The categories of each categorical variable are aggregated into a single per-individual contribution before averaging, so that each bar reports the importance of the underlying variable rather than of an individual encoded category.

Figure 3 reports the mean absolute SHAP contribution of each background characteristic to estimated upward mobility. For each individual, the SHAP value of a variable measures how much that variable moves the individual's estimated CER_{25} value away from the sample-average estimate, holding the contributions of all other variables fixed in the additive decomposition. Averaging the absolute values of these contributions across individuals yields a measure of how strongly each

variable shifts the estimates, irrespective of the direction of the shift. The most striking result is the overwhelming importance of race. Its mean absolute SHAP contribution is approximately 2.27 rank points. That is, race accounts on average for about 2.27 percentile ranks of the deviation of a child's estimated CER_{25} from the sample-average estimate. This contribution is more than three times larger than that of the second-ranked variable, childhood health. This finding closely mirrors the extensive literature documenting persistent racial disparities in upward mobility (Chetty et al. 2020; Jácome et al. 2025; Chang et al. 2025). The results indicate that race remains the most important variable for explaining heterogeneity in upward mobility even after accounting for an unusually rich set of childhood demographic, socioeconomic, and family characteristics.

Beyond race, the SHAP ranking differs markedly from the ordering implied by the univariate correlations. Childhood health emerges as the second most important variable, followed by parental age at birth, birth year, shared family meals, and number of siblings. Childhood health contributes substantially more to the model than conventional socioeconomic indicators such as parental education or apartment ownership. This result provides strong support for the childhood health literature, which argues that early-life health constitutes an important mechanism through which family background shapes adult socioeconomic outcomes (Case and Paxson 2006; Currie 2009). Rather than merely proxying parental socioeconomic status, childhood health appears to capture an independent dimension that remains important after taking the broader childhood environment into account.

While household structure, parental education and apartment ownership showed strong correlations with upward mobility, their importance declines sharply once the overlap among family characteristics is removed. This pattern suggests that much of their association with upward mobility reflects characteristics they share with other variables rather than an independent relationship. The finding is consistent with Bloome (2017), who concludes that childhood family structure explains relatively little variation in income mobility after conditioning on broader family socioeconomic circumstances. In contrast, parental age at birth remains important in the SHAP analysis, consistent with Chang et al. (2025), who identify parental childbearing age as one of several family-level bottlenecks that contribute to persistent intergenerational inequality. Birth year also emerges as an important source of heterogeneity in upward mobility, consistent with Jácome et al. (2025), who show that upward mobility rates vary across birth cohorts. In contrast to other data sources, the PSID provides information that allows for the inclusion of proxies for parenting practices, highlighting the importance of the frequency with which parents and children share the main meal of the day. More recently, García and Heckman (2023) identified an independent association between parenting and upward mobility, consistent with the findings presented here.

Overall, the SHAP decomposition substantially alters the relative importance of background

characteristics compared with their unconditional correlations. Variables capturing family resources and living arrangements appear strongly associated with upward mobility in isolation but contribute comparatively little once their shared variation with other characteristics is taken into account. By contrast, race, childhood health, and parental age at birth remain important for characterizing heterogeneity in upward mobility.

5.4.4 Direction and Dispersion of Variable Contributions

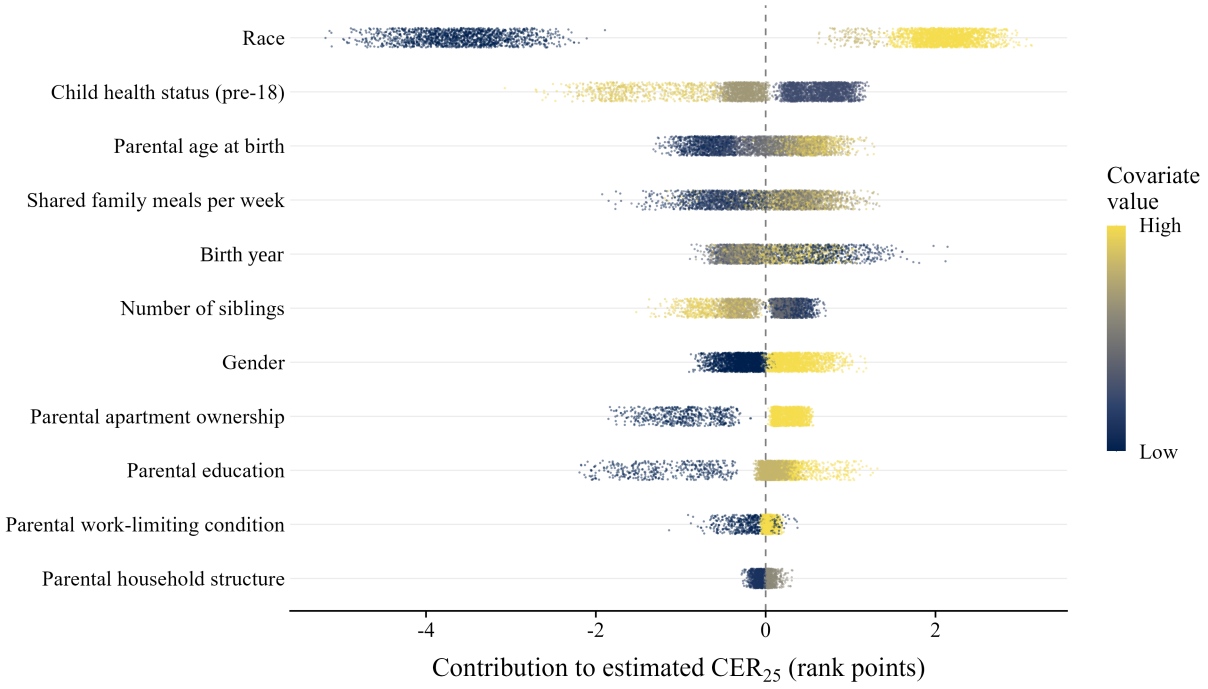


Figure 4: Individual-level SHAP contributions to estimated upward mobility (CER_{25})

Notes: Each point is an individual. The horizontal position is that individual's SHAP contribution (in rank points) from the covariate in the corresponding row, and the colour encodes the covariate's value from low (blue) to high (yellow). Parental occupation and childhood region are not shown because their categories have no natural low-to-high ordering and therefore cannot be represented on the colour scale.

Figure 4 complements the aggregate importance rankings by showing the distribution and direction of individual Shapley values for each variable. Whereas Figure 3 identifies which variables matter most overall, the beeswarm plot illustrates how each characteristic shifts estimated upward mobility for individual children. For categorical variables, the contributions of the individual categories are aggregated into a single per-individual value. Each point in the figure represents one individual. Its horizontal position shows how much the variable in that row shifts the individual's estimated CER_{25} above or below the sample-average estimate, and its colour shows the individual's value of that variable, from low (blue) to high (yellow). For instance, a yellow point to the left of zero describes

an individual with a high value of the variable for whom this value lowers estimated upward mobility. Read together, position and colour therefore show whether higher values of a variable tend to be associated with higher or lower estimated upward mobility. Parental occupation and childhood region are not shown, as their categories have no natural ordering.

Race again dominates the distribution of contributions. Its Shapley values split into two clearly separated clusters: a large negative cluster, in which upward mobility is reduced by roughly two to five rank points, and a positive cluster, in which it is raised by roughly one to three rank points. This pattern indicates that African-American children face substantially lower estimated upward mobility. These patterns closely parallel the evidence of Chetty et al. (2020), who document substantially lower upward mobility among African-American children even conditional on neighbourhood environment, as well as the nationally representative estimates of Jácome et al. (2025), which show persistent racial disparities in upward mobility across successive birth cohorts.

Childhood health also shows substantial dispersion: poor health before age 18 lowers estimated mobility by up to almost three rank points, while good health raises it by up to just over one rank point. Older parental age at birth and more frequent shared family meals contribute positively to upward mobility, with contributions of roughly one rank point in either direction, while having more siblings contributes negatively. Birth year, by contrast, shows no clear direction: individuals from both earlier and later birth cohorts appear on either side of zero, with contributions ranging from about -1 to $+2$ rank points. Gender separates into two distinct groups: being male raises estimated upward mobility, while being female lowers it, with contributions of up to about one rank point in either direction. Parental apartment ownership shows a pronounced asymmetry: children whose parents did not own their apartment receive negative contributions of up to almost two rank points, whereas ownership raises estimated mobility only modestly.

The socioeconomic indicators also follow expected gradients. Low parental education produces a distinct cluster of negative contributions of up to about two rank points, whereas higher parental education contributes positively, by up to just over one rank point. Parental work-limiting conditions and parental household structure show only small contributions in either direction, largely within one rank point of zero. Overall, the individual SHAP contributions reveal not only the direction of each variable's association with estimated upward mobility but also how its importance is distributed across individuals. Contributions are often asymmetric. For race and childhood health, the negative contributions are considerably larger in magnitude than the positive ones. The mean absolute contribution in Figure 3 conceals this distinction, which becomes visible only in the beeswarm plot.

6 Discussion and Conclusion

Research on intergenerational mobility has traditionally assessed economic opportunity using aggregate mobility statistics or comparisons across a small number of predefined subgroups. Both approaches necessarily average over much of the heterogeneity that shapes individuals' opportunities. This paper takes a more fine-grained approach to examining heterogeneity in upward mobility, without requiring the ex-ante specification of relevant dimensions of heterogeneity or assumptions about functional relationships. To do so, I introduce the Conditional Mobility Forest (CMF), a nonparametric estimator that produces individualized estimates of the Conditional Expected Rank at the 25th percentile of the parental income distribution (CER_{25}). Applying the method to data from the Panel Study of Income Dynamics, I document three main findings.

First, average measures of upward mobility conceal substantial heterogeneity in economic opportunity. Estimated CER_{25} values span roughly twenty percentile ranks. Validation exercises demonstrate that this dispersion reflects stable and well-calibrated differences rather than estimation noise. Aggregate social mobility statistics therefore provide only a coarse description of the opportunity structure within a society as children face very different prospects for upward mobility from the bottom.

Second, the paper provides new evidence on the sources of heterogeneity in upward mobility. Descriptive correlational analysis across an unusually rich set of demographic, socioeconomic, and family background characteristics broadly reproduces the patterns documented in the existing mobility literature. Race exhibits by far the strongest association with upward mobility, while parental education, apartment ownership, parental occupation, childhood health, and parental age at birth also display sizable unconditional relationships. These descriptive associations provide a useful benchmark but do not account for the dependence structure among variables.

Third, applying SHAP analysis substantially changes this picture. By approximating each variable's contribution to estimated upward mobility while jointly accounting for the full set of observed variables and the nonlinear relationships captured by the CMF, the decomposition provides a considerably richer descriptive assessment of variable importance than unconditional correlations alone. Race remains by far the most important variable, followed by childhood health, parental age at birth, birth year, and shared meal frequency. In contrast, household structure and apartment ownership, despite their strong unconditional associations with upward mobility, contribute comparatively little once their overlap with other variables is accounted for. The SHAP results thus have important implications for understanding which dimensions are most closely linked to heterogeneous upward mobility prospects. In particular, the prominence of childhood health—whose contribution exceeds that of parental education or apartment ownership—suggests

that early-life health captures an important dimension of disadvantage beyond family socioeconomic resources alone. Likewise, the fact that race remains the dominant variable of upward mobility even after accounting for an unusually rich set of observed childhood characteristics underscores the magnitude of racial disparities in economic opportunity. To my knowledge, this constitutes the first joint assessment of variable importance for heterogeneity in upward mobility across such a broad set of characteristics. While unobserved factors prevent a complete account, the resulting ranking approximates which dimensions of childhood circumstances are most closely linked to constrained opportunity. This matters for policy. Programs aimed at promoting upward mobility typically target children by parental income alone. The results suggest that targeting could be sharpened by also considering dimensions such as early-life health, which identify children with particularly low prospects even at the same level of parental income.

Beyond these empirical findings, the paper contributes a general framework for studying heterogeneity in social mobility. Unlike conventional approaches, which require researchers to specify ex ante both the dimensions along which social mobility may vary and the functional relationships governing this heterogeneity, the proposed framework allows both to be determined flexibly from the data. Like conventional subgroup and interaction analyses, the CMF is descriptive rather than causal and is best understood as a data-adaptive extension of conventional subgroup analysis. Comparing the CMF with exhaustive searches over predefined subgroups illustrates its advantages. Although conventional searches can identify groups with levels of upward mobility comparable to those identified by the CMF, they quickly encounter a precision frontier as the depth of interactions increases. The use of honest sample splitting further separates subgroup discovery from estimation, mitigating the specification-search concerns inherent in exhaustive searches. The framework can also be readily extended to other measures of intergenerational persistence. As increasingly rich administrative and longitudinal data become available, this approach offers a valuable complement to traditional methods of heterogeneity analysis.

Several limitations bear emphasis. First, the analysis is descriptive rather than causal. Second, the estimator does not provide formal pointwise inference for individual CER_{25} estimates. Instead, its reliability is assessed through empirical validation exercises. Third, the analysis of heterogeneity is limited to variables observed in the PSID, leaving scope for unmeasured family, neighbourhood, and institutional factors to account for additional heterogeneity in upward mobility. Finally, the findings pertain to income mobility in the United States and may not generalize to other institutional settings or to upward mobility in wealth, education, or occupation. Extending the framework to richer administrative data and comparative international settings is therefore a natural next step.

Overall, the results show that aggregate mobility statistics conceal substantial heterogeneity in economic opportunity that cannot be adequately captured by a small set of predefined subgroups.

Assessing the prospect of upward mobility therefore requires going beyond summary measures and conventional subgroup comparisons. By bringing flexible machine learning to the study of upward mobility, the Conditional Mobility Forest provides a more granular picture of how opportunities vary across individuals and which observed characteristics are most strongly associated with upward mobility.

Appendix A: Conditional Mobility Forest Estimator

A.1 Objective and Motivation

I develop a custom mobility forest-based estimator designed to flexibly capture heterogeneity in the intercept of the following baseline model:

$$Y_c = \alpha(X_c) + \beta(X_c)\,(Y_p - x_0) + \varepsilon_c, \tag{3}$$

where Y_c denotes the child's income rank, Y_p is the parental income rank, and X_c is a vector of observed covariates. The central object of interest is the intercept function $\alpha(X_c)$, which represents the expected child income rank at a fixed parental rank benchmark x_0. The slope $\beta(X_c)$ may also vary with the covariates. It is estimated locally alongside the intercept but is not itself an object of interest. Heterogeneity in $\alpha(X_c)$ reflects differences in social mobility across individuals. In the application in the paper, I set $x_0 = 25$, corresponding to the bottom quartile of the parental income distribution.

A.2 Estimation Strategy

The estimator is an ensemble of honest regression trees that target heterogeneity through recursive partitioning and adaptive kernel weighting. The procedure unfolds as follows:

1. **Subsampling:** For each tree, I draw a subsample $\mathcal{P} \subseteq \{1, \ldots, n\}$ of 70% of the data (without replacement).

2. **Honest Sample Splitting:** The subsample $\mathcal{P}$ is divided at random into two disjoint subsets of equal size:

    - A splitting sample $\mathcal{P}_{\text{split}}$, used exclusively to determine splits.
    - An estimation sample $\mathcal{P}_{\text{est}}$, which is used exclusively to identify neighbors and construct the corresponding weights for out-of-bag observations.

This honesty principle ensures that the model does not reuse data for both structure learning and parameter estimation, reducing overfitting bias.

3. **Candidate Splits and Local Fits:** Categorical variables enter as one-hot indicators, one for each category (no reference category is omitted), so every indicator is a separate candidate covariate and a split on it separates one category from all others. At each node, a random number of m_{try} covariates is considered, where m_{try} is drawn from a Poisson distribution with mean 7 and truncated to lie in $[1, p]$, with p the number of available covariates. Categorical variables with many categories are therefore offered as candidates more often than binary ones. For each candidate covariate j and threshold θ, the splitting sample is partitioned into:

$$\mathcal{P}_L = \{i \in \mathcal{P}_{\text{split}} : X_{ij} \le \theta\}, \qquad \mathcal{P}_R = \{i \in \mathcal{P}_{\text{split}} : X_{ij} > \theta\}.$$

Within each child node, a survey-weighted linear regression of Y_c on Y_p—using the PSID longitudinal individual weights—is fit and evaluated at x_0, giving the local estimates $\hat{\alpha}_L(x_0)$ and $\hat{\alpha}_R(x_0)$. Weighting the split-stage fits ensures that the partition itself is learned on a sample that is representative of the national population rather than of the oversampled PSID design.

4. **Split Criterion:** Each candidate split is scored based on the squared difference in estimated outcomes at $x_0 = 25$, adjusted for balance and, optionally, their size:

$$\text{Score}(j, \theta) = \big(\hat{\alpha}_L(x_0) - \hat{\alpha}_R(x_0)\big)^2 \cdot \underbrace{\frac{\text{ESS}_L \cdot \text{ESS}_R}{(\text{ESS}_L + \text{ESS}_R)^2}}_{\text{balance penalty}} - \frac{\lambda}{\sqrt{\text{ESS}_L + \text{ESS}_R}}.$$

Here, ESS_L and ESS_R are the effective sample sizes of the left and right child nodes, defined as $\text{ESS} = (\sum_i w_i)^2 / \sum_i w_i^2$, a standard measure of kernel concentration, where w_i are the PSID longitudinal individual weights. Node sizes and balance are therefore judged in survey-weighted terms. The second term penalizes small nodes. In the application $\lambda = 0$, so candidate splits are ranked by the balance-weighted squared difference alone. A candidate split is admissible only if each child node has an effective size of at least min_size and carries at least a share κ of the combined effective sample size. A split is accepted only if the penalized heterogeneity score exceeds a minimal threshold proportional to the variance of Y_c in the node's splitting sample. For a continuous covariate j, the candidate thresholds θ are the empirical quantiles of X_{ij} in the splitting sample, evaluated on an equally spaced grid between the 10th and 90th percentiles, so that extreme cut points generating very small child nodes are never considered.

To ensure common support, I additionally require that the evaluation point x_0 lies within the central 90% range (5th–95th percentile) of Y_p in the splitting sample of both child nodes. This prevents extrapolation and ensures that both subgroups contribute meaningful local information at the point of interest.

5. **Recursive Partitioning and Leaf Estimation:** The best-scoring split is applied recursively. Partitioning stops when:

   - Either child node has effective size below min_size,
   - The tree reaches max_depth, or
   - No candidate split improves the heterogeneity score beyond the threshold.

6. **Sparse OOB Weight Construction:** After all trees are grown, personalized neighbourhoods are constructed for each child using a sparse, honest kernel. For each tree:

   - Leaf assignment for every observation is identified.
   - For each out-of-bag (OOB) observation i — that is, one not used in the tree's construction — estimation-sample observations that share the same leaf as i are collected.

   These estimation-sample units form i's honest neighbors in this tree and each receives weight $1/|L_t(i)|$, where $L_t(i)$ is the set of estimation-sample units in i's leaf of tree t. Across all trees, the kernel weights are averaged and normalized by the number of trees in which i was OOB and had non-empty neighbourhoods. The resulting kernel measures how often each honest neighbor shares a leaf with i. It captures which observations are comparable to i. This ensures that each child's local regression is based only on honest predictions while remaining representative of the national population.

7. **Forest Aggregation and Final Estimation:** For each child i, a weighted regression is fit over i's sparse forest neighbourhood:

$$Y_c = \hat{\alpha}_i + \hat{\beta}_i(Y_p - x_0) + \varepsilon_c,$$

   where each neighbor's weight is the product of the honest forest kernel and that neighbor's PSID longitudinal individual weight (renormalized to sum to one). Survey weighting therefore enters at every stage of the procedure—split selection, effective-node-size and balance conditions, and the final estimation step—so that the reported estimates are representative of the national population rather than of the oversampled PSID sample. The estimated intercept $\hat{\alpha}_i$ represents the expected child rank at $x_0 = 25$.

A.3 Implementation Details

The estimator is implemented in base R and supports the following tunable hyperparameters:

- `min_size`: Minimum effective node size (default: 5).
- `max_depth`: Maximum depth of each tree (default: 20).
- `alpha_param` (κ): Minimum share of the combined effective sample size required in each child node (default: 0.00).
- `min_heterogeneity_fraction`: Minimum gain in heterogeneity required to split (default: 0.00).
- `n_thresholds`: Number of split thresholds evaluated per continuous covariate (default: 20).
- `m_try_mean`: Number of covariates randomly drawn at each node from a Poisson distribution (default: 7).
- `n_trees`: Number of trees in the forest (default: 1000).
- `subsample_frac`: Share of the sample drawn for each tree (default: 0.7).
- `lambda` (λ): Weight of the node-size penalty in the split score (default: 0).

A.4 Comparison to Generalized Random Forests

My approach is conceptually related to the Generalized Random Forest (GRF) framework of Athey et al. (2019), but differs in several key respects. First, while GRFs are typically designed to estimate heterogeneity in treatment effects or slope parameters in conditional moment models, my method is tailored to uncover heterogeneity in an intercept-like quantity: the expected child income rank at the fixed parental income rank $x_0 = 25$. Second, the splitting criterion departs from the GRF approach by directly evaluating the squared difference in estimated outcomes at x_0, rather than optimizing a gradient-based approximation or pseudo-outcome. This ensures that the tree-growing process is explicitly aligned with the target parameter, albeit at the cost of higher computational complexity. Third, I explicitly impose a common support condition when evaluating candidate splits, requiring that the evaluation point x_0 lies within the central range (5th to 95th percentile) of the parent outcome Y_p in both resulting child nodes. This guards against extrapolation in rank-rank regressions, where extreme values (particularly from the top of the parental distribution) could otherwise disproportionately influence split decisions and lead to unstable or misleading heterogeneity estimates. Fourth, while GRFs are designed to support formal statistical inference through

influence-function-based standard errors and confidence intervals, my primary goal is not pointwise estimation or inference. Instead, I aim to identify broader patterns of heterogeneity in upward mobility across the covariate space. As such, I do not currently implement any formal variance estimation procedures. Because the asymptotic theory underlying GRF inference rests on honesty, regularity, and subsampling conditions rather than on a particular splitting criterion, pointwise inference for the CMF could in principle be obtained along the lines of (Athey and Wager 2019). Doing so would require modifying the baseline configuration, in particular imposing a minimum share of observations per child node and growing trees in groups on half-samples, and is left for future work. Overall, while the CMF is inspired by the GRF framework—incorporating honest sample splitting, recursive partitioning, sparse kernel weighting, and out-of-bag estimation—it is specifically adapted to the empirical task of analyzing heterogeneity in upward mobility.

A.5 Software and Reproducibility

The estimator is implemented in modular `R` code using the `future.apply` package for parallelization. All code to reproduce the estimator, including diagnostics, will be made publicly available upon publication. A full replication package is available from the author upon request.

Disclaimer:
During the preparation of this work, the author used GPT-6 for text editing and Claude Opus 5.5 for programming assistance. The author reviewed and revised all AI-assisted outputs and takes full responsibility for the final content of the manuscript.

References

Athey, S., J. Tibshirani, and S. Wager (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2.

Athey, S. and S. Wager (2019). *Estimating Treatment Effects with Causal Forests: An Application*.

Becker, G. S. and N. Tomes (1979). "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility". In: *Journal of Political Economy* 87.6, pp. 1153–1189.

Bloome, D. (2017). "Childhood Family Structure and Intergenerational Income Mobility in the United States". In: *Demography* 54.2, pp. 541–569.

Blundell, J. and E. Risa (2018). "Do Rank-Rank Income Mobility Measures Fully Capture Broader Parental Influence on Child Income?" In: *SSRN Electronic Journal*.

Case, A. and C. H. Paxson (2006). "Children's Health and Social Mobility". In: *The Future of Children* 16.2, pp. 151–173.

Chang, Y. et al. (2025). *Accounting for Individual-Specific Heterogeneity in Intergenerational Income Mobility*. w33349. Cambridge, MA: National Bureau of Economic Research, w33349.

Chernozhukov, V. et al. (2025). "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India". In: *Econometrica* 93.4, pp. 1121–1164.

Chetty, R. et al. (2014a). "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility". In: *American Economic Review* 104.5, pp. 141–147.

Chetty, R. et al. (2014b). "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States *". In: *The Quarterly Journal of Economics* 129.4, pp. 1553–1623.

Chetty, R. et al. (2020). "Race and Economic Opportunity in the United States: an Intergenerational Perspective*". In: *The Quarterly Journal of Economics* 135.2, pp. 711–783.

Chetty, R. et al. (2025). "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility". In: *American Economic Review (Forthcoming)*.

Cholli and Durlauf (2022). "INTERGENERATIONAL MOBILITY". In: *SSRN Electronic Journal*.

Corak, M. (2020). "The Canadian Geography of Intergenerational Income Mobility". In: *The Economic Journal* 130.631, pp. 2134–2174.

Currie, J. (2009). "Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development". In: *Journal of Economic Literature* 47.1, pp. 87–122.

Davis, J. and B. Mazumder (2018). "Racial and Ethnic Differences in the Geography of Intergenerational Mobility". In: *SSRN Electronic Journal*.

— (2026). "The Decline in Intergenerational Mobility after 1980". In: *Review of Economics and Statistics* 108.1, pp. 1–15.

Deutscher, N. and B. Mazumder (2023). "Measuring Intergenerational Income Mobility: A Synthesis of Approaches". In: *Journal of Economic Literature* 61.3, pp. 988–1036.

García, J. L. and J. J. Heckman (2023). "Parenting Promotes Social Mobility Within and Across Generations". In: *Annual Review of Economics* 15.1, pp. 349–388.

Jácome, E., I. Kuziemko, and S. Naidu (2025). "Mobility for All: Representative Intergenerational Mobility Estimates over the Twentieth Century". In: *Journal of Political Economy* 133.1, pp. 306–354.

Kenedi, G. and L. Sirugue (2023). "Intergenerational income mobility in France: A comparative and geographic analysis". In: *Journal of Public Economics* 226, p. 104974.

Kurer, T. and B. Van Staalduinen (2022). "Disappointed Expectations: Downward Mobility and Electoral Change". In: *American Political Science Review* 116.4, pp. 1340–1356.

Lundberg, S. and S.-I. Lee (2017). *A Unified Approach to Interpreting Model Predictions*. arXiv.

Lundberg, S. M. et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 56–67.

Mazumder, B. (2016). “Estimating the Intergenerational Elasticity and Rank Association in the United States: Overcoming the Current Limitations of Tax Data”. In: *Research in Labor Economics*. Ed. by L. Cappellari, S. W. Polachek, and K. Tatsiramos. Vol. 43. Emerald Group Publishing Limited, pp. 83–129.

Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable*. 3:e upplagan. OCLC: 1520125425. München: Christoph Molnar. 379 pp.

Nybom, M. and J. Stuhler (2025). “Geographic Variation in Multigenerational Mobility”. In: *Sociological Methods & Research* 54.4, pp. 1532–1575.

Shapley, L. S. (1953). “17. A Value for n-Person Games”. In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by H. W. Kuhn and A. W. Tucker. Princeton University Press, pp. 307–318.