International Inequalities Institute

WORKING PAPER 160

Inherited inequality and the distribution of opportunities in the United States, China, India, and South Africa

Paolo Brunori
Università di Firenze; International Inequalities Institute, LSE

Francisco H.G. Ferreira
International Inequalities Institute, LSE; IZA - Institute of Labor Economics

Pedro Salas-Rojo
CUNEF Universidad, Madrid; EQUALITAS

JANUARY 2026

Paolo Brunori
Università di Firenze; International Inequalities Institute, LSE

Francisco H.G. Ferreira
International Inequalities Institute, LSE; IZA - Institute of Labor Economics

Pedro Salas-Rojo
CUNEF Universidad, Madrid; EQUALITAS

This paper extends and supersedes III Working Paper No. 107. Please use this version instead.

In addition to our working papers series, all our publications are available to download free from our website: www.lse.ac.uk/III

International Inequalities Institute
The London School of Economics and Political Science, Houghton Street,
London WC2A 2AE

E inequalities.institute@lse.ac.uk
W www.lse.ac.uk/III

Inherited inequality and the distribution of opportunities in the United States, China, India, and South Africa

Paolo Brunori, Francisco H.G. Ferreira, and Pedro Salas-Rojo[1]

18 January 2026

Abstract: Researchers have sought to quantify the extent of inequality that is inherited from previous generations in multiple ways, including a large body of work on intergenerational mobility and inequality of opportunity. Many of the most frequently used approaches to measuring mobility or inequality of opportunity fit within a general framework which involves, as a first step, an estimation of the extent to which inherited personal characteristics can predict current incomes. We suggest a new method, within that broad framework, which is sensitive to differences across the entire conditional distributions of relevant population subgroups, rather than just in their means – a feature that makes it particularly well-suited to measuring ex-post inequality of opportunity. Sensitivity to differences in higher moments of the conditional distributions allow for a more comprehensive assessment of inherited inequality. We apply this approach to household income distributions in China, India, South Africa, and the United States, to illustrate how the method performs in different settings. We find that inherited inequality accounts for large shares of total inequality, from 36% in the United States to 59% in China, 62% in India, and 81% in South Africa.

Keywords: Inherited inequality; opportunity; mobility; transformation trees; China; India; South Africa; United States.

JEL Codes: D31, D63, J62

[1] Brunori is at Università di Firenze and the London School of Economics. Ferreira is at the London School of Economics and is also affiliated with IZA. Salas-Rojo is at CUNEF Universidad in Madrid and EQUALITAS. Correspondence to Pedro Salas-Rojo: pedro.salas@cunef.edu. We are grateful to Torsten Hothorn and Achim Zeileis for very helpful advice, and to Pedro Torres for superb research assistance. We also thank two anonymous reviewers, and seminar participants at the European Commission, the London School of Economics, the World Bank, and the Universities of Bari, Bicocca, la Laguna, Leeds, Queen Mary, Stockholm and Tilburg for useful comments on previous versions of this paper. All remaining errors are ours.

1. Introduction

People's educational and professional achievements, incomes, and wealth are generally not independent of their background. Various attributes that are determined at or before birth or during childhood – such as sex at birth; race, ethnicity, or caste; parental income and other aspects of family background – are powerful predictors of a person's own economic outcomes later in life. Large bodies of work have sought to quantify the extent to which these inherited or pre-determined characteristics shape people's life outcomes, and to compare results across societies or over time, including the literatures on intergenerational mobility, inequality of opportunity (IOp), and sibling correlations.

This paper contributes to those literatures in two ways. First, we note that most of these approaches rely on using observed inherited characteristics (often termed 'circumstances') to *predict* future outcomes – hereafter incomes, for simplicity. We suggest a simple general framework for the measurement of inherited inequality which relies on comparisons of inequality in observed and predicted income distributions and show that a wide range of measures in current use are special cases.

In this general framework, we define the hypothetical situation in which there is no inherited inequality as one in which inherited circumstances are *not* predictive of outcomes later in life – that is, current-generation income (y) is distributed independently from those circumstances ($\boldsymbol{c}$): $F(y|\boldsymbol{c}) = F(y), \forall \boldsymbol{c}$. Although this independence condition requires that the full conditional distributions of income be identical across groups of people that share the same circumstances, most commonly used approaches require only a weaker condition on conditional means: $E(y|\boldsymbol{c}) = E(y)$ for all $\boldsymbol{c}$. This second condition is implied by – but does not imply – the stronger independence condition.

Our second and main contribution is therefore to propose a new approach to measuring inherited inequality that captures the extent to which circumstances predict full conditional distributions – rather than just averages – for different population subgroups, and that does so in a statistically efficient manner. Given the central role of prediction in the general framework, we draw on new data-driven (supervised machine learning) techniques, which have been shown to be more accurate predictors than many standard econometric approaches used historically (see, e.g., Mullainathan and Spiess, 2017). Specifically, we propose to use *transformation trees*: a variant of regression trees proposed by Hothorn and Zeileis (2021) which generates a data-

driven partition of the population into groups with homogeneous inherited characteristics, while also predicting their conditional distribution functions.

This tool is ideally suited to estimating inequality of opportunity – especially what is known as *ex-post inequality of opportunity* – a concept that draws on a rich theoretical tradition in normative economics. In that approach, equal opportunity is defined as a situation in which all individuals who exert the same degree of effort or responsibility achieve the same outcomes, regardless of inherited circumstances (see, e.g., Roemer, 1993, 1998; Fleurbaey, 1994, 2008). Under some assumptions, the theory suggests that the appropriate degree of effort, once cleansed of the effects of circumstances, can be proxied by the relative position – the quantile – of the individual in the income distribution of the group of people that have the same inherited circumstances as she does – her "type". (Please see Roemer, 1998, for details).

Although this perspective – same efforts, same rewards – has considerable theoretical appeal (see, e.g., Fleurbaey and Peragine, 2013), it has hitherto faced serious empirical challenges which have limited its use in practice. Group-specific conditional distributions were used to detect inequality of opportunity by Lefranc, Pistolesi and Trannoy (2009), and empirical estimates of ex-post inequality of opportunity were first computed by Checchi and Peragine (2010). These pioneering approaches faced two main practical challenges: First, the partition of the population into types – population subgroups sharing the same circumstances – was *ad hoc* and therefore unlikely to adequately balance the model selection trade-off between (downward) omitted circumstance biases and (upward) overfitting biases (see Section 3 below for details). Second (and relatedly), predicting full conditional distribution functions for each type in data-scarce settings – such as household surveys – requires considerable parsimony both in defining types and in selecting quantiles, leading to coarseness in both contexts.

Our transformation tree-based approach can significantly alleviate both these challenges. The algorithm is designed to select partitions optimally – in a well-defined statistical sense – given the available data: it trades off the upward and downward biases so as to maximize a weighted sum of log likelihood functions. (See Section 3 below.) In addition, by using Bernstein polynomials to fit parametric estimates of type conditional distributions, the method uses data more efficiently and leads to much finer quantile groupings than was possible in earlier approaches. We argue that this combination makes this new approach the state-of-the-art method to estimate ex-post inequality of opportunity.

That said, the attractiveness of the approach does not require adherence to the specific normative views embodied in the theoretical literature on inequality of opportunity. Our choice of method arises primarily from the objective of capturing departures from the strong statistical independence condition described above - $F(y|\boldsymbol{c}) = F(y)$ - rather than from the weaker condition on means. It is therefore suitable for any empirical exercise where the objective is to identify the heterogeneity of conditional distributions across latent groups. Our results can also be interpreted in the spirit of alternative inequality decompositions, in which the between-groups term is not independent of within-group inequality.[2]

We apply this approach to four countries over many years: China (every two years between 2010 and 2018), India (2005 and 2012), South Africa (2008, 2012, 2015, and 2017), and the United States (every two years between 1968 and 2018). These four countries include the world's two largest nations by population (India and China), as well as the world's two largest economies by GDP (US and China). South Africa is a significant developing country and arguably the world's most unequal nation. These are four large economies characterized by very different social structures and territorial features. Using biological sex, parental education, parental occupation, place of birth, and ethnicity/race/caste as circumstances, we construct transformation trees to estimate the full conditional distributions of circumstance-homogeneous groups (types) for each country and to compute summary measures of inherited inequality as ex-post inequality of opportunity. We find a substantial amount of inequality of opportunity across the four countries, ranging from 14 Gini points (or 36% of total inequality) in the United States to 50 points (or 81% of the total) in South Africa. China and India display intermediate levels of IOp, but the structure of the distribution of opportunities differs: China shows a more pronounced concentration of opportunity at the very top of the distribution, while India has a large share of the population with extremely limited access to opportunity.

The use of transformation trees also allows us to extend the analysis of inherited inequalities beyond the estimation of a single summary index. For example, the final partition generated by the tree – although primarily a means to estimating the share of inequality that is inherited – can also be informative of the social structure in different countries. In addition, the estimates of type-specific empirical cumulative distribution functions (ECDF) enable us to directly estimate the social maximand proposed in John Roemer's original theory (1993, 1998): the level of (moneymetric) opportunity accessible to the lower envelope of types (see below). We find that

[2] See Foster and Shneyerov (2000) and Ebert (2010) for discussions of why it might make sense to account for differences in the full distributions within groups – rather than just the means – when defining the between-group term of the decomposition.

this level of opportunity ranges from 30% to 70% of the average income in the population across our four countries, with China performing much better than India and South Africa, and the United States showing a significant catch-up of the worst-off groups relative to the country's average income over the last two decades.

Transformation trees can also be used to assess the relative importance of different circumstances in shaping the income distribution. By aggregating (bagging) hundreds of trees, we derive Shapley values that quantify the average relative importance of each ascriptive characteristic. Our results indicate that all circumstances play a meaningful role, with race and caste being especially influential in South Africa and India, respectively. Moreover, the "marginal effects" of each category within each circumstance are obtained from the predicted conditional distributions. These partial effects find high premia associated with being White in South Africa, being born in specific regions of China (e.g., Shanghai and Zhejiang), and reporting higher parental education in India. In the United States, the most negative marginal effect is associated with identifying as Black.

The paper proceeds as follows. The next section briefly describes a general framework for the estimation of inherited inequality, of which the most common approaches in the measurement of mobility and inequality of opportunity are shown to be special cases. Section 3 discusses the key model selection challenge faced by these methods and introduces our own approach to estimating inherited inequality using transformation trees as another special case within the same general framework.

Section 4 describes the data and Section 5 presents results. These results include not only summary estimates of inherited inequality in the four selected countries, but also several complementary statistical and visualization tools to help the reader understand the complexity of the phenomenon: (i) a schematic description of the population partition that reveals the most salient cleavages in each society (again, in a well-defined statistical sense); (ii) estimates of the conditional cumulative distribution functions by type; (iii) a Shapley-Shorrocks decomposition of the average predictive importance of individual circumstances in the overall decomposition; (iv) a calculation of the marginal influence of each individual characteristic in predicting opportunities; and (v) an estimate of the lower-envelope of the type quantile functions, which corresponds to the maximand in Roemer's (1998) original policy objective. This rich set of byproducts of the headline estimates of inherited inequality is another advantage of our proposed approach: taken together, this set of statistical tools enable a deeper understanding of inherited inequality based on survey data. Section 6 concludes.

2. Inherited inequality: a simple general framework

Consider a population of *N* individuals, indexed by $i \in \mathcal{N} = \{1, \dots, N\}$, each of whom is characterized by a current-generation outcome y_i, $y \in \mathbb{R}$, and a set of inherited characteristics, which we call circumstances (following Roemer, 1998). For individual *i*, these circumstances are represented by a k-dimensional vector $\boldsymbol{c_i}$. Let ***y*** denote the N-dimensional outcome (or income) vector with entries y_i, $i \in \mathcal{N}$, and **C** denote the $N \times k$ matrix with rows $\boldsymbol{c_i}$.

In general, many people may share the same vector of circumstances, so many of the rows of the matrix **C** may be identical**.** Without loss of generality, let the number of *distinct* rows of **C** be denoted by M, $M \leq N$. If a "type" is defined as a group of individuals who share identical circumstances, this means that there are M types. The population can then be exhaustively partitioned into a set of types, $T = \{\tau_1, \dots, \tau_m, \dots, \tau_M\}$, where $\tau_m := \{\forall i | \boldsymbol{c_i} = \boldsymbol{c_m}\}$. Let $C = \{c_1, \dots, c_m, \dots, c_M\}$ denote the corresponding set of circumstance vectors, and ***c*** denote the generic random vector in C. Let $T \in \mathbb{T}$, which denotes the set of all possible partitions of the set $\mathcal{N}$, and $C \in \mathbb{C}$, the corresponding set of all possible type circumstance vectors. Note that an exhaustive partition implies that $\bigcup_1^M \tau_m = \mathcal{N}$ and $\bigcap_1^M \tau_m = \emptyset$.

That's the basic setup. Let us now define the benchmark situation in which there is no inherited inequality as one in which ***y*** and ***C*** are stochastically independent, in the sense that there are no differences across the conditional income distributions of types:

$$F(y|\boldsymbol{c_l}) = F(y|\boldsymbol{c_m}), \forall \boldsymbol{c_l}, \boldsymbol{c_m} \in C \quad (1)$$

Given full stochastic independence, it is clear that if (1) holds, C has no predictive power over y. Conversely, if (1) does not hold, then the associations between **C** and ***y*** across the population imply that circumstances C have (some) predictive power over *y*. I.e., there exist non-constant prediction functions,

$$y = f(\boldsymbol{c}, \varepsilon), f \in \mathcal{F} \quad (2)$$

that outperform constant functions in predicting *y* out of sample. In (2), ε denotes a random variable that captures other influences on y and is the residual term in the prediction model, and $\mathcal{F}: \mathbb{C} \to \mathbb{R}$ denotes the set of possible prediction functions linking circumstances to outcomes.

Since the benchmark situation of zero inherited inequality is characterized by C having no predictive power over *y*, then it is natural to think of inherited inequality as the extent to which circumstances do, in fact, predict the outcome *y* in a particular society. In other words, given a prediction function $f \in \mathcal{F}$, absolute inherited inequality can be defined simply as $I_n^A(y, \boldsymbol{c}, f) = I(\hat{y})$, where $\hat{y} = \hat{f}(\boldsymbol{c})$. Absolute inherited inequality is simply inequality in the distribution of predicted incomes, when incomes are predicted by circumstances. One can also define relative measures of inherited inequality as $I_n^R(y, \boldsymbol{c}, f) = \frac{I(\hat{y})}{I(y)}$, where $\hat{y} = \hat{f}(\boldsymbol{c})$. Relative inherited inequality is the ratio – or potentially a monotonically increasing function of the ratio – of inequality in predicted incomes to inequality in observed incomes.

Indeed, it turns out that most methods for estimating the intergenerational transmission of advantage currently in use – including relative measures of intergenerational mobility and inequality of opportunity – revolve around estimating prediction models of the general form (2), using different functions in the set of possible functions $\mathcal{F}$, and then computing objects analogous to $I_n^A(y, \boldsymbol{c}, f)$ or $I_n^R(y, \boldsymbol{c}, f)$, often using different inequality indices *I(.)*.[3]

Special cases

Suppose, for example, that the only inherited characteristic that really matters is parental income, y_p. Then the vector of circumstances reduces to a scalar: $\boldsymbol{c} = y_p$. If, in addition, we choose a prediction function $f(\boldsymbol{c}) = f(y_p)$ of the form $y = e^{\alpha + \beta \log y_p + \varepsilon}$, which can be estimated through the standard Galtonian regression $\log y = \alpha + \beta \log y_p + \varepsilon$, then we are clearly in the world of intergenerational mobility measurement. See, e.g., Solon (1992) and Chetty et al. (2014) for classic references.

In that standard formulation, predicted incomes are given by $\hat{y} = e^{\hat{\alpha} + \hat{\beta} \log y_p}$. Although $\hat{\beta}$, an estimate of the intergenerational elasticity of income, is a common measure of mobility, another frequently used measure is the correlation coefficient between $\log y$ and $\log y_p$, which can be written as $\hat{\beta} \sqrt{\frac{var \log y_p}{var \log y}}$. But this is precisely a monotonic function of $\frac{I(\hat{y})}{I(y)}$ when the inequality index is the variance of logarithms.

Similarly, absolute and relative estimates of inequality of opportunity can also be written as examples of $I_n^A(y, \boldsymbol{c}, f)$ or $I_n^R(y, \boldsymbol{c}, f)$, but typically with circumstance vectors with *k > 1*. One

[3] See Ferreira and Brunori (2024) for a more in-depth discussion of the concept of inherited inequality and of the relationship between intergenerational mobility, inequality of opportunity, and this broader concept.

frequently used (ex-ante) relative measure is $\frac{I(\hat{y}_{EA})}{I(y)}$, where $\hat{y}_{EA} = \hat{f}_{EA}(c) = e^{\hat{\alpha}+C\hat{\gamma}}$, estimated from an OLS regression of the form $\log y = \alpha + C\gamma + \varepsilon$. (See, e.g., Ferreira and Gignoux, 2011, or Niehues and Peichl (2014) for a fixed-effects specification for panel data.) A non-parametric analogue proposed by Checchi and Peragine (2010) uses a prediction function that simply computes type means for each cell in the partition $T = \{\tau_1, \dots, \tau_m, \dots, \tau_M\}$. Its prediction function is therefore:

$$\hat{y}_{EA(n)} = \hat{f}_{EA(n)}(c) = \int_0^1 y dF(y|\boldsymbol{c_m}), \forall m \tag{3}$$

Equation (3) simply yields the conditional means for all those who share the same vector of circumstances $\boldsymbol{c}$. So $I\left(\hat{y}_{EA(n)}\right)$ is computed over the *smoothed distribution* where individual incomes are replaced by the average incomes of individuals who share the same vector of circumstances – that is, individuals in the same type.[4] It is analogous to the OLS regression-based estimate, but without imposing a linearity assumption.

Ex-post measures of inequality of opportunity are also special cases of the inherited inequality framework. Checchi and Peragine (2010), for instance, propose to estimate ex-post IOp by aggregating income differences across the quantiles of the conditional distributions, while abstracting from level differences between tranches.[5] Denoting the overall mean income, $E(y)$, by μ and the mean income for a given quantile *q* across types by μ_q,[6] their prediction function is given by:

$$\hat{y}_{EP} = \hat{f}_{EP}(\boldsymbol{c}) = \frac{\mu}{\mu_q} F^{-1}(q|\boldsymbol{c}) \tag{4}$$

Denote the income at quantile *q* of the conditional distribution on circumstances $\boldsymbol{c}$, $F^{-1}(q|\boldsymbol{c})$, by y_{qc}. Then their absolute IOp measure $I(\hat{y}_{EP}) = \mathrm{I}\left(\frac{\mu}{\mu_q} y_{qc}\right)$ is simply $I_n^A(y, \boldsymbol{c}, f_{EP})$. Similarly, the relative version is $I_n^R(y, \boldsymbol{c}, f_{EP}) = \frac{I(\hat{y}_{EP})}{I(y)}$. In other words, Checchi and Peragine (2010) compute inequality in predicted incomes by dividing the income of each observation y_{qc} by the average, across types, of all incomes occupying that same quantile *q* in their own conditional distributions and then computing inequality across the resulting ratios. Relative IOp is, once again, the ratio of inequality in predicted incomes to observed inequality.

[4] See Foster and Shneyerov (2000) for a definition of the smoothed distribution.

[5] A 'tranche' denotes those individuals exerting the same degree of relative effort. Under Roemer's (1998) identifying assumption, a tranche is therefore given by the set $\Upsilon_q := \{\forall i | F(y_{ic}|c) = q, \forall c\}$.

[6] $\mu_q = \frac{1}{M}\sum_{m=1}^{M} F^{-1}(q|\boldsymbol{c_m})$

3 Estimating IOp using Transformation Trees

The model selection problem

Empirical applications of all three variants of the prediction problem described above – intergenerational mobility, ex-ante IOp, and ex-post IOp – may suffer from a variety of challenges, including data availability, measurement error (particularly in variables such as parental income or occupation), small sample sizes, etc. More fundamentally, though, they suffer from a model selection problem in the presence of two competing biases. This is particularly true in the IOp literature, where many inherited circumstance variables are typically used in the analysis, often with multiple categories each.

The first bias arises from the partial observability of circumstances. It is rather common for data sources that contain information about individual outcomes to also contain various variables describing inherited circumstances such as sex, race, and socioeconomic background. But the set of available information is almost certainly a strict subset of all background circumstances which society does not wish to hold individuals responsible for. Omission of the unobserved circumstances, or indeed of interactions between categories of variables one does observe, tends to bias estimates of IOp downwards (Ferreira and Gignoux, 2011; Roemer and Trannoy, 2016).[7]

On the other hand, a second source of bias arises from the classic overfitting problem, whereby saturating the model with a large number of independent variables and their multiple interactions leads to an upward bias in the estimates of goodness of fit. This is a problem for both parametric and non-parametric methods. In a non-parametric setting, the problem manifests as exploding sampling variation around cell means as cell sizes decline below a certain level. This problem introduces noise in the predictions. This noise has the effect of inflating the estimation of explained variance, introducing an upward bias in the measurement of the variation predicted by circumstances (Chakravarty and Eichhorn, 1994), that is IOp, and an attenuation bias in the case in which predictions are used as regressors, that is when circumstances are used to predict parental income to estimate intergenerational mobility adopting a two-sample-two-stage approach (Bloise et al., 2021).[8]

[7] This bias is also a concern for estimates of intergenerational mobility if they are to be interpreted as measures of inherited inequality – except in the unlikely event that parental income is a sufficient statistic for all circumstances.

[8] Note that these biases are connected to the bias-variance trade-off central to supervised machine learning. Assuming that our objective is to estimate to what extent observable circumstances are

Although this problem was recognized from the outset, most of the early literature failed to address the trade-off between the two kinds of bias in a systematic way.[9] The early studies that proposed either parametric or non-parametric methods to estimate IOp relied on ad-hoc specifications, either of the regression model or of the type partition. Yet, changing the number of regressors in such a model can substantially affect the final estimates of IOp.

Obtaining a meaningful estimate of $I(\hat{y})/I(y)$ therefore depends crucially on selecting the 'right' model for the prediction function $y = f(\boldsymbol{c}, \varepsilon)$. Of course, what the 'right' model is depends on the nature and purpose of the exercise. If one is estimating a structural model, guidance from the theory being tested is indispensable, and econometric methods suitable for the estimation of structural parameters should be used. However, when the model is used for prediction, as is the case here, it may very well be that machine-learning methods from data science perform better. See Mullainathan and Spiess (2017) for an excellent discussion of the role of machine learning in economics and its advantages in prediction problems.

Indeed, machine learning methods have recently been applied to the measurement of *ex-ante* (but not *ex-post*) inequality of opportunity. In particular, Brunori, Hufe, and Mahler (2023) have used conditional inference trees and random forests (CITF), introduced by Hothorn, Hornik, and Zeileis (2006), to estimate inequality of opportunity in 31 European countries.[10] CITF partition a regressor space with the aim of predicting a dependent variable via the estimation of subgroup means. This feature makes them ideally suited to choosing a type-partition in an ex-ante framework, because each binary split is chosen by identifying the most significant difference between means in the two resulting nodes. Since the ex-ante approach to IOp involves computing inequality among type means, such an algorithm is a conceptually attractive approach to selecting the partition and estimating Equation (3).

predictive of outcomes later in life: choosing a model that underfits the data, that is minimizing the variance of the model but introducing a large bias, would result in an underestimate of inherited inequality. Conversely, minimizing the bias by fitting a very complex model would result in a large variance that, in expectation, will exaggerate the share of inequality that can be correctly predicted by observing innate circumstances. Supervised machine learning methods can therefore be used to trade-off the two sources of errors and to obtain the most accurate estimate of inherited inequality (Brunori, Peragine, Serlenga, 2019).

[9] Ferreira and Gignoux (2011), for example, note that "As sampling variance is high for cells containing few observations, estimated between-type inequality may become inflated, thereby inducing an overestimation of inequality of opportunity." (p.640). However, their proposed solution is to exercise "considerable parsimony in the partitioning of the population…" (p.642). They selected categories arbitrarily and restricted the number of types to a maximum of 108, but there was no sense in which that particular number represented an optimal choice between the downward bias from omitting certain interactions between the variables and categories, and the upward bias from including too many.

[10] See also Li Donni, Rodriguez and Dias (2015) who use finite mixture models to define types.

But precisely because conditional inference trees focus on differences between means, they are not well suited to assessing deviations from the stricter criterion of equal CDFs (Equation 1), whether one interprets such equality as ex-post equality of opportunity or simply as the absence of inherited inequality. An alternative data-driven approach is needed and, in what follows, we propose the use of one such approach, namely transformation trees.

Transformation trees and ex-post IOp

As noted in Section 2, the ex-post approach to inequality of opportunity consists of measuring inequality across the types' conditional distributions functions at each quantile, and then appropriately aggregating across quantiles. The key ingredient for the approach, therefore, is to estimate the income level at quantile q in type τ_m, that is: the conditional quantile function $y_{qc_m} = F^{-1}(q|\boldsymbol{c_m})$, for all $\boldsymbol{m}$. When data on the joint distribution {**y, C**} is not observed for the full population, estimating these conditional quantile – or their inverse, distribution – functions from a sample notionally involves two steps.

First, an optimal type partition $C \in \mathbb{C}$ needs to be selected, trading off the downward bias that arises from combining sub-types into types against the upward bias from overfitting that arises from an excessively fine partition, (i.e., by subdividing types into sub-types). See Brunori, Peragine, and Serlenga (2019). Second, given a partition $C \in \mathbb{C}$, the conditional quantile functions must be estimated, either parametrically or non-parametrically. Once that has been done, the resulting estimates $\{\tilde{y}_{qc}\}$ can be used to compute quantile-specific inequality levels (across types), which are then suitably aggregated across quantiles.

Previous attempts to compute ex-post IOp (e.g., Checchi and Peragine, 2010) have typically suffered from two shortcomings. First, the partition $C \in \mathbb{C}$ was chosen arbitrarily. Second, quantiles were computed at a highly aggregated level, e.g., quartiles or deciles, so as to ensure that there were sufficient observations in each quantile (or "tranche") for a meaningful computation of inequality across types to take place. Indeed, the fact that the ex-post approach to IOp requires information on the entire conditional distribution $F\left(y_{qc}|c\right)$, rather than merely the mean μ_c of that distribution for each type, makes it more data-intensive and has been one of the reasons why the ex-ante approach has dominated empirical applications.

These combined requirements – to choose an optimal type-partition given the available dataset and to estimate conditional distribution functions for each of those types in a data scarce

environment – make this problem well-suited to a new variety of tree-based estimator, recently developed by Hothorn and Zeileis (2021). This estimator, known as a transformation tree (TrT), was specifically designed to estimate conditional distributions for terminal nodes of trees.

TrT relies on the assumption that there exist "good enough" parametric approximations to $F(y_{qc}|c)$. In the limit, they assume that there exist parameters $\theta \in \Theta$ such that:

$$F(y_{qc}|c) \cong F\left(\tilde{y}_{qc}, \theta(c)\right), \theta: \mathbb{C} \to \Theta \tag{5}$$

$\theta(c)$ is known as the conditional parameter function, which maps from the space of all possible circumstance vectors on to the space of possible distributional parameters. Under this assumption, the problem of estimating conditional distribution functions for types in the optimal partition, and hence $\{\tilde{y}_{qc}\}$, reduces to the problem of selecting the optimal parameter estimates, $\hat{\theta}$, given the data {y, C}. TrT uses an adaptive local likelihood maximization approach for that purpose. Specifically, it selects $\hat{\theta}$ as:

$$\hat{\theta}^N(c) = \arg\max_{\theta \epsilon \Theta} \sum_{i=1}^{N} w_i(c) \ell_i(\theta) \tag{6}$$

where $i \in \{1, \dots, N\}$ denotes each observation in the data set and $\ell_i(\theta)$ denotes the log-likelihood contribution of i, when the parameters are given by θ. The recursive binary splitting process that creates a transformation tree is implemented by choosing weights:

$$w_i(c) = \sum_{b=1}^{B} I(c \in \mathcal{B}_b \wedge c_i \in \mathcal{B}_b) \tag{7}$$

The indicator function takes the value 1 when observation *i* is sufficiently "close" to c, so the weights in (7) simply count the number of observations in each bin $\mathcal{B}_b$. At the terminal nodes, $\mathcal{B}_b$ corresponds to a type, so the maximization process in (6)-(7) allocates each observation to a type and sums the local likelihood functions across types. The type partition and the parameter vector θ are chosen so as to maximize that weighted sum of likelihoods. That is, given the available data {**y, C**} and the recursive splitting approach to weights, the likeliest set of types and income distributions conditional on type is that given by $F\left(\tilde{y}_{qc}, \hat{\theta}^N(c)\right)$. So, our prediction function under this method is given by:

$$\hat{y}_T = \hat{f}_T(c) = \frac{\mu}{\mu_q} \tilde{y}_{qc} \quad \text{where} \quad \tilde{y}_{qc} = F^{-1}\left(q, \hat{\theta}^N(c)\right) \tag{8}$$

The Transformation Tree estimate of absolute ex-post inequality of opportunity is then $I_n^A(y, \boldsymbol{c}, f_T) = I(\hat{y}_T)$, while the relative measure is analogously given by:

$$I_n^R(y, \boldsymbol{c}, f_T) = \frac{I(\hat{y}_T)}{I(y)} \tag{9}$$

Details of how the likelihood maximization is implemented (using Bernstein polynomials to fit the conditional distribution functions at each node) are given in Appendix 1A. In practice, the process can be summarized by the following seven-step algorithm:

1. set a confidence level $(1 - \alpha)$ and a minimum size for final nodes (n_{min});
2. choose a polynomial order (M);
3. estimate the unconditional distribution function with a Bernstein polynomial of order M;
4. test the null hypothesis of polynomial parameter stability for all possible partitions based on each element of the circumstance vector $\boldsymbol{c}$, and store $p - values$.
5. If, for all $\boldsymbol{c}$ and each possible partition, either the Bonferroni-adjusted $p - value > \alpha$ or $n_c < n_{min}$, exit the algorithm;
6. otherwise, choose the variable and the splitting value producing the smallest $p - value$ to obtain two subgroups.
7. Repeat step 4-6 for the resulting subgroups, until exiting everywhere.

In our application, we follow statistical convention and set α to 0.01 and n_{min} to 1% of the sample size. Then, we choose M, the order of the Bernstein Polynomial. The selection of M is not as simple as that of α, because how well a polynomial of a certain order interpolates the distribution is intrinsically data dependent. An order too small might result in a poor approximation of the distribution, while too high an order would translate into a loss of degrees of freedom and high computational costs.[11]

To find an appropriate order, we tune the algorithm by estimating the out-of-sample log-likelihood, after a 5-fold cross validation, for several order values of the Bernstein Polynomial (ranging between 2 and 10). We select the lowest order for which the relative improvement of

[11] The confidence level $(1 - \alpha)$ and the order of the polynomial (M) interact in determining the depth of the tree, and thus the complexity of the final partition. For a given sample size, fixing a higher polynomial order implies using more degrees of freedom in each test, leading to a lower probability of rejecting the null hypothesis of equal distributions. Consequently, the resulting partition is more parsimonious. Similarly, for a given polynomial order and confidence level, a larger sample size results in a more detailed partition of types and likely a higher level of between-group inequality. In the empirical application, we recommend verifying the sensitivity to sample size, as we do below, when drawing conclusions about estimates.

the log-likelihood that would be obtained by estimating an additional parameter is smaller than 0.1%.[12] In step 3, an unconditional CDF for our sample is thus estimated with a Bernstein polynomial of order 8.

The key step is then step 4, where the M-fluctuation test is performed to detect instability of the parameters in the conditional distribution functions across potential types (see Appendix 1A). To intuitively illustrate this key test, Appendix 1B provides a simple example of the procedure, using made-up data. Further details can be found in Hothorn and Zeileis (2021) and Kopf, Augustin, and Strobl (2013).

After following steps 4-7 we obtain an estimated Transformation Tree and, from that tree, a number of outputs that are described in Section 5. Before presenting those results, we briefly describe our datasets in Section 4.

4. Data

We apply this method to four countries: China, India, the United States, and South Africa. These countries were selected for their relative importance in different parts of the global economy and because they represent substantial heterogeneity in both the structure of inequality and the availability of data. For all countries, our samples comprise all adult individuals (aged over 18) observed in nationally representative surveys. The outcome of interest is equivalized disposable household income, using the square root equivalence scale (Buhmann et al., 1988; OECD, 2013). We also age-adjust the income to account, at least in part, for life-cycle dynamics. The adjustment consists of regressing our income variable on age and age squared, and using the sum of the constant and the individual residual from that regression as the adjusted variable (see, e.g., Palomino et al., 2022).

We select a set of circumstance variables available in each survey to estimate inherited inequality. Balancing the desirability both of including the most relevant circumstances, and to preserve a certain degree of comparability across countries, we selected the following variables: biological sex[13], place of birth, mother's and father's education, mother's and father's

[12] Note that the order of the polynomial does not turn out to be a key determinant of the estimated level of inequality of opportunity. Figure B1 in Appendix 2 shows, for the case of South Africa, the stability of the ex-post IOp when the selected Bernstein polynomial order varies.

[13] Given the choice of household income as the outcome of interest, the role of sex among the circumstances will necessarily be limited, since intra-household inequalities are ignored. The relative unimportance of sex in the analysis that follows should therefore be treated cautiously. It does not imply that these countries provide equal opportunities to men and women in other dimensions. On the other

occupation, and an “ethnicity” variable, which includes religion and a coarse caste classification in India. Even for this relatively limited set of inherited characteristics, only two countries (China and the US) contain information on all seven. Birthplace is missing in South Africa and mother’s and father’s occupation are missing in India, and comparisons should be interpreted accordingly. Detailed descriptive statistics for each country are available upon request.

For China, we use the China Family Panel Studies (CFPS), carried out by the Institute of Social Science Survey (ISSS) of Peking University every two years from 2010 to 2018 with a sample size of individuals with complete information ranging between 16,000 and 21,000 observations. CFPS has been already used to study aspects of the intergenerational persistence of income and inequality of opportunity in China (e.g. Fan, Yi, and Zhang, 2021; Emran et al, 2023). The dataset contains data on important inherited characteristics. We include sex; ethnicity, classified into 11 categories ("Han", "Mongol", "Hui", "Tibetan", "Miao", "Yi", "Zhuang", "Bouyei", "Korean", "Manchu", "Other"); 24 birth area categories[14]; mother’s and father’s education in eight categories[15]; and mother’s and father’s occupation (10 ISCO categories plus one category for unemployment).

For India, we use the India Human Development Survey (IHDS), conducted by researchers from the University of Maryland and the National Council of Applied Economic Research (NCAER). These are large representative samples of the 2005 and 2012 populations, each containing over 100,000 observations, resulting in analysis samples of 78,000 and 98,000 complete observations, respectively. IHDS has previously been used to study intergenerational mobility and inequality of opportunity in India (e.g. Asher, Novosad and Rafkin, 2024; Kundu and Lefranc, 2020). Circumstance variables included are: six castes/religions (“Forward caste”, “Other

hand, omitting sex as a circumstance variable would have caused us to miss some non-negligible consequences of differences in household composition across countries.

[14] Namely "Hebei", "Shanxi", "Liaoning", "Jilin", "Heilongjiang", "Shanghai", "Jiangsu", "Zhejiang", "Anhui", "Fujian", "Jiangxi", "Shandong", "Henan", "Hubei", "Hunan", "Guangdong", "Guangxi Zhuang Autonomous Region", "Sichuan", "Guizhou", "Yunnan", "Shaanxi", and "Gansu”, plus “not available” and “other”.

[15] Namely "Illiterate/Semi-literate", "Primary school", "Junior high school", "Senior high school/secondary school/technical school/vocational senior school", "3-year college", "4-year college/Bachelor’s degree", "Master’s degree", "Doctoral degree")

Backward castes", "Dalit", "Adivasi", "Muslim", "Christian, Sikh, and Jain"); 23 birth areas[16]; and mother's and father's education in seven categories.[17]

For the United States, we employ the Panel Study of Income Dynamics (PSID), a well-known data source for researchers interested in the intergenerational transmission of income and status, as it is the longest-running longitudinal household survey in the world. The survey started in 1968 and is currently managed by the University of Michigan. It has been used in many studies of intergenerational mobility and inequality of opportunity in the US (e.g. Mazumder, 2018; Pistolesi, 2009). We use every other wave (even years) between 1968 and 2018.

The PSID includes data related to employment, income, wealth, expenditures, and a number of other background characteristics that could be used as circumstance variables. However, to preserve a modicum of comparability with the other three countries, we restrict inclusion to the seven variables listed above. Ethnicity is described in six possible categories ("White", "Black", "American Indian, Aleut, or Eskimo", "Asian or Pacific Islander", "Hispanic", "Other"). The area of birth is also classified into six categories ("Northeast", "Northwest", "South", "West", "Alaska & Hawaii", "foreign country"). Mother's and father's occupation are coded in a variable based on ISCO codes (High: includes ISCO 1, 2, and 3; Medium: includes ISCO 4, 5, and 6; and Low: ISCO 7, 8, 9, and 0).[18] Finally, mother's and father's education are recoded in eight categories ("0-5 grades", "6-8 grades", "9-11 grades", "high school", "12 grades and non-academic training", "college, no degree", "college degree", "advanced college").

For South Africa, we rely on the National Income Dynamics Study (NIDS 1-5) survey, carried out by the Southern Africa Labour and Development Research Unit (SALDRU). NIDS is a longitudinal survey, collected in 2008, 2010/11, 2012, 2014/5 and 2017. It is an interesting dataset for studying the inheritance of inequality because it is a reliable and extensive source of information about incomes and circumstances for arguably the world's most unequal country. Inequality of opportunity and mobility have already been analysed in South Africa using the NIDS, e.g. by Piraino (2015) and Brunori, Ferreira, and Peragine (2021). The circumstance variables that we include in the analysis are: ethnicity ("African", "Asian or Indian", "coloured", and "white"),

[16] Namely "Jammu and Kashmir", "Himachal Pradesh", "Punjab", "Another State", "Uttarakhand", "Haryana", "Delhi", "Rajasthan", "Uttar Pradesh", "Bihar", "Overseas", "Northeast", "West Bengal", "Jharkhand", "Orissa", "Chhattisgarh", "Madhya Pradesh", "Gujarat", "Maharashtra", "Andhra Pradesh", "Karnataka", "Kerala", "Tamil Nadu".

[17] Namely, "none", "incomplete primary", "complete primary", "incomplete secondary", "completed secondary", "higher secondary", and "post-secondary or higher".

[18] ISCO codes are grouped in this way following Hufe et al. (2022). Results are very similar if each ISCO occupational classification is entered separately.

fathers' and mothers' education (13 categories, ranging from "not educated" to "Grade 12 or more") and fathers' and mothers' occupation (11 categories, 10 associated to the 1-Digit ISCO and one extra including other categories, such as out of the labour force, deceased or other unclassified occupations).[19]

In all four countries, the final sample used for our analysis includes only complete observations, in the sense that information is not missing for any of the outcome or circumstance variables described above. Of course, item non-response can be a serious issue in data containing retrospective information on respondent's parents. We are able to alleviate this problem somewhat by matching individuals across waves and by filling some missing information with answers reported by the same individual in other waves. Nonetheless, the process of dropping observations with incomplete information does reduce our sample sizes, and may do so in a selected way. While we cannot rule out sample selection, for each country and wave we examined the pattern of missing information and calculated the difference between average income and its inequality, both including and excluding observations with missing circumstances. Results do not seem particularly alarming and are available upon request.

Table 1 shows some basic descriptive income statistics for the analysis samples of the most recent survey used. The four countries differ both in their level of development and in the nature of their inequality. The United States combines high income levels with a moderate level of inequality (at least relative to this group of countries). In contrast, South Africa, despite mid-level average incomes, is marked by high inequality, with a Gini coefficient exceeding 0.6. India and China, while showing similar overall levels of inequality, differ in the structure of that inequality. In China, income is more concentrated at the top: the top 1% accounts for 14% of total income, compared to 12% in India. However, India faces a more pressing issue at the lower end of the distribution. The bottom 40% of the population in India receives just 9% of total income, a figure not only lower than China's 13%, but also below that of South Africa. Table A1 in Appendix 2 contains the same set of summary descriptive statistics for all earlier waves.

[19] Note that the question refers to current (or last recorded) occupation of the parents. We exploit the panel structure of NIDS and look at information about circumstances reported by the same individuals in previous waves. Whenever a circumstance variable that is missing is available for the same individual in previous waves, we use the oldest available value of the circumstance, on the ground that it was reported closest to when the respondent was young.

Table 1: Descriptive Income Statistics for the most recent waves

Country	Year	Mean	Gini	MLD	Top 1%	Top 10%	Bottom 40%
China	2018	9,998	0.497	0.459	0.137	0.400	0.129
India	2012	3,196	0.527	0.518	0.123	0.439	0.089
South Africa	2017	13,429	0.610	0.690	0.157	0.444	0.113
USA	2018	48,420	0.389	0.301	0.063	0.290	0.179

Note: Income units are in 2017 US dollars at PPP exchange rates. MLD stands for Mean Log Deviation. The three columns on the left represent the share of income received by the Top 1%, the Top 10%, and the Bottom 40% in the income distribution. Source: CFPS (2018), IHDS (2012), NIDS (2017), and PSID (2018).

5. Results: Inequality of Opportunity in China, India, South Africa and United States

Transformation trees and Type-specific Cumulative Distribution Functions

Applying the algorithm outlined in Section 3 to solve Equations 14-15, with the key stopping rule parameters set to α = 0.01 and n_{min} = 1% of the sample (as described), yields the transformation trees shown as Figure B2 in Appendix 2. These stopping rules are quite conservative and the nodes of those trees are used for our IOp estimates described below. However, as the full trees are fairly deep and complex, Figure 1 below shows the trees pruned to a maximum depth of four levels, to make the output more readable.[20]

Consider first the pruned tree for the United States, in Figure 1a. The splitting process generated by the algorithm should be read from left to right. The first split divides the population into a group consisting of just under one-third of the sample, whose fathers had at least some college education (Node 15), and the rest of the sample (Node 2). As we move to the right, other circumstances further partition the population following the algorithm, until the final nodes – types – are reached. An interesting symmetry emerges straight away: ethnicity appears as the second splitting circumstance in both subtrees, producing identical splits (into nodes 16 and 19 for those with more educated parents, and 3 and 8 for the remainder). In both cases, the categories 'Black' and 'American Indian, Aleut, Eskimo' cluster together in the poorer sub-branch, while all other groups fall into the more affluent sub-branch. Subsequent splits are determined by the father's education and occupation, the mother's education, birth area, and sex. At level four, where the tree is trimmed, we find a partition into ten types, and the Figure shows the parametrically estimated density function for each of them, as well as indicating the

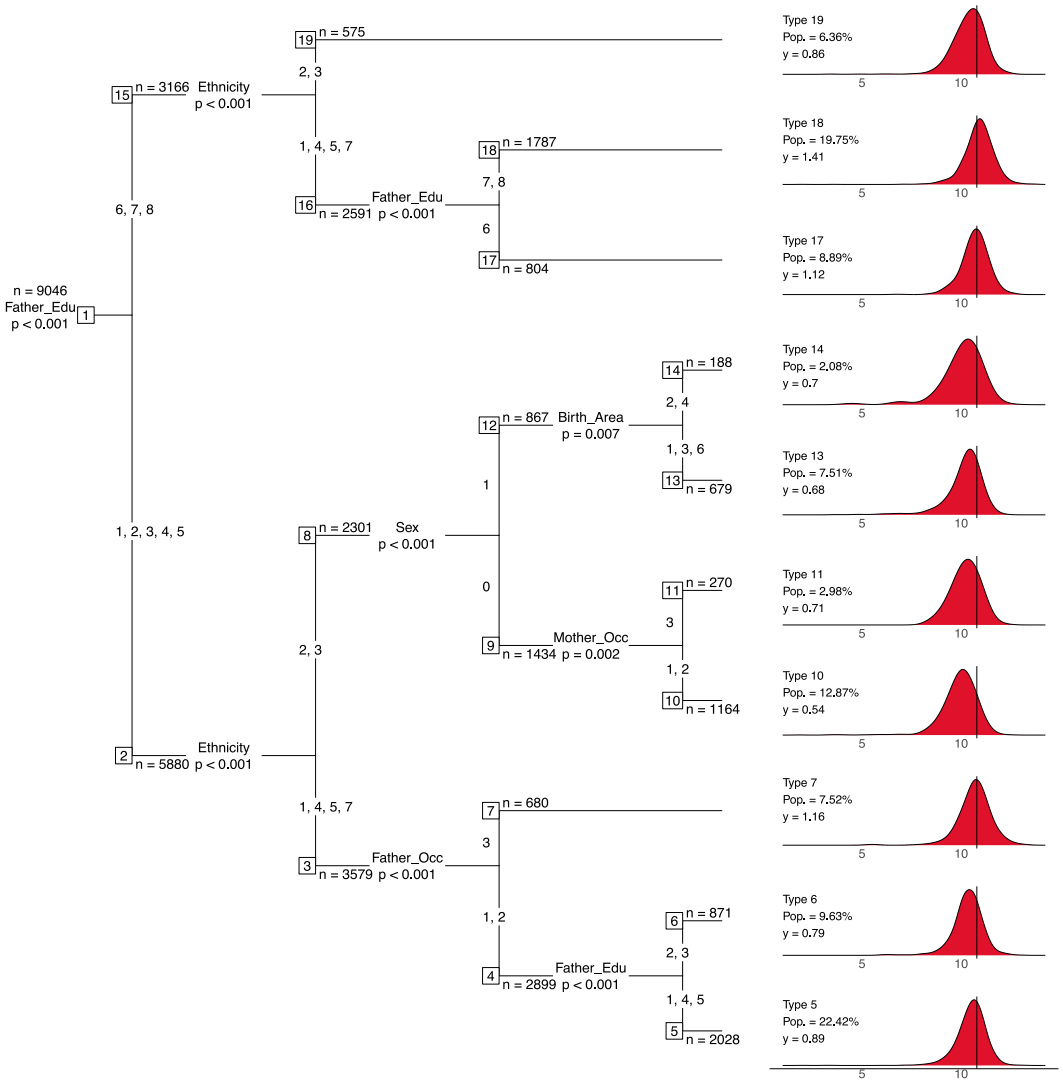
[20] This is implemented through an additional stopping rule that prevents any observation from being involved in more than four splits.

population share accounted for by each type and its mean income as a multiple or share of the overall mean.[21]

In terms of the model selection challenge discussed in Section 3, the algorithm partitioned the population into these ten groups (and fit CDFs to them) so as to maximize the likelihood of fitting the data, under the restrictions $f \in \mathcal{F}_\tau$, with $\mathcal{F}_\tau$ being the class of recursive binary TrT estimators. The partition can be thought of as the product (or interactions) of various dummy variables defined over the circumstances. Type 10, for example, which is the poorest type in terms of expected income (54% of the national average) and comprises 13% of the population, consists of black or indigenous women whose fathers never went to college and whose mothers worked in specific occupations, corresponding to the interaction of dummy variables $x_{10} = \mathbf{1}_{race=\text{"black"} or \text{ "American Indian,Aleut,Eskimo "}} \times \mathbf{1}_{father\ education=below\ college} \times \mathbf{1}_{sex=\text{"}female\text{"}} \times \mathbf{1}_{mother\ occupation=\text{"}ISCO\ categories\ 1\ or\ 2\text{"}}$. Type 18, which is the richest type and includes 20% of the sample, consists of White, Hispanic and Asian people whose fathers are college graduates. corresponding to the interaction $x_{18} = \mathbf{1}_{race=White,\ Hispanic\ and\ Asian} \times \mathbf{1}_{father\ education=College\ gradueted\ or\ more}$. And so on.

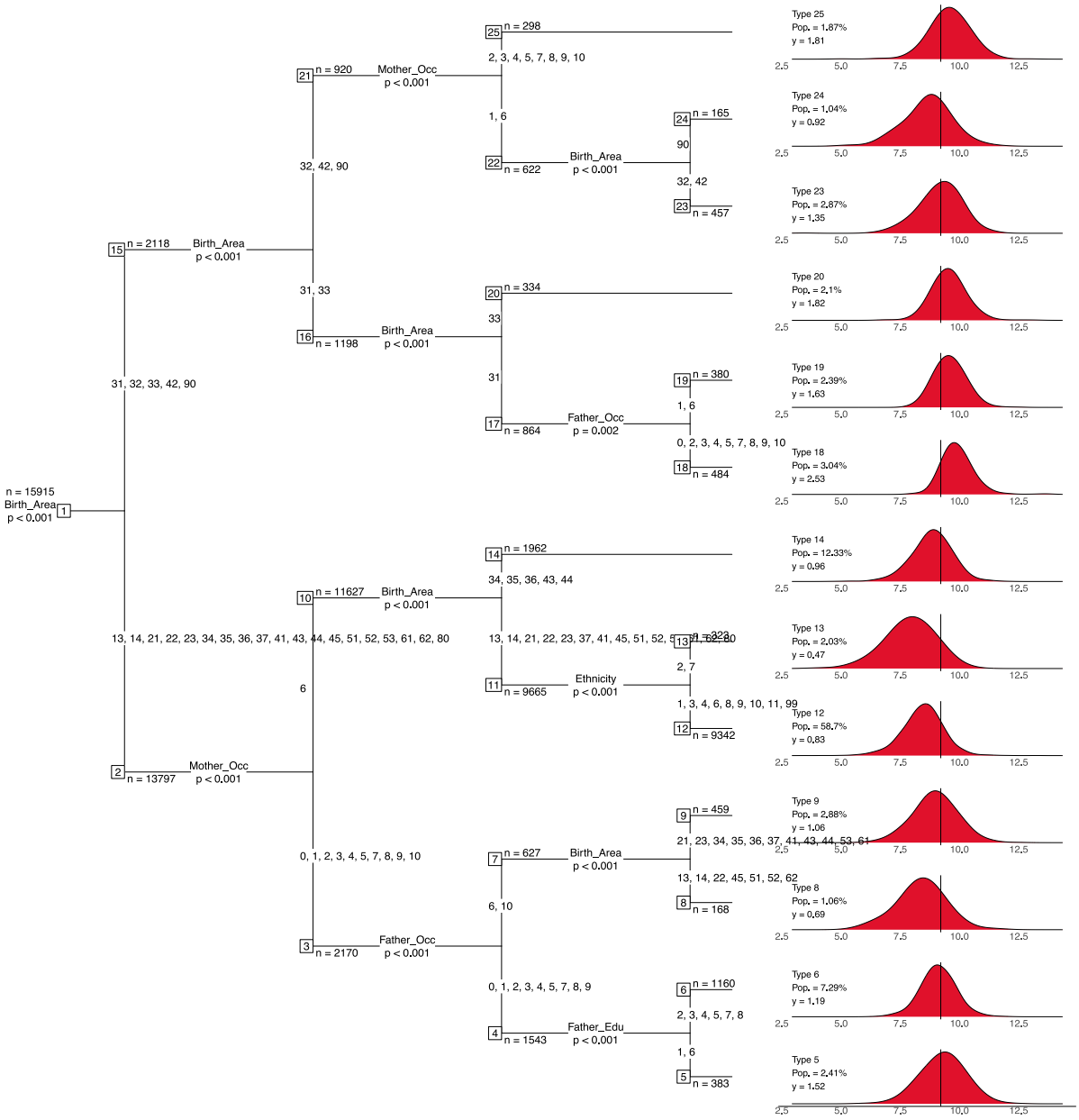
[21] Although we use income in levels to compute all our measures, we plot the density of log incomes for ease of visualization.

Figure 1a: Transformation Tree for the United States



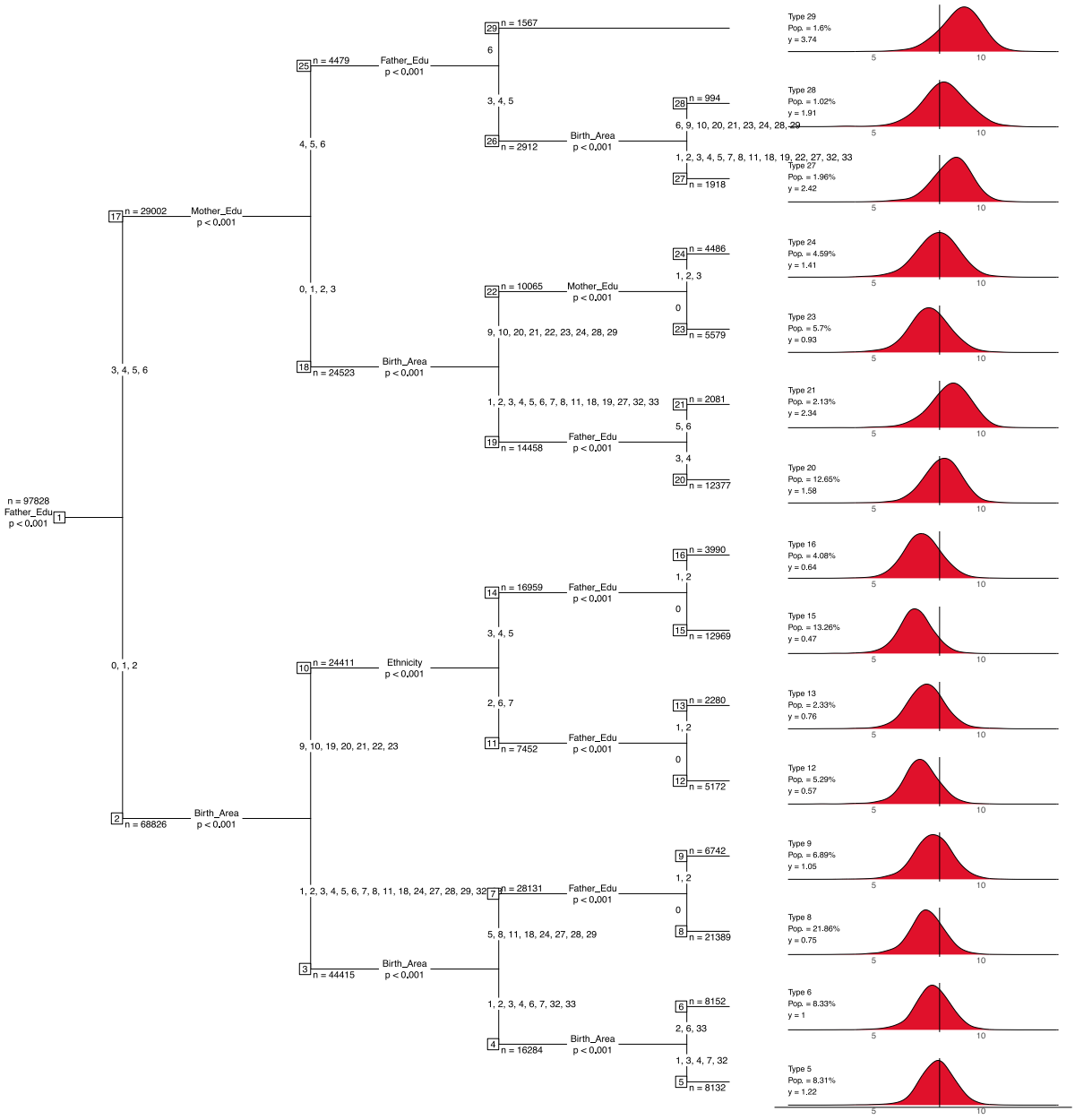
Note: Splitting nodes show their sample size and the p-value associated with the split. Circumstance categories are Gender (0 Male, 1 Female), Ethnicity (1 White, 2 Black, 3 American Indian/Aleut/Eskimo, 4 Asian/Pacific Islander, 5 Hispanic, 7 Other), Region of upbringing (1 Northeast, 2 North Central, 3 South, 4 West, 5 Alaska/Hawaii, 6 Foreign country), Parents' education (1 0–5 Grades, 2 6–8 Grades, 3 9–11 Grades, 4 High school, 5 12+ Grades + non-academic training, 6 Some college, 7 College degree, 8 Advanced college degree), and Parents' occupation (ISCO) (1 Basic, 2 Middle, 3 High). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: Own elaboration from the PSID (2018).

Figure 1b: Transformation Tree for China



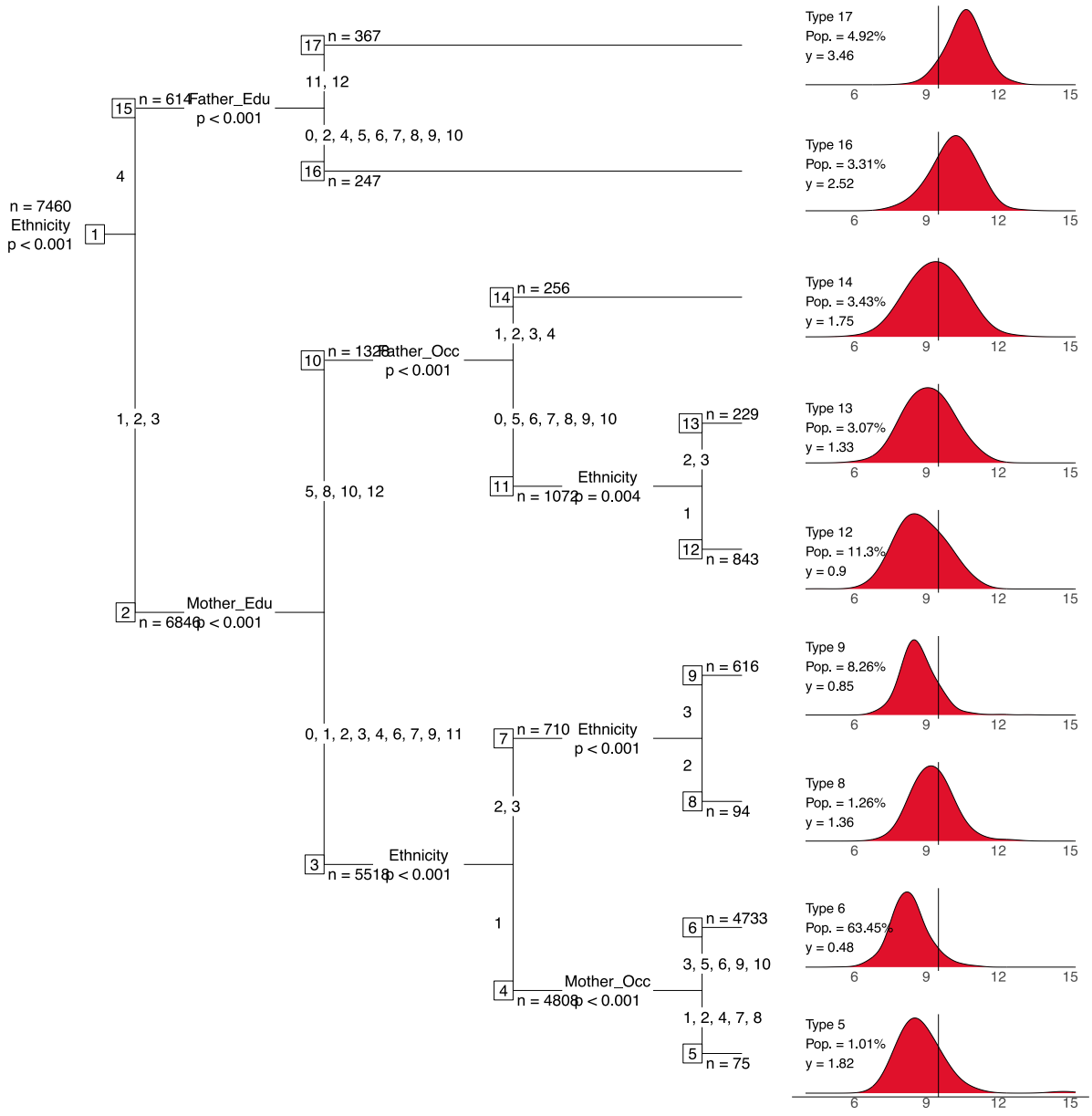
Note: Splitting nodes show their sample size and the p-value associated with the split. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (1 Han, 2 Mongol, 3 Hui, 4 Tibetan, 5 Miao, 7 Yi, 8 Zhuang, 9 Bouyei, 10 Korean, 11 Manchu, 99 Other), Birth Area (13 Hebei, 14 Shanxi, 21 Liaoning, 22 Jilin, 23 Heilongjiang, 31 Shanghai, 32 Jiangsu, 33 Zhejiang, 34 Anhui, 35 Fujian, 36 Jiangxi, 37 Shandong, 41 Henan, 42 Hubei, 43 Hunan, 44 Guangdong, 45 Guangxi Zhuang Autonomous Region, 51 Sichuan, 52 Guizhou, 53 Yunnan, 61 Shaanxi, 62 Gansu, 80 Not available, 90 Other), Parent's education (1 Illiterate/Semi-literate, 2 Primary school, 3 Junior high school, 4 Senior high school/secondary school/technical school/vocational senior school, 5 3-year college, 6 4-year college/Bachelor's degree, 7 Master's degree, 8 Doctoral degree), Parent's occupation (0 Armed forces, 1 Managers, 2 Professionals, 3 Technicians and Associate professionals, 4 Clerks, 5 Services and Sales workers, 6 Agricultural, Forestry and Fishery workers, 7 Craft and trade workers, 8 Plant and machine operators and assemblers, 9 Elementary occupations, 10 Unemployed). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: CFPS (2018).

Figure 1c: Transformation Tree for India



Note: Splitting nodes show their sample size and the p-value associated with the split. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (2 Forward caste, 3 Other Backward castes (OBC), 4 Dalit, 5 Adivasi, 6 Muslim, 7 Christian, Sikh, Jain), Parent's Education (0 None, 1 Incomplete primary, 2 Complete primary, 3 Incomplete secondary, 4 Complete secondary, 5 Higher secondary, 6 Post-secondary or higher), Birth Area (1 Jammu & Kashmir, 2 Himachal Pradesh, 3 Punjab, 4 Another State, 5 Uttarakhand, 6 Haryana, 7 Delhi, 8 Rajasthan, 9 Uttar Pradesh, 10 Bihar, 11 Overseas, 18 Northeast, 19 West Bengal, 20 Jharkhand, 21 Orissa, 22 Chhattisgarh, 23 Madhya Pradesh, 24 Gujarat, 27 Maharashtra, 28 Andhra Pradesh, 29 Karnataka, 32 Kerala, 33 Tamil Nadu). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: IHDS (2012).

Figure 1d: Transformation Tree for South Africa



Note: Splitting nodes show their sample size and the p-value associated with the split. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (1 African, 2 Asian/Indian, 3 Coloured, 4 White), Parent's Education (0 Zero, 1 Grade 1, 2 Grade 2, 3 Grade 3, 4 Grade 4, 5 Grade 5, 6 Grade 6, 7 Grade 7, 8 Grade 8, 9 Grade 9, 10 Grade 10, 11 Grade 11, 12 Grade 12), Parent's Occupation (0 Military, 1 Managers, 2 Professionals, 3 Technicians and Professionals, 4 Clerical Support, 5 Service and sales, 6 Farm, Forest, Fishery, 7 Craft, 8 Operators, 9 Elementary, 10 Others). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: NIDS (2017).

In China (Figure 1b), the four-level tree yields a thirteen-type partition, with incomes ranging from 47% to 253% of the mean income. The birth area variable plays a critical role in the transformation tree: it is the first splitting factor, and it also appears in five additional splitting nodes, highlighting the importance of the geography of birth for the distribution of life chances in China. For example, the two most advantaged types consist exclusively of individuals born in Shanghai. Other influential variables include parental occupation, education, and, to a lesser extent, ethnicity (the worst-off type contains only Mongol and Yi individuals). [22]

In India (Figure 1c), the four-level tree yields fifteen types, with incomes ranging from 47% (Type 15) to 374% (Type 29) of the mean. Father's education determines the first split in the transformation tree, followed by the mother's education. The structure of the tree indicates a dominant role for the interaction between these two variables. The wealthiest group, whose expected income is nearly four times the sample mean, is composed of individuals whose parents are both at the top of the educational distribution. Additionally, birth area and ethnicity (which here consists of caste and religious identities) also contribute substantially to predicting the shape of the income distribution, which is heterogeneous not only in terms of its mean but also in its higher-order moments. While the poorest types tend to follow a clearly log-normal distribution, the richer types exhibit distributions skewed to the right, suggesting greater variability, and possibly some downward risk (in the long left tails) for these higher-income groups. These insights into the shape of each type's distribution are one benefit of our approach to measuring inequality of opportunity using transformation trees.

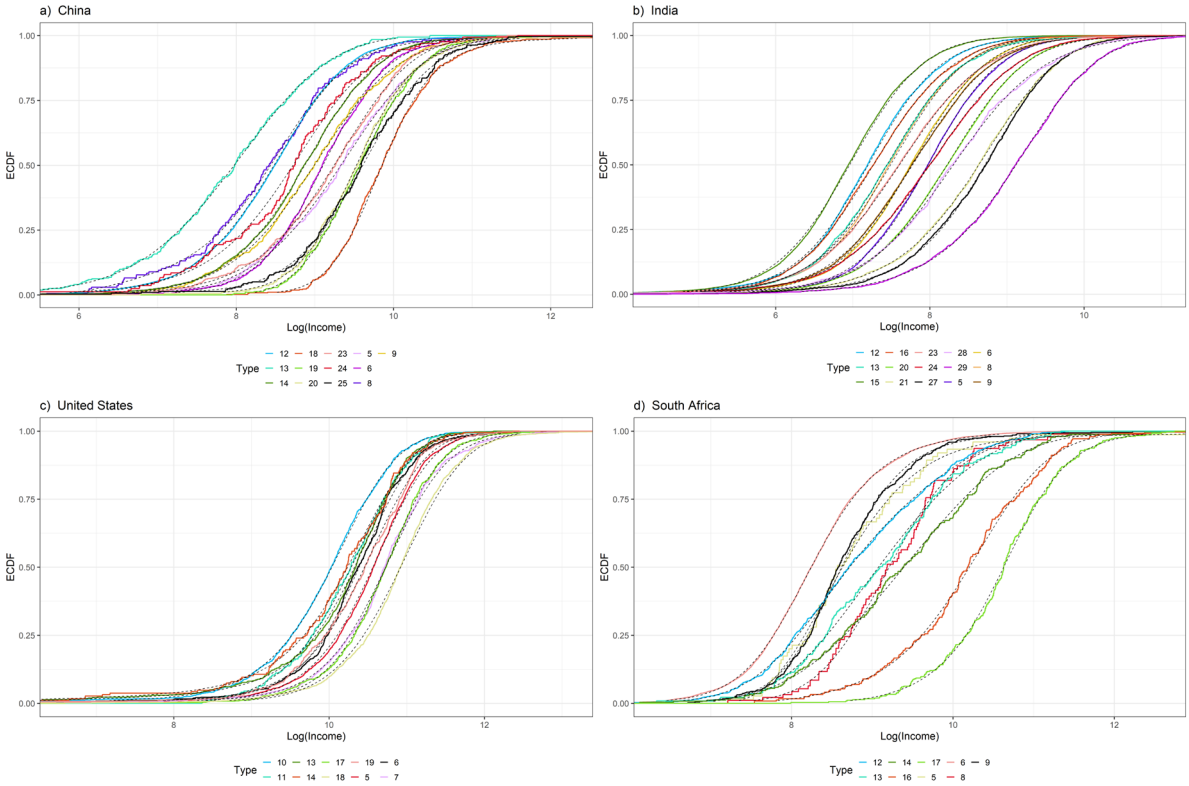
Finally, in South Africa (Figure 1d), the pruned tree has nine nodes or types. Unsurprisingly, its structure is determined primarily by ethnicity. The white population, which is exclusively concentrated in Types 16 and 17, displays an income distribution so markedly shifted to the right that it appears as though it might have been drawn from a different country. Among the non-white groups, those including Asian and Coloured individuals have a higher expected income

[22] The split at Node 4 in the Chinese tree is worth a comment, as it groups parents with a Master's degree together with the least educated category, in Type 5. This illustrates a feature of the algorithm when there are very few observations in a particular category. When the algorithm uses a certain circumstance to divide the sample, it must place all individuals from the node that originates the split into either one subgroup or the other. If there are very few respondents who have a specific value for the characteristic in question – there are only four individuals in the entire sample reporting a father with Master's degree - the assignment to the group can be almost random. Naturally, because this happens only when the number of individuals is very small, the consequences for summary *IOp* estimation are minimal.

and tend to follow a log-normal distribution. In contrast, groups composed by African individuals exhibit a distribution with a density mass skewed to the left.

Besides the parameterized density functions shown to the right of the trees in Figure 1, type distributions can also be visualized as cumulative distribution functions. Figure 2 shows both the empirical CDFs for each type (as solid coloured lines) and the corresponding predicted CDFs (ECDFs: $F\left(\tilde{y}_{qc}, \hat{\theta}^{N}(c)\right)$) generated by the Bernstein polynomials, as dashed lines.

Figure 2. ECDF of the type-partition (pruned trees) for the most recent waves



Note: here we present the ECDF of the types obtained from the pruned Transformation Trees displayed in Figure 1a (USA), Figure 1b (China), Figure 1c (India), and Figure 1d (South Africa). Solid lines represent the ECDF for each type and are coloured consistently with the corresponding types in the trees. Dashed lines represent the corresponding CDF predicted with the Bernstein polynomial. Figures. Source: CFPS (2018), IHDS (2012), NIDS (2017) and PSID (2018).

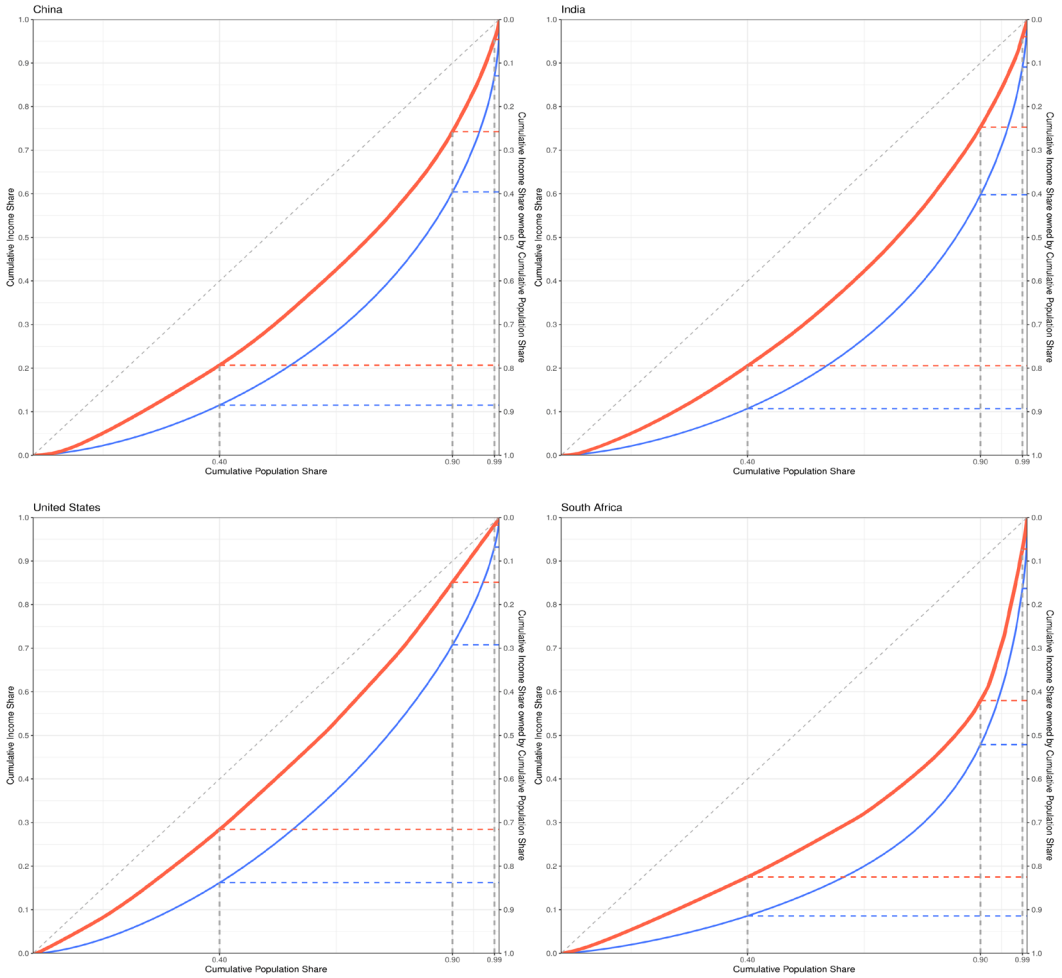
These CDF plots provide a striking visual depiction of the structure and extent of inequality of opportunity in each country. Compared to the United States, type distributions are much further apart in the three developing countries in our sample, particularly South Africa, where the two White types stochastically dominate all others by a large margin. We can also observe interesting crossings, such as that between types 5 and 24 in India. These two types have relatively similar

means (122% and 141% of the overall average, respectively), but Type 24 has a more unequal distribution, so the CDFs cross and the welfare of the two types cannot be unambiguously ranked (in terms of first-order dominance). Similar crossings can be observed in all four countries, revealing that types differ not only in terms of their first moments, but also in their higher-order moments.

From the trees to Lorenz Curves and scalar measures of inherited inequality

As discussed, Figures 1 and 2 draw on four-level pruned trees. To compute IOp, we rely on the full trees shown in Figure B2 in Appendix 2. Those trees generate finer types, with their own income predictions, $\tilde{y}_{qc}$. These are then adjusted for differences across tranche means as in Equation (8) to yield the distribution of predicted incomes $\hat{y}_T$, which is used for computing our proposed absolute and relative measures of inherited inequality or ex-post inequality of opportunity (Eq. 9). We can also define a Lorenz Curve of these predicted incomes, $L(\hat{y}_T)$, to which we refer as the Opportunity Lorenz Curve. Figure 3 displays both Opportunity Lorenz curves and regular Lorenz curves, $L(y)$, for the most recent surveys for each country. As expected, the Opportunity Lorenz curve dominates that for income, and the area between $L(\hat{y}_T)$ and the line of equality captures the extent of inequality of opportunity in each society. The dashed lines indicate the shares of income and opportunities held by the bottom 40%, 90%, and 99%, corresponding to the shares presented in Table 2 below.

Figure 3 Income and Opportunity Lorenz Curves for China, India, South Africa and the US



Note: The red line corresponds to the Lorenz curve of $\hat{y}_T$, while the blue line corresponds to the Lorenz curve of income. The dashed lines correspond to the income (Table 1) and $\hat{y}_T$ (Table1) shares received by the Bottom 40%, Top 10% and Top 1%. Source: CFPS (2018), IHDS (2012), NIDS (2017), and PSID (2018).

To go from the Lorenz Curves in Figure 3 to scalar measures of inequality, we choose two different indices, namely the Gini coefficient and the mean log deviation (MLD). Values for income inequality were presented in Table 1 above, and Table 2 (Panel A) presents the corresponding inequality of opportunity measures, once again for the latest available survey waves. Although we report both Ginis and MLDs in Table 2, we focus the subsequent discussion on the Gini estimates.[23] For both indices, we report the absolute measure of inherited inequality, $I_n^A(y, \boldsymbol{c}, f_T) = I(\hat{y}_T)$, as well as the relative, $I_n^R(y, \boldsymbol{c}, f_T) = \frac{I(\hat{y}_T)}{I(y)}$. In the tables, the latter is

[23] As noted by Brunori, Palmisano, and Peragine (2019), the Gini coefficient is more sensitive to the central parts of the distribution, where group means tend to cluster, rather than to the lower tail. In that sense, the Gini is better suited to studying IOp than the MLD.

denoted $\boldsymbol{IOR_T}$. The last three columns of the Table also report predicted income (or opportunity) shares for the top 1%, top 10%, and bottom 40%.

Table 2: Inequality of Opportunity results for the most recent waves

Panel A: Ex-post Inequality of Opportunity

Country	Year	Gini ($\hat{y}_T$)	IOR_T (Gini)	MLD ($\hat{y}_T$)	IOR_T (MLD)	Top 1%	Top 10%	Bottom 40%
China	2018	0.292	58.8%	0.172	37.5%	0.079	0.27	0.245
India	2012	0.327	62.0%	0.207	40.0%	0.040	0.278	0.168
South Africa	2017	0.495	81.1%	0.413	59.9%	0.039	0.313	0.209
USA	2018	0.141	36.3%	0.052	17.3%	0.017	0.148	0.372

Note: The three columns on the right represent the share of $\hat{y}_T$ accruing to the Top 1%, the Top 10%, and the Bottom 40% in the $\hat{y}_T$ distribution. Source: CFPS (2018), IHDS (2012), NIDS (2017), and PSID (2018).

Panel B: Ex-ante Inequality of Opportunity

Country	Year	Gini ($\hat{y}_{BHM}$)	IOR_{EA} (Gini)	MLD ($\hat{y}_{BHM}$)	IOR_{EA} (MLD)	Top 1%	Top 10%	Bottom 40%
China	2018	0.219	44.1%	0.076	16.6%	0.063	0.245	0.299
India	2012	0.279	52.9%	0.123	23.7%	0.048	0.269	0.184
South Africa	2017	0.468	76.7%	0.36	52.2%	0.041	0.292	0.219
USA	2018	0.154	39.6%	0.037	12.3%	0.014	0.149	0.34

Note: IOR_{EA} is the Brunori et al., (2023) estimate and $\hat{\boldsymbol{y}}_{\boldsymbol{BHM}}$ denotes the incomes received by types obtained using that ex-ante method. Source: CFPS (2018), IHDS (2012), PSID (2018), and NIDS (2017).

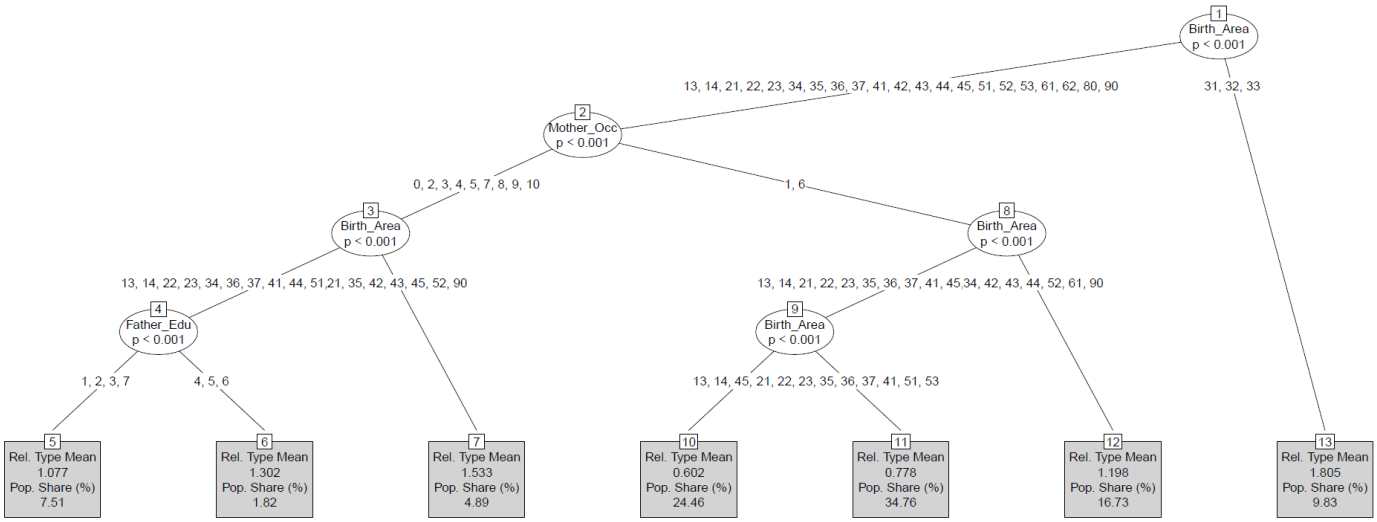
For these years, the opportunity Gini coefficient ranges from 0.141 in the US to 0.495 in South Africa. The latter is a remarkable number: the opportunity Gini for South Africa is higher than the overall income Gini coefficient of the United States and almost as high as total inequality in China. Indeed, inherited inequalities account for a remarkable 81% of the (very high) income inequality in South Africa. India has the second-highest level of IOp, with a Gini of 0.33, accounting for 62% of total inequality, with China not far behind. However, the shape of the opportunity distribution differs between those two countries, mirroring what we have already seen for the distribution of incomes. In India, the top 1% receives 4% of total opportunities, half of the share in China, where the top 1% holds 8%. Yet, India's top 10% captures about the same as in China, while the bottom 40% receives significantly less (17% compared to 25% in China), even than in South Africa (21%). This suggests that, although India exhibits less concentration at the very top of the distribution of opportunities compared to China, it has a much larger share of the population that is extremely opportunity deprived. Figure B3 in the Appendix presents

the time trends in ex-post IOp across all four countries this century, showing that IOp was largely stable in the United States during the 2000 – 2018 period, while it rose both in India (between 2005-2012) and China (particularly since 2014). South Africa followed a U-shaped pattern, with inequality of opportunity declining until 2015 and then rising again in 2017.

For comparison, Panel B of Table 2 contains benchmark estimates from applying an ex-ante approach to our data. Specifically, we follow the approach of Brunori, Hufe and Mahler (BHF, 2023) to construct conditional inference trees and random forests. The resulting ex-ante IOp estimates are typically lower than our ex-post results, both in absolute and relative terms: they are lower in all four cases for the MLD, including by a factor of less than 0.5 for China. They are also lower for the Gini in all countries except the United States. Top 10% shares are also lower in the ex-ante estimates, except again for the US, where they are basically the same. The picture is a little less clear for the top 1% share, where the ex-ante estimate is higher in both India and South Africa. The overall pattern, though, is that even when comparing our approach to the closest possible ex-ante alternative – another tree-based machine learning approach, but using differences in means rather than full distributions to split types – overall IOp levels and shares tend to be higher in the ex-post approach.

The difference is particularly marked in the case of China, where the ex-post Opportunity Gini coefficient is 1.33 times the ex-ante. Such large differences between ex-ante and ex-post estimates tend to arise whenever considering only the first moment of the conditional distribution is not sufficient to understand the entire conditional distribution of types. To provide an intuition for this, Figure 4 shows the ex-ante conditional inference tree (CIT) for China in 2018, estimated with same four-level stopping rule as the transformation tree in Figure 1b. Consider Type 5, which represents 7.5% of the sample. Individuals in this group have mothers who are neither Agricultural, Forestry, and Fishery workers nor managers, and their fathers have low levels of education. This type has an expected income close to the population mean. However, when examining the *ex-post* types to which individuals from ex-ante Type 5 are mapped, a clear divergence emerges.

Figure 4: Conditional Inference Tree to evaluate ex-ante IOp in China

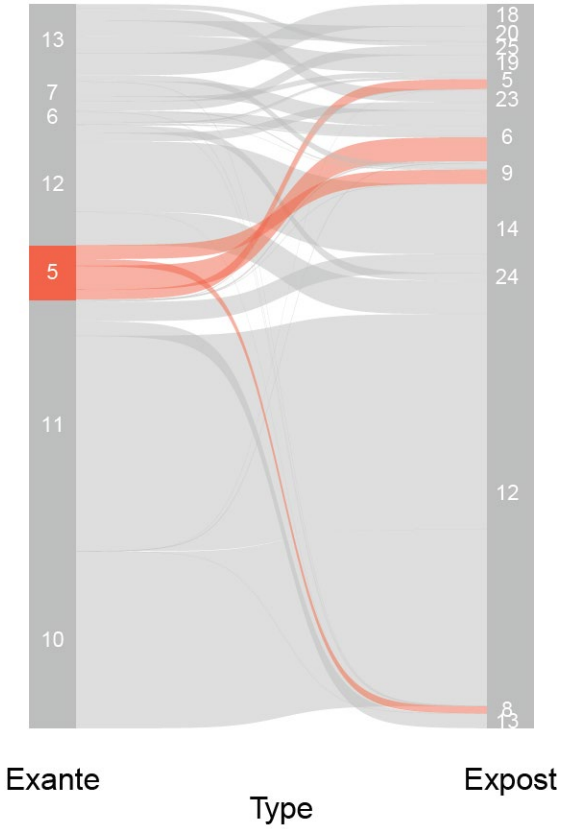


Note: Splitting nodes show the p-value associated to the splitting. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (1 Han, 2 Mongol, 3 Hui, 4 Tibetan, 5 Miao, 7 Yi, 8 Zhuang, 9 Bouyei, 10 Korean, 11 Manchu, 99 Other), Birth Area (13 Hebei, 14 Shanxi, 21 Liaoning, 22 Jilin, 23 Heilongjiang, 31 Shanghai, 32 Jiangsu, 33 Zhejiang, 34 Anhui, 35 Fujian, 36 Jiangxi, 37 Shandong, 41 Henan, 42 Hubei, 43 Hunan, 44 Guangdong, 45 Guangxi Zhuang Autonomous Region, 51 Sichuan, 52 Guizhou, 53 Yunnan, 61 Shaanxi, 62 Gansu, 80 Not available, 90 Other), Parent's education (1 Illiterate/Semi-literate, 2 Primary school, 3 Junior high school, 4 Senior high school/secondary school/technical school/vocational senior school, 5 3-year college, 6 4-year college/Bachelor's degree, 7 Master's degree, 8 Doctoral degree), Parent's occupation (0 Armed forces, 1 Managers, 2 Professionals, 3 Technicians and Associate professionals, 4 Clerks, 5 Services and Sales workers, 6 Agricultural, Forestry and Fishery workers, 7 Craft and trade workers, 8 Plant and machine operators and assemblers, 9 Elementary occupations, 10 Unemployed). The panels at the bottom indicate the share of the population each type represents, and their average income relative to the overall sample mean. Source: CFPS (2018).

The conditional inference tree in Figure 4 and the transformation tree in Figure 1b are estimated in the exact same sample, so it is possible to map individuals to the types to which they belong in each exercise. This is what the Sankey (or alluvial) plot in Figure 5 shows. The left margin shows the ex-ante partition obtained using the CIT (as plotted in Figure 4), while the right margin displays the partition derived from the TrT (Figure 1b). In both margins, types are ranked from top to bottom in descending order of expected incomes. Ex-ante Type 5 is highlighted in light orange to illustrate how individuals from this group are distributed across four different ex-post types. These destination types differ significantly in terms of income. For instance, ex-post Type 5 has an average income that is 52% higher than the population average, whereas ex-post Type 8 has an average income that is less than half that: about 70% of the population mean. These regroupings arise when there are meaningful differences in the higher moments of the conditional distributions of potential types, so that the transformation tree and the CIT

algorithm yield quite different partitions. This example illustrates how a substantial divergence between ex-ante and ex-post partition can arise in practice.

Figure 5: An example of heterogeneity in ex-ante and ex-post partition in China



Note: The Sankey plot maps the same individuals according to two different type partitions. On the left-hand side, the ex-ante partition (Figure 4), and on the right-hand side, the ex-post partition (Figure 1b). We have highlighted Type 5 in the ex-ante partition, that splits into four very different types in the ex-post partition. Source: CFPS (2018).

It is somewhat harder to compare our main results with previous studies, which have employed various different statistical approaches and often used different samples and income definitions. For example, in the case of the United States, Pistolesi (2009) report a similar level of IOp in terms of MLD for the year 2000 to what we obtain. However, their analysis focuses on earnings and is restricted to working males, which limits comparability. In most other cases, previous estimates are generally based on ad-hoc ex-ante methods, and these tend to be lower – often much lower – than ours. For India, Kundu and Lefranc (2020) estimate that ex-ante IOp in 2012 ranges from 8% to 39%, depending on the set of regressors used in a parametric model. Their estimate using conditional inference regression trees is 32%, which compares to our ex-post Gini share of 62%. For China, Wu (2018), using the same data source and inequality index (Gini) as our study, reports relative IOp levels between 30% and 40% for the years 2010 and 2012,

whereas we observe relative IOp levels of 53% in 2010 and 42% in 2012. Finally, Piraino (2015) employs an ex-ante approach and two econometric methods to estimate IOp in gross employment earnings in South Africa, using up to 54 Roemerian types. Using data for male workers in 2008 and 2012, he finds IOp shares ranging from 17% to 24% of total inequality, as measured by MLD. These estimates compare to our MLD-based estimate of 57% in 2012. Once again, although comparisons are hampered by differences in samples and/or income definitions, most previous IOp estimates for our sample of countries tend to be of the ex-ante variety, and considerably lower than ours, consistently with the broad pattern of the comparison between Panels A and B of Table 2.

The role of individual circumstances

Because the prediction function in Equation (16) is highly non-linear in circumstances, any assessment of the relative contribution of individual circumstances to inequality in predicted incomes, $I(\hat{y}_T)$, cannot rely on marginal effects. As in other cases in inequality analysis, the decomposition method most suitable to our application is the Shapley-Shorrocks decomposition (Shapley, 1953; Shorrocks, 2013). Intuitively, this decomposition computes the total contribution of a particular circumstance variable c_k to predicted inequality as the average reduction in the latter when c_k is omitted from the prediction, with the average taken across all possible combinations of circumstances that originally include c_k. (See Shorrocks, 2013). A description of the algorithm used to compute the decomposition also helps clarify its logic:

A) Draw a subsample of the full sample;[24]
B) Estimate IOp in this subsample, as described in Section 3, but setting $\alpha = 1$; $n_{min} = 0$
C) Further, estimate IOp in the subsample for all possible permutation sequences that eliminate circumstance c_k. This elimination is performed by replacing c_k with a constant vector **1**;
D) Estimate a tree and IOp after each elimination sequence and store results;
E) Average IOp across all permutation sequences. The difference between overall IOp and this average is the specific contribution of c_k;
F) Repeat steps A-E *z* times, to account for different potential data-generating processes. In our case, we set *z* = 100;
G) Estimate the contribution of c_k to IOp as the average contribution across these *z* repetitions;

[24] Following the convention often used in tree bagging procedures, we draw subsamples of 63.2% of the original sample size (see Hothorn, Hornik, and Zeileis, 2006).

H) Repeat the algorithm for each c_k, $k \in \{1, \dots, K\}$.

This algorithm grows trees on subsamples of the initial population, permitting each tree to attain significant depth. These two adjustments enable all circumstances with predictive power to contribute to defining the partition of types, at least in some iterations, making the assessment of the relative contribution of each circumstance more robust to the typical problem of the variance of estimates based on a single tree. Table 3 presents the results of the Shapley-Shorrocks decomposition for our four countries and the seven circumstance variables available in the most recent wave. Results are presented as percentage shares of the ex-post opportunity Gini coefficients reported in Table 2.

Table 3: Ex-post Shapley value decomposition (as %) for the most recent waves

Circumstances	China	India	South Africa	USA
Birth Area	15.41	20.87	-	14.25
Ethnicity	8.65	16.32	32.41	15.24
Father Education	10.46	27.74	16.43	19.17
Father Occupation	16.32	-	13.78	15.42
Mother Education	17.26	27.89	16.07	16.5
Mother Occupation	29.53	-	16.06	12.29
Sex	2.36	7.17	5.25	7.13

Note: Values in these tables represent the relative contribution (as %) of circumstances to the ex-post IOp estimates reported in Table 2, Column Gini ($\hat{y}_T$). The sum of values within columns adds up 100%. Missing values (-) correspond to circumstances that are not available in the data. Source: CFPS (2018), IHDS (2012), NIDS (2017), and PSID (2018).

Notice that due to the unobservability of parental occupation in India and area of birth in South Africa, results are only fully comparable for China and the United States.[25] Looking across those two countries, it is interesting that, despite the larger number of ethnic groups in China, the Shapley value for race in the US is twice as large. While father's occupation and mother's education have similar values in both countries, father's education appears significantly more

[25] Even in this case, some might argue that comparability is limited by how the same circumstances are coded across countries. For example, in the USA there are five racial categories, while in China there are twelve ethnic groups. However, such classifications are linked to a country's structure of opportunity. We can and should not impose identical categories across countries, which differ in terms of their territories, cultural diversity, and social structure. These aspects of a country's social organization are part and parcel of its opportunity distribution, and they should be reflected in the data used — without attempting to make them uniform. It is for the learning algorithm to select the most salient binary splits across categories in each case.

important in the US, whereas mother's occupation plays a larger role in China. The influence of sex also appears to be much more pronounced in the US.

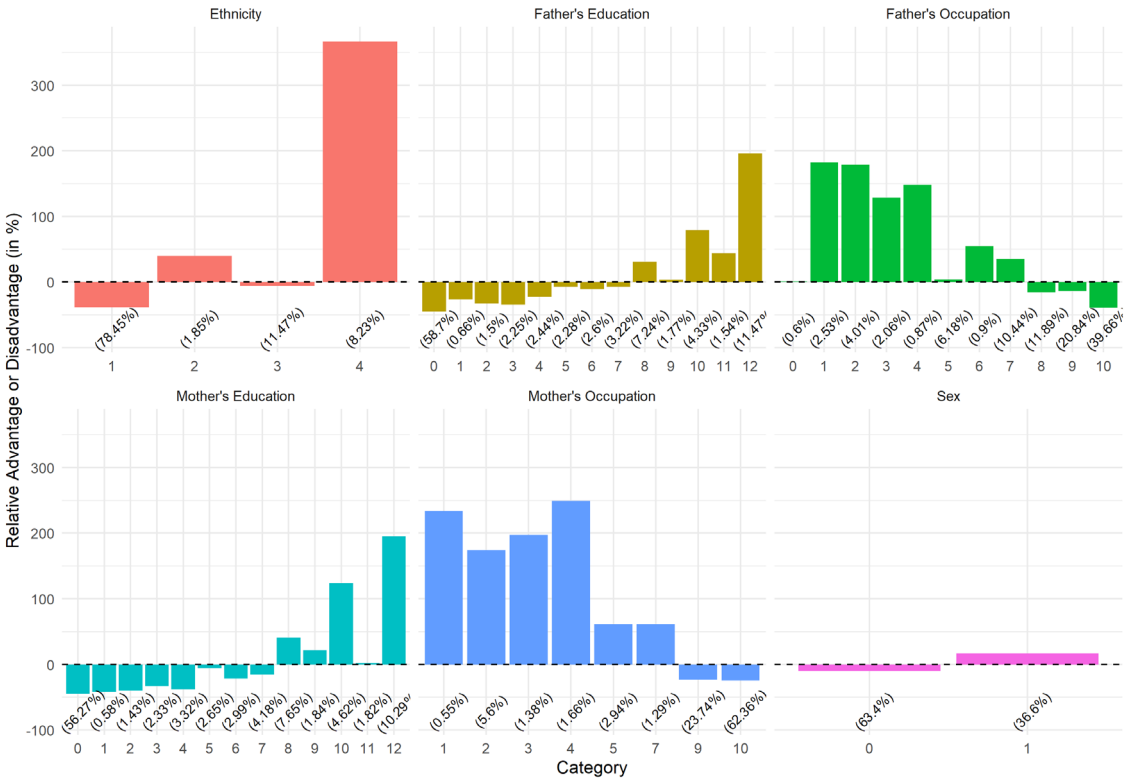
When including India and South Africa in the comparison, the dominant role of race emerges clearly in South Africa: it is comparable to the combined effect of caste and area of birth in India. Interestingly, the Shapley value for parents' education in India is not too far from the sum of the values for both occupation and education of the parent in South Africa, suggesting that the presence of unobservable circumstances may inflate the Shapley value of observable and correlated circumstances. Regarding the role of sex, it is important to recall that our analysis is based on equivalized household income. Therefore, we expect sex to play a significant role only in contexts where single-parent households are not uncommon and where income disparities between male- and female-headed households are substantial. This is the case in India and the United States. In contrast, the difference is smaller in our sample for China, where female-headed households earn about 95% of what their male-headed counterparts earn. Naturally, it should go without saying that, in keeping with the measurement-using-prediction spirit of our analysis, these decompositions are purely descriptive.

Moreover, when commenting on the role of circumstances in predicting the conditional distribution of income, we should consider that Shapley values measure the reduction in predicted inequality when a specific circumstance is removed from the analysis. This value is influenced by the distribution of the circumstance itself. For example, in a society where most individuals are Black and only few are White, removing the race variable will not significantly reduce the model's explanatory power. This is because, for the majority, the conditional distribution closely aligns with the unconditional one, and only for a small subset does race influence the income distribution. However, from the point of view of the minority, that characteristic may matter a great deal. In other words: while the average contribution of individual circumstances contains valuable information, so would an estimate of the marginal importance of belonging to a specific circumstance category.

We therefore complement Shapley values by estimating the marginal effect (on predicted incomes) of being affected by a specific characteristic, for example, being White in South Africa. We do this by regressing individuals' predicted incomes, $\hat{y}_T$, on a set of dummy variables, each representing a category of a given circumstance. This exercise is similar to the estimation of Partial Dependence Plots (PDPs), which are frequently used in machine learning to complement decomposition techniques such as Shapley values. Figure 6 presents the marginal effects in the most recent wave for South Africa. Both the signs and magnitudes of the effects are broadly

consistent with expectations. The substantial positive effect of being White is clearly visible in the first panel: being White is associated with an opportunity premium of 350%. A substantial "college premium" – more precisely, of completing or going beyond Grade 12 – can also be observed for both fathers' and mothers' education (Category 12).

Figure 6: Marginal effect of circumstances on opportunities in the South Africa



Note: Values on the y-axis represent the relative advantage or disadvantage associated with each category, computed as $100 \times \frac{\text{average income (category)}}{\text{average income (sample)}}$. *Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (1 African, 2 Asian/Indian, 3 Coloured, 4 White), Parent's Education (0 Zero, 1 Grade 1, 2 Grade 2, 3 Grade 3, 4 Grade 4, 5 Grade 5, 6 Grade 6, 7 Grade 7, 8 Grade 8, 9 Grade 9, 10 Grade 10, 11 Grade 11, 12 Grade 12), Parent's Occupation (0 Military, 1 Managers, 2 Professionals, 3 Technicians and Professionals, 4 Clerical Support, 5 Service and sales, 6 Farm, Forest, Fishery, 7 Craft, 8 Operators, 9 Elementary, 10 Others). Numbers in parentheses denote population shares in each category. We are not showing categories populated by less than 0.5% of the sample size. Source: NIDS (2017).*

Figures B4–B6 in Appendix 2 show analogous marginal effects for China, India, and the United States. Noteworthy findings include the strong positive effect of being born in certain regions of China—for instance, being born in Shanghai is estimated to contribute a marginal advantage of more than 150% to predicted incomes. In both India and the U.S., having well-educated parents is associated with significantly higher opportunities. In India, the premium is particularly high for maternal education: for the small minority reporting a mother with education above the

secondary level, the estimated marginal advantage exceeds 200%. In the U.S., the effect is higher for father's education. A sizable negative effect is observed for the large Black community in the United States, with an estimated penalty of approximately 25%.

The lower envelope of quantile functions

As noted in the Introduction, another advantage of our conditional CDF-based approach is that it enables us to compute estimates of the social objective function proposed in the original theory of equal opportunity (Roemer, 1993, 1998). In proposing a normative objective function, egalitarians must contend with the *levelling-down objection*: if the objective were simply to eliminate inequality in predicted incomes, $I(\hat{y})$, and thus immobility or inequality of opportunity, this might be achieved by setting all incomes to zero – or some other very low but constant value. Policies might be arranged in such a way that there was no inherited inequality, but everyone lived in abject poverty.

The standard normative response to this objection is Rawls's proposal that inequalities should be tolerated only insofar as they are to the benefit of the worst-off (Rawls, 1971). This gives rise to the familiar Rawlsian maximin objective functions and, indeed, various versions of maximin objectives have been proposed in the context of inequality of opportunity.[26] A dominant early version, due to Roemer (1998), is to arrange society and choose policies so as to maximize the (average of the) lowest incomes at each quantile, across the conditional distribution functions of all types. Recalling from the general framework in Section 2, that there are M types, $\tau_m := \{\forall i | \boldsymbol{c_i} = \boldsymbol{c_m}\}$, whose conditional cumulative distribution functions are of the form $F(y|\boldsymbol{c_m})$, define the lower envelope of the joint distribution {y, **c**} as:

$$\Lambda(q) = \min_{\tau_m} F^{-1}(q, \boldsymbol{c_m}) \tag{10}$$

And choose policies so as to:

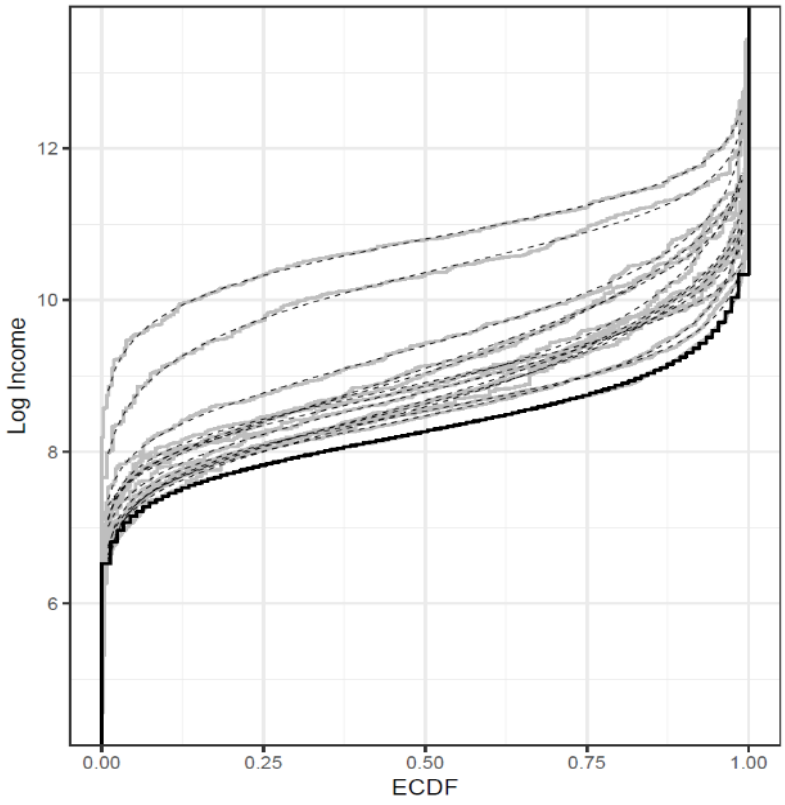
$$Max \int_0^1 \Lambda(q) dq \tag{11}$$

As Roemer and Trannoy (2016) put it: "*We do not simply want to render the functions identical at a low level, so we need to adopt some conception of 'maxi-minning' these functions. [...] A natural approach is therefore to maximize the area under the lower envelope of the [*quantile*] functions.*" (p. 231).

[26] See, e.g., Van de Gaer (1993) and Bourguignon, Ferreira, and Walton (2007).

Equation (10) defines the lower envelope of the set of quantile functions (inverse functions of the distribution function). Graphically, the type quantile functions, shown in Figure 7 below for South Africa, are obtained by inverting the conditional CDFs in Figure 2d. $\Lambda(q)$ defines the lowest points in the graph at each quantile. If the poorest type were first-order stochastically dominated by all other types, then the lower envelope would simply be its quantile function, and Equation (11) would mandate maximizing its average income, equal to the area under the quantile function. When quantile functions cross at the bottom of the graph, Equation (11) mandates maximizing the average income of the lower envelope of the quantile functions. If there were no inequality of opportunity, all of society would be one type and $\int_0^1 \Lambda(q)dq$ would be its average income. Therefore, the value of the maximand in (11) is informative per se, as a measure of the opportunity floor in a society, and is interesting also in relative terms, as a measure of how close that floor is to the average income, $Lenv_R = \frac{\int_0^1 \Lambda(q)dq}{\int_0^1 F^{-1}(q)dq}$.

Figure 7: Type quantile functions and the lower envelope for South Africa

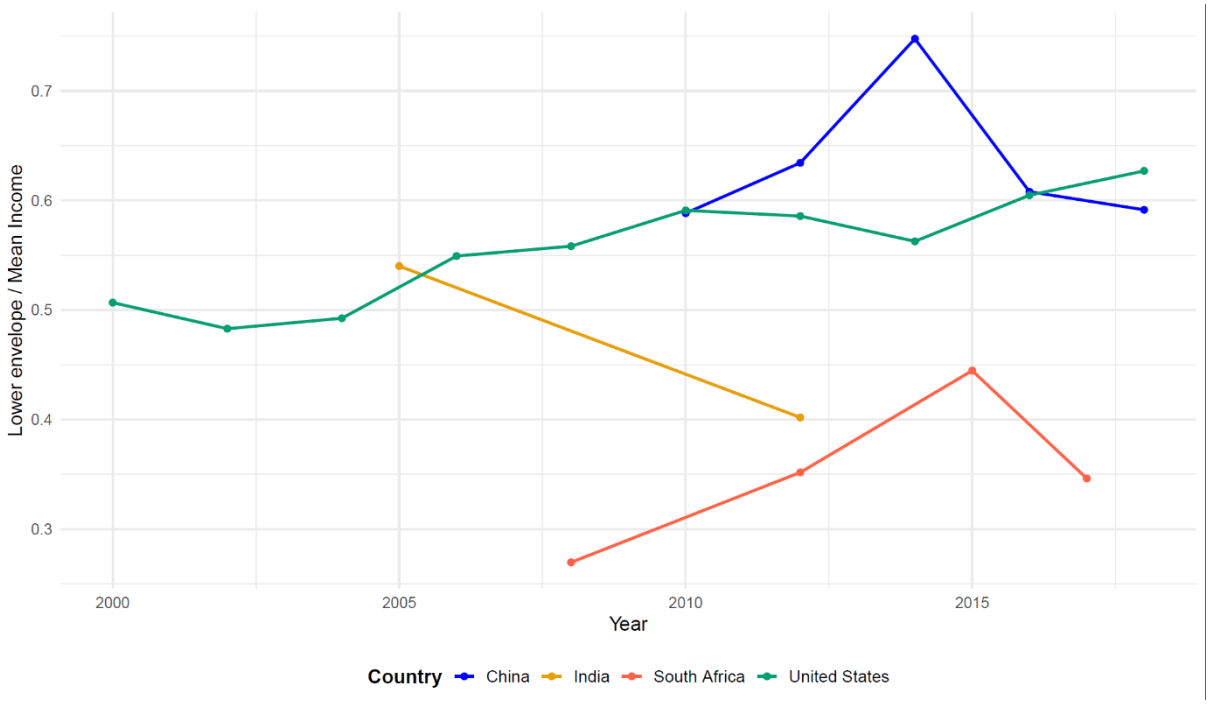


Note: Solid lines represent the ECDF for each type and are labelled consistently with the corresponding types in the trees. Dashed lines represent the corresponding CDF predicted with the Bernstein polynomial. Source: NIDS (2017).

In practice, a literal computation of $\int_0^1 \Lambda(q)dq$ (Lenv, hereafter) might be over-sensitive to small types detected in a particular sample. We therefore propose a robust version of the lower envelope which consists, in each quantile, of the average income across the worst-off types adding up to at least 10% of the population. The heavier line in Figure 7 shows the robust lower envelope for South Africa in 2017, against the full set of conditional quantile functions.

Figure 8 presents the evolution of $\mathrm{Lenv}_{\mathrm{R}}$ for the four countries, over the 2000-2018 period. A striking result is the high value recorded for China, where the poorest types appear to have been catching up with the average income until 2014 (75%), but then lost some ground thereafter. In the most recent wave, China's $\mathrm{Lenv}_{\mathrm{R}}$ is comparable to that of the United States, at around 60% of average income in our sample. In terms of trends, the United States shows a rising trajectory in the first two decades of the century, whereas India exhibits a downward trend between 2005 and 2012, falling from 54% to 40%. The improvement in South Africa between 2008 and 2015 is not sufficient to converge toward the other countries. Its $\mathrm{Lenv}_{\mathrm{R}}$ in 2017 is still around 35%.

Figure 8: Lenv_R over time in China, India, South Africa and the US.



Source: CFPS, IHDS, PSID, and NIDS.

Figure B7 in Appendix 2 shows an analogous graph for *Lenv* in absolute terms. The United States' area under the lower envelope rises from approximately $22,000 in 2000 to about $30,000 in 2018. The other three countries exhibit significantly lower and more closely aligned levels. In

India, Lenv declines from around $3,500 to $1,500 between 2005 and 2011. China and South Africa, in contrast, display similar upward trajectories in the 2000s.

6. Conclusions

The extent to which economic advantage is inherited from previous generations and shaped by pre-determined circumstances is a matter of both positive and normative interest. Many, if not most, approaches to quantifying this phenomenon rely on prediction exercises, essentially assessing how well incomes can be predicted by pre-determined circumstances such as parental income, biological sex, race, or other indicators of family background. We showed that many commonly used measures of intergenerational mobility and inequality of opportunity can be written as functions of the ratio of inequality in these predicted incomes to inequality in observed current-generation incomes.

We then proposed a new approach for measuring inherited inequality that is sensitive to differences across the full conditional income distributions – rather than just the means – of subpopulations that share the same inherited characteristics – "types" in the IOp literature. This method, based on transformation trees (Hothorn and Zeileis, 2021), represents an improvement over previous approaches to estimating inequality of opportunity because it is designed to partition the population and estimate distribution functions optimatlly, given the trade-off between a downward omitted-variable bias and an upward overfitting bias that is inherent to the model selection problem in this literature.

We applied this method to thirty-six representative household surveys from four large and systemically important countries, namely the United States (25 waves, between 1970 and 2018), China (five waves, between 2010 and 2018), India (two waves, 2005 and 2012) and South Africa (four waves, between 2008 and 2017). We found high absolute levels of inherited inequality, measured as ex-post inequality of opportunity, with Opportunity Gini coefficients of 0.14 in the US, 0.29 in China, 0.33 in India, and 0.50 in South Africa in the latest available years. These correspond to substantial shares of total income inequality – 36% for the US, 59% for China, 62% for India, and 81% for South Africa – attesting to the heavy weight of inherited characteristics in predicting current economic success in all four countries, but particularly the three poorer ones.

Comparing these estimates both to state-of-the-art ex-ante methods[27] applied to the very same data and using the same income and circumstance variables, as well as to earlier estimates in

[27] Random forest estimates obtained using the approach of Brunori et al. (2023).

the literature that use other income definitions and statistical approaches, we found our estimates to be generally, and sometimes substantially, higher. For China, for example, we found an ex-post inherited inequality share of 59%, whereas the ex-ante estimate on the exact same sample was 44%. These differences reflect both differences in the type partitions generated by the two algorithms, and the fact that the ex-post method aggregates horizontal differences for all quantiles along the (adjusted) type cumulative distribution function, whereas the ex-ante method relies on differences in means only. We illustrated the subtle partition differences that can occur with a Sankey plot mapping Chinese individuals from their ex-ante to their ex-post types.[28]

We also estimated both average and marginal contributions of specific circumstances (and categories, in the marginal case) to overall inherited inequality. The relative importance of these circumstances varied substantially across countries, reflecting their different histories and socio-economic structures. Race was unsurprisingly dominant in post-Apartheid South Africa, whereas area of birth was important in both India and China, where being born in cities such as Shanghai yields great advantage. But it was the occupation of one's mother that seemed to be the most descriptively important inherited characteristic in China, whereas the educational attainment of one's father played that role in the United States. Being Black in the US commands a significant (25%) opportunity penalty.

Finally, the granular estimation of quantile functions for each type inherent in this approach also allows us to investigate how the worst-off types – and the lower envelope of such types – are performing over time. In absolute levels, the average income of this lower envelope rose in the United States, China and Souh Africa, but fell in India. Relative to average incomes, this opportunity floor was much higher in China and the United States (at around 60% in the latest year), than in India and South Africa (35-40%).

Most of these insights into the extent and nature of the inheritance of inequality across generations, and of the distribution of opportunities across these four very different countries, were only possible through a comprehensive approach that incorporates many circumstance variables and looks beyond averages when assessing how predictive they are of observed living standards.

[28] At the same time, it is important to note that these larger ex-post estimates are not mechanical. As illustrated by the US case, it is possible that ex-ante partitions generate higher IOp estimates, for example when the higher power of tests on mean differences allow for a finer partition.

References

Asher, S., Novosad, P. and Rafkin, C. 2024. "Intergenerational mobility in India: New measures and estimates across time and social groups." *American Economic Journal: Applied Economics*, 16, 66–98.

Bloise, F., Brunori, P. and Piraino, P. 2021. "Estimating intergenerational income mobility on sub-optimal data: a machine learning approach." *Journal of Economic Inequality*, 19, 643–665.

Bourguignon, François, Francisco H. G. Ferreira, and Michael Walton. 2007. "Equity, Efficiency and Inequality Traps: A research agenda". *Journal of Economic Inequality*, 5: 235-256.

Brunori, Paolo, Francisco H. G. Ferreira, and Vito Peragine. 2021. "Prioritarianism and Equality of Opportunity." in Matthew Adler and Ole Norheim (ed.), *Prioritarianism in Practice*, Cambridge University Press.

Brunori, Paolo, Paul Hufe, and Daniel Gerszon Mahler. 2023. "The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees." *Scandinavian Journal of Economics*, https://onlinelibrary.wiley.com/doi/abs/10.1111/sjoe.12530.

Brunori, Paolo, Flaviana Palmisano, and Vito Peragine. 2019. "Inequality of Opportunity in Sub-Saharan Africa." *Applied Economics*, 51, 6428-6458.

Brunori, Paolo, Vito Peragine, and Laura Serlenga. 2019. "Upward and Downward Bias When Measuring Inequality of Opportunity." *Social Choice and Welfare*, 52, 635-661.

Buhmann, Brigitte, Lee Rainwater, Guenther Schmaus, and Timothy M. Smeeding. 1988. "Equivalence scales, well-being, inequality, and poverty: sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database." *Review of Income and Wealth*, 34, 115-142.

Chakravarty, Satya R., and Wolfgang Eichhorn. 1994. "Measurement of Income Inequality Observed Versus True Data." In Wolfgang Eichhorn (ed.), *Models and measurement of welfare and inequality*, Springer.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014. "Is the United States still a land of opportunity? Recent trends in intergenerational mobility." *American Economic Review*, 104, 141-147.

Checchi, Daniele, and Vito Peragine. 2010. "Inequality of Opportunity in Italy." *Journal of Economic Inequality*, 8, 429-450.

Ebert, Udo. 2010. "The decomposition of inequality reconsidered: Weakly decomposable measures." *Mathematical Social Sciences*, 60, 94-103.

Emran, Shahe, Francisco H. G. Ferreira, Yajing Jiang and Yan Sun (2023): "Occupational dualism and intergenerational educational mobility: evidence from China and India", *Journal of Economic Inequality,* 21: 743-773.

Fan, Y., Yi, J. and Zhang, J. 2021. "Rising intergenerational income persistence in China." *American Economic Journal: Economic Policy*, 13, 202–230.

Ferreira, Francisco H.G. and Paolo Brunori. 2024. "Inherited Inequality, Meritocracy, and the Purpose of Economic Growth", International Inequalities Institute Working Paper 147, LSE.

Ferreira, Francisco H.G., and Jérémie Gignoux. 2011. "The Measurement of Inequality of Opportunity: Theory and an Application to Latin America." *Review of Income and Wealth*, 57, 622-657.

Fleurbaey, Marc. 1994. "On fair compensation." *Theory and Decision*, 36, 277-307.

Fleurbaey, Marc. 2008. *Fairness, responsibility and welfare*. Oxford University Press.

Fleurbaey, Marc, and Vito Peragine. 2013. "Ex Ante Versus Ex Post Equality of Opportunity." *Economica*, 80, 118-130.

Foster, James and Artyom Shneyerov. 2000. "Path Independent Inequality Measures." *Journal of Economic Theory*, 91, 199-222.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, 15, 651-674.

Hothorn, Torsten, and Achim Zeileis. 2021. "Predictive Distribution Modeling Using Transformation Forests." *Journal of Computational and Graphical Statistics*, 30, 1181-1196.

Hufe, P., Kanbur, R. and Peichl, A. 2022. Measuring unfair inequality: Reconciling equality of opportunity and freedom from poverty. *The Review of Economic Studies*, *89*, 3345-3380.

Kopf, Julia, Thomas Augustin, and Carolin Strobl. 2013. "The Potential of Model-Based Recursive Partitioning in the Social Sciences: Revisiting Ockham's Razor." In *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Routledge.

Kundu, T. and Lefranc, A. 2020. "Inequality of opportunity in Indian society." *CSH-IFP Working Papers*, No. 14. Institut Français de Pondichéry / Centre de Sciences Humaines.

Lefranc, Arnaud, Nicolas Pistolesi, and Alain Trannoy. 2009. "Equality of Opportunity and Luck: Definitions and Testable Conditions, with an Application to Income in France." *Journal of Public Economics*, 93, 1189-1207.

Li Donni, Paolo, Juan Gabriel Rodriguez, and Pedro Rosa Dias. 2015. "Empirical Definition of Social Types in the Analysis of Inequality of Opportunity: A Latent Classes Approach." *Social Choice and Welfare*, 44, 673–701.

Mazumder, B. 2018. "Intergenerational mobility in the United States: What we have learned from the PSID." *Annals of the American Academy of Political and Social Science*, 680, 213–234.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine Learning: An applied econometric approach." *Journal of Economic Perspectives*, 31, 87-106.

Niehues, Judith and Andreas Peichl. 2014. “Upper bounds of inequality of oppoprtunity: theory and evidence for Germany and the US.” *Social Choice and Welfare*, 43, 73-99.

OECD. 2013. “OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth”, OECD Publishing, Paris

Palomino, Juan C., Gustavo A. Marrero, Brian Nolan, and Juan Gabriel Rodriguez. 2022. “Wealth inequality, intergenerational transfers and family background.” *Oxford Economic Papers*, 74, 643-670.

Piraino, Patrizio. 2015. “Intergenerational Earnings Mobility and Equality of Opportunity in South Africa.” *World Development*, 67, 396–405.

Pistolesi, N. 2009. "Inequality of opportunity in the land of opportunities, 1968–2001." *Journal of Economic Inequality*, 7, 411–433.

Rawls, John. 1971. *A Theory of Justice*. Harvard University Press.

Roemer, John E. 1993. “A pragmatic theory of responsibility for the egalitarian planner.” *Philosophy and Public Affairs*, 22, 146-166.

Roemer, John E. 1998. *Equality of Opportunity*. Harvard University Press.

Roemer, John E., and Alain Trannoy. 2016. “Equality of Opportunity: Theory and Measurement.” *Journal of Economic Literature* 54 (4): 1288–1332.

Shapley, Lloyd Sowell. 1953. “A value for n-person games.” In Harold Kuhn and Albert W. Tucker (ed.), *Contributions to the Theory of Games*, Princeton University Press.

Shorrocks, Anthony. 2013. “Decomposition procedures for distributional analysis: a unified framework based on the Shapley value.” *Journal of Economic Inequality*, 11, 99-126.

Solon, Gary. 1992. “Intergenerational income mobility in the United States.” *American Economic Review*, 82, 393-408.

Van De Gaer, Dirk. 1993. “Equality of Opportunity and Investment in Human Capital.” Ph.D. Dissertation, Katholieke Universiteit Leuven.

Wu, Dongjie. 2018. “Inequality of Opportunity in China.” Doctoral thesis, University of Queensland. https://core.ac.uk/download/pdf/189933279.pdf. Accessed on 09/01/2026.

Appendix 1: Technical details of the transformation tree algorithm

1A. The likelihood maximization using Bernstein polynomials

In practice, implementation of the likelihood maximization is facilitated by using a monotonic transformation function of y, $z = h(y)$, with $h'(y) > 0, \forall y$. Monotonicity ensures that $F(y) = F_z\big(h(y)\big)$. We follow Hothorn and Zeileis (2021) in using Bernstein polynomials of order M to construct the transformation function: $h(y) = a(y)^T\theta$. Note that a(y) is a polynomial of order M in y. The choice of M implies the choice of the dimension of the parameter vector, P=M+1. The higher that order, the greater the flexibility with which $F\left(y_{qc}, \theta(c)\right)$ can be modelled, and the greater the degree to which differences in their higher moments affect the partition and the estimation. Bernstein polynomials are a particular application of this transformation function, in which:

$$a_M(y) = \frac{\left(\phi_{1,M+1}(y), \dots, \phi_{M+1,1}(y)\right)}{M+1} \tag{A.1}$$

where $\phi_{m,M}$ denote the density of the Beta distribution with parameters m and M. Using this particular vector for the polynomial in $h(y)$ implies a simple log likelihood function that can be used for the maximization implicit in (5):

$$\ell_i(\theta) = \log[f_z(a(y)^T\theta)] + \log(a(y)^T\theta) \tag{A.2}$$

With this specific functional form for $\ell_i(\theta)$, all that is needed to solve Equations 14-15 (in the main text) and thus have the parameter estimates to model the conditional income distributions for all types in the tree terminal nodes is the algorithm to split the sample into types. This proceeds sequentially. Start from the case when $w_i(c) = 1, \forall i$. This corresponds to no splits: all observations are in a single bin, and have the same weight in the log likelihood maximization. The parameter estimates obtained under that assumption are the simple maximum likelihood estimates:

$$\hat{\theta}_{ML}^N(c) = \arg\max_{\theta \epsilon \Theta} \sum_{i=1}^{N} \ell_i(\theta) \tag{A.3}$$

To decide whether or not a split can improve prediction, test the null hypothesis:

$$H_0 : s\big(\hat{\theta}_{ML}^N | y\big) \perp C \tag{A.4}$$

where $s(\hat{\theta}|y)$ denotes the gradient contribution of observation i. For continuous distributions, the score contribution is simply the derivative of the log density with respect to θ. Differentiating (A.2) we obtain:

$$s(\hat{\theta}|y) = \mathrm{a}(y)\frac{f'_z(a(y)^T\theta)}{f_z(a(y)^T\theta)} + \frac{a'(y)}{a'(y)^T\theta} \qquad (\mathrm{A}.5)$$

There are a number of methods to test (A.4), and we follow Hothorn and Zeileis (2021) in using M-fluctuation tests. When these tests reject H_0, the algorithm implements a binary split in the circumstance x (an element of the vector c) that has the most significant association with the P x P score matrix, measured by the marginal multiplicity adjusted p-value (see Hothorn, Hornik, and Zeileis. 2006).

The algorithm is then repeated by testing hypotheses analogous to (A.4) in each of the resulting cells, and so on recursively, until H_0 can no longer be rejected. At this point, the algorithm has identified the optimal partition of the population into types: $\mathfrak{I} = \bigcup_{b=1,\dots B}\mathcal{B}_b$. Over this final partition, the likelihood function given by (A.2) and the weights given by (15) are used to solve (14), yielding the final parameter vector $\hat{\theta}^N(c)$, which fully characterizes the conditional distribution $F\left(y_{qc}, \theta(c)\right)$ in each type (terminal node) $\mathcal{B}_b$.

These parametric conditional distributions can then be inverted to yield the estimated type quantile functions $\tilde{y}_{qc} = F^{-1}\left(q,\ \hat{\theta}^N(c)\right)$.

1B: An illustration of the M-fluctuation test using made-up data

The algorithm employs an M-fluctuation test of parameter stability to determine node splits. Purely as an example, we show how the algorithm performs the type partition in a simplified hypothetical case in which father's occupation is the only circumstance and the logarithm of income is the outcome of interest.[29] The objective is testing whether the parameters defining the income distribution are significantly different when the population is split in two subgroups.

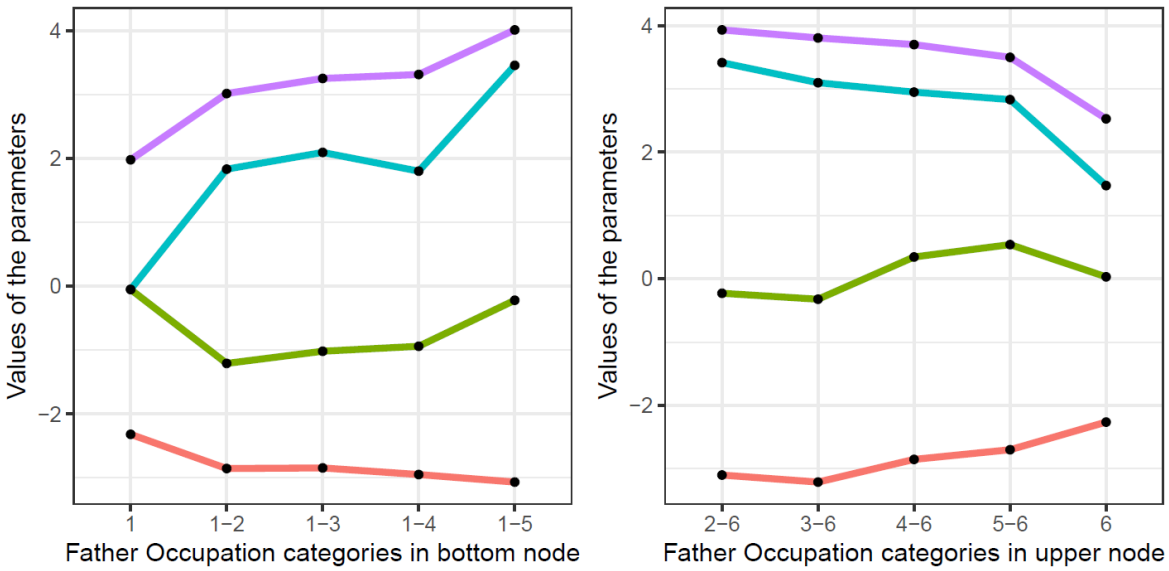
Following the steps described in the main text, we set a confidence level (α = 0.01) and, in order to obtain a graphical intuition of the instability of the parameters, a lower order of the polynomial (ω = 3), hence using four parameters to estimate the log-income distribution. We generate a mock dataset to split incomes according to father occupation, which takes 6

[29] Ours is a different version of a similar example proposed by Kopf, Augustin, and Strobl (2013).

categories ordered from smaller associated expected income to higher associated expected income.

In Figure B.1 below, we show the values of the parameters in the Bernstein polynomial associated with each split. Beginning from the left-hand side in both plots, the first four points represent the parameters associated with the nodes created when we split the population in two groups: those whose father occupation is 1 (right-hand plot) and the rest, that is, those whose father's occupation is 2 to 6 (left-hand plot). As we move to the right through the X-axis, we generate other splits, moving observations associated to categories in fathers' occupation from one node to the other, changing the resulting conditioned distributions. It is evident from Figure A.1 that, when transitioning observations from one terminal node to another, parameters undergo a change in magnitude. However, it is not immediately apparent which partition exhibits the most statistically significant parameter instability. That is, which occupational category should be selected as splitting point.

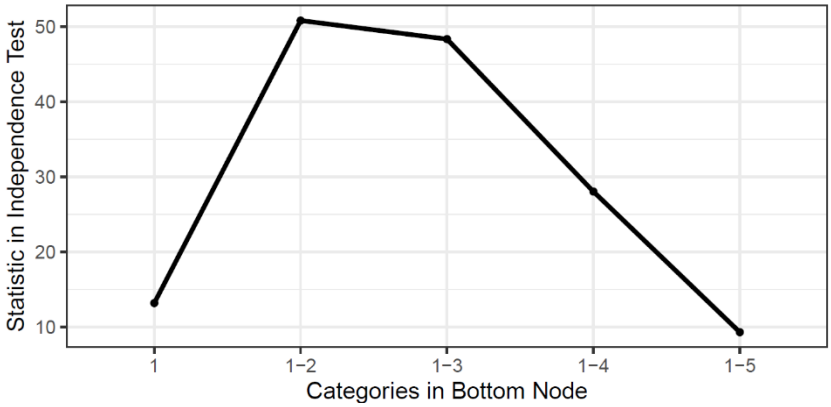
Figure A1. Values for the Parameters of the Bernstein Polynomial in each node



Source: Own Elaboration on NIDS 5

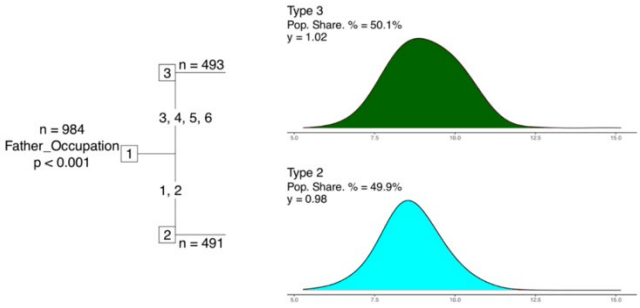
That selection is guided by the M-fluctuation test. Figure A.2 shows the value of the statistics for the tests described in step 4. The higher value (associated with a smaller p-value) is achieved when the bottom node has categories 1 and 2. That is the splitting point, as confirmed in Figure A.2. The population is thereby divided in two groups: those with father's occupation equal to 2 or less, and the rest, generating the simple tree in Figure A.3.

Figure A2. M-fluctuation quadratic test Statistics



Source: Own Elaboration on NIDS 5

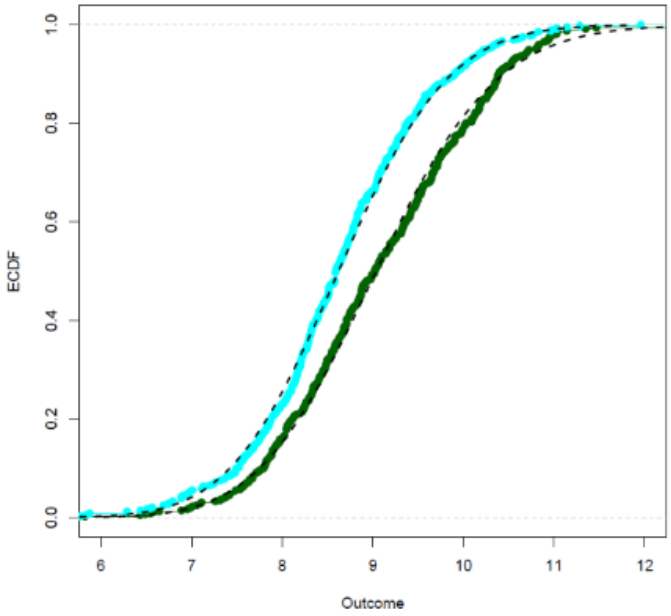
Figure A3. Transformation Tree (example)



Source: Own Elaboration on NIDS 5

This partition into two types allows us, for instance, to graphically explore Roemer's theory by plotting the cumulative density functions (CDF) of the outcome of interest by types (Figure A4). Here, the colored lines represent the empirical cumulative density functions (ECDF), while the dashed lines represent the interpolation of the distribution predicted with the polynomial approximation.

Figure A4. ECDFs (example)



Source: Own Elaboration

Appendix 2: Additional tables and figures

A. Additional tables

Table A1: Descriptive Income Statistics for previous waves

Country	Year	Mean	Gini	MLD	Top 1%	Top 10%	Bottom 40%
China	2010	4,878	0.496	0.475	0.109	0.397	0.119
China	2012	6,657	0.503	0.557	0.105	0.368	0.108
China	2014	6,168	0.483	0.526	0.154	0.410	0.101
China	2016	8,272	0.519	0.512	0.182	0.402	0.126
India	2005	5,903	0.499	0.457	0.104	0.414	0.092
South Africa	2008	11,383	0.647	0.839	0.067	0.376	0.118
South Africa	2012	10,974	0.617	0.711	0.075	0.330	0.131
South Africa	2015	10,764	0.574	0.6	0.087	0.331	0.137
USA	1974	35,535	0.299	0.155	0.037	0.205	0.268
USA	1976	36,104	0.293	0.152	0.031	0.192	0.276
USA	1978	36,230	0.294	0.152	0.037	0.196	0.273
USA	1980	34,390	0.328	0.192	0.064	0.222	0.251
USA	1982	32,762	0.317	0.179	0.037	0.207	0.253
USA	1984	36,046	0.345	0.214	0.057	0.227	0.245
USA	1986	36,627	0.345	0.210	0.047	0.220	0.248
USA	1988	41,626	0.389	0.270	0.085	0.267	0.221
USA	1990	39,470	0.365	0.236	0.058	0.249	0.227
USA	1992	40,078	0.376	0.259	0.076	0.269	0.216
USA	1994	39,161	0.379	0.269	0.073	0.272	0.215
USA	1996	39,837	0.367	0.256	0.065	0.284	0.197
USA	1998	42,838	0.392	0.296	0.084	0.286	0.195
USA	2000	44,494	0.384	0.268	0.069	0.287	0.195
USA	2002	45,158	0.395	0.297	0.095	0.302	0.190
USA	2004	46,249	0.406	0.308	0.077	0.284	0.194
USA	2006	45,494	0.431	0.369	0.082	0.299	0.184
USA	2008	45,110	0.403	0.305	0.070	0.276	0.197
USA	2010	43,192	0.393	0.296	0.065	0.268	0.197
USA	2012	44,111	0.386	0.286	0.057	0.254	0.203
USA	2014	45,608	0.395	0.290	0.052	0.250	0.200
USA	2016	48,057	0.393	0.315	0.054	0.271	0.180

Note: Income units are in 2017 US dollars at PPP exchange rates. MLD stands for Mean Log Deviation. The three columns on the left represent the share of income received by the Top 1%, the Top 10%, and the Bottom 40% in the income distribution. Source: CFPS, IHDS, NIDS, and PSID.

Table A2: Ex-post IOp estimates for previous waves

Country	Year	Gini ($\hat{y}_T$)	IOR_T (Gini)	MLD ($\hat{y}_T$)	IOR_T (MLD)	Top 1%	Top 10%	Bottom 40%
China	2010	0.266	53.6%	0.133	28.0%	0.046	0.292	0.247
China	2012	0.218	43.3%	0.124	22.3%	0.049	0.237	0.233
China	2014	0.207	42.9%	0.104	19.8%	0.058	0.238	0.288
China	2016	0.293	56.5%	0.209	40.8%	0.051	0.287	0.246
India	2005	0.306	61.3%	0.178	39.0%	0.046	0.278	0.172
South Africa	2008	0.533	82.4%	0.534	63.6%	0.037	0.279	0.182
South Africa	2012	0.479	77.6%	0.405	57.0%	0.032	0.237	0.231
South Africa	2015	0.385	67.1%	0.248	41.3%	0.024	0.193	0.278
USA	1970	0.137	45.7%	0.032	20.1%	0.013	0.122	0.401
USA	1972	0.139	45.7%	0.034	20.5%	0.014	0.121	0.402
USA	1974	0.125	41.8%	0.027	17.4%	0.011	0.108	0.409
USA	1976	0.117	39.9%	0.024	15.8%	0.013	0.110	0.429
USA	1978	0.117	39.8%	0.024	15.8%	0.011	0.109	0.428
USA	1980	0.141	43.0%	0.035	18.2%	0.018	0.122	0.411
USA	1982	0.130	41.0%	0.030	16.8%	0.013	0.114	0.418
USA	1984	0.138	40.0%	0.034	15.9%	0.013	0.116	0.414
USA	1986	0.147	42.6%	0.037	17.6%	0.012	0.115	0.411
USA	1988	0.164	42.2%	0.048	17.8%	0.016	0.126	0.393
USA	1990	0.147	40.3%	0.037	15.7%	0.013	0.121	0.397
USA	1992	0.153	40.7%	0.043	16.6%	0.013	0.113	0.383
USA	1994	0.178	47.0%	0.081	30.1%	0.018	0.130	0.377
USA	1996	0.162	44.1%	0.051	19.9%	0.017	0.139	0.342
USA	1998	0.195	49.7%	0.086	29.1%	0.025	0.141	0.332
USA	2000	0.158	41.1%	0.046	17.2%	0.018	0.142	0.348
USA	2002	0.192	48.6%	0.079	26.6%	0.024	0.148	0.330
USA	2004	0.200	49.3%	0.090	29.2%	0.018	0.139	0.331
USA	2006	0.192	44.5%	0.108	29.3%	0.027	0.145	0.372
USA	2008	0.180	44.7%	0.076	24.9%	0.018	0.133	0.371
USA	2010	0.168	42.7%	0.057	19.3%	0.018	0.127	0.362
USA	2012	0.180	46.6%	0.113	39.5%	0.025	0.141	0.378
USA	2014	0.165	41.8%	0.054	18.6%	0.013	0.121	0.391
USA	2016	0.175	44.5%	0.080	25.4%	0.024	0.151	0.339

Note: The three columns on the left represent the share of $\hat{y}_T$ received by the Top 1%, the Top 10%, and the Bottom 40% in the $\hat{y}_T$ distribution. Source: CFPS, IHDS, NIDS, and PSID.

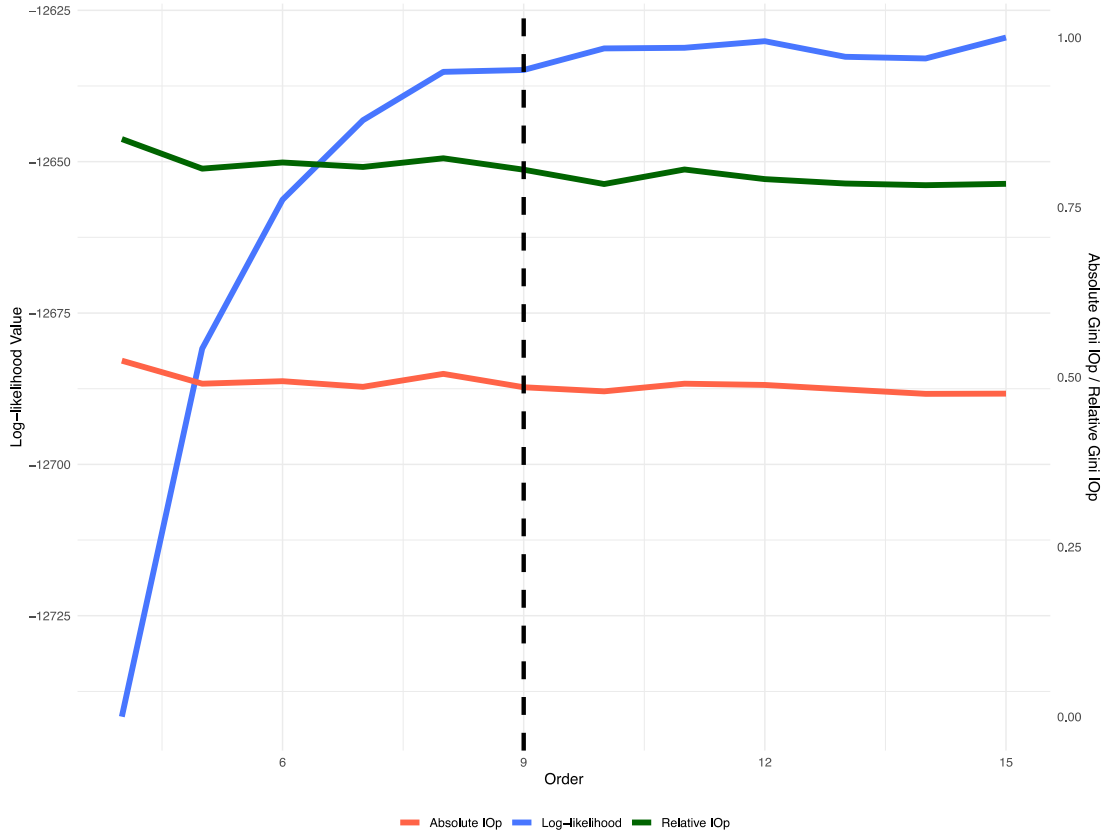
Table A3: Ex-ante IOp estimates for previous waves

Country	Year	Gini ($\hat{y}_{BHM}$)	IOR_{EA} (Gini)	MLD ($\hat{y}_{BHM}$)	IOR_{EA} (MLD)	Top 1%	Top 10%	Bottom 40%
China	2010	0.227	45.8%	0.082	17.3%	0.033	0.251	0.297
China	2012	0.176	35.0%	0.052	9.3%	0.037	0.220	0.308
China	2014	0.197	40.8%	0.066	12.5%	0.100	0.259	0.23
China	2016	0.205	39.5%	0.067	13.1%	0.069	0.248	0.308
India	2005	0.267	53.5%	0.111	24.3%	0.045	0.260	0.182
South Africa	2008	0.479	74.0%	0.385	45.9%	0.022	0.254	0.231
South Africa	2012	0.430	69.7%	0.301	42.3%	0.025	0.205	0.257
South Africa	2015	0.350	61.0%	0.204	34.0%	0.04	0.250	0.251
USA	1970	0.135	45.0%	0.030	18.9%	0.014	0.121	0.408
USA	1972	0.139	45.7%	0.032	19.3%	0.015	0.123	0.403
USA	1974	0.131	43.8%	0.028	18.1%	0.014	0.120	0.416
USA	1976	0.120	41.0%	0.024	15.8%	0.013	0.114	0.429
USA	1978	0.118	40.1%	0.023	15.1%	0.013	0.113	0.437
USA	1980	0.132	40.2%	0.029	15.1%	0.017	0.120	0.425
USA	1982	0.137	43.2%	0.031	17.3%	0.014	0.116	0.418
USA	1984	0.146	42.3%	0.035	16.4%	0.014	0.117	0.413
USA	1986	0.155	44.9%	0.039	18.6%	0.015	0.123	0.409
USA	1988	0.173	44.5%	0.051	18.9%	0.022	0.136	0.387
USA	1990	0.159	43.6%	0.042	17.8%	0.015	0.126	0.394
USA	1992	0.156	41.5%	0.041	15.8%	0.015	0.128	0.387
USA	1994	0.161	42.5%	0.043	16.0%	0.015	0.129	0.391
USA	1996	0.164	44.7%	0.045	17.6%	0.015	0.139	0.342
USA	1998	0.170	43.4%	0.050	16.9%	0.021	0.137	0.354
USA	2000	0.170	44.3%	0.047	17.5%	0.016	0.142	0.347
USA	2002	0.189	47.8%	0.059	19.9%	0.026	0.158	0.335
USA	2004	0.175	43.1%	0.051	16.6%	0.018	0.136	0.366
USA	2006	0.177	41.1%	0.051	13.8%	0.019	0.141	0.375
USA	2008	0.165	40.9%	0.045	14.8%	0.015	0.128	0.386
USA	2010	0.167	42.5%	0.045	15.2%	0.018	0.133	0.378
USA	2012	0.170	44.0%	0.047	16.4%	0.017	0.131	0.387
USA	2014	0.164	41.5%	0.044	15.2%	0.014	0.119	0.394
USA	2016	0.159	40.5%	0.040	12.7%	0.017	0.133	0.359

Note: IOR_{EA} is the Brunori, Hufe and Mahler (2023) estimate, and $\hat{y}_{BHM}$ denotes the incomes predicted by circumstances using that ex-ante method. So*urce: CFPS, IHDS, PSID, and NIDS.*

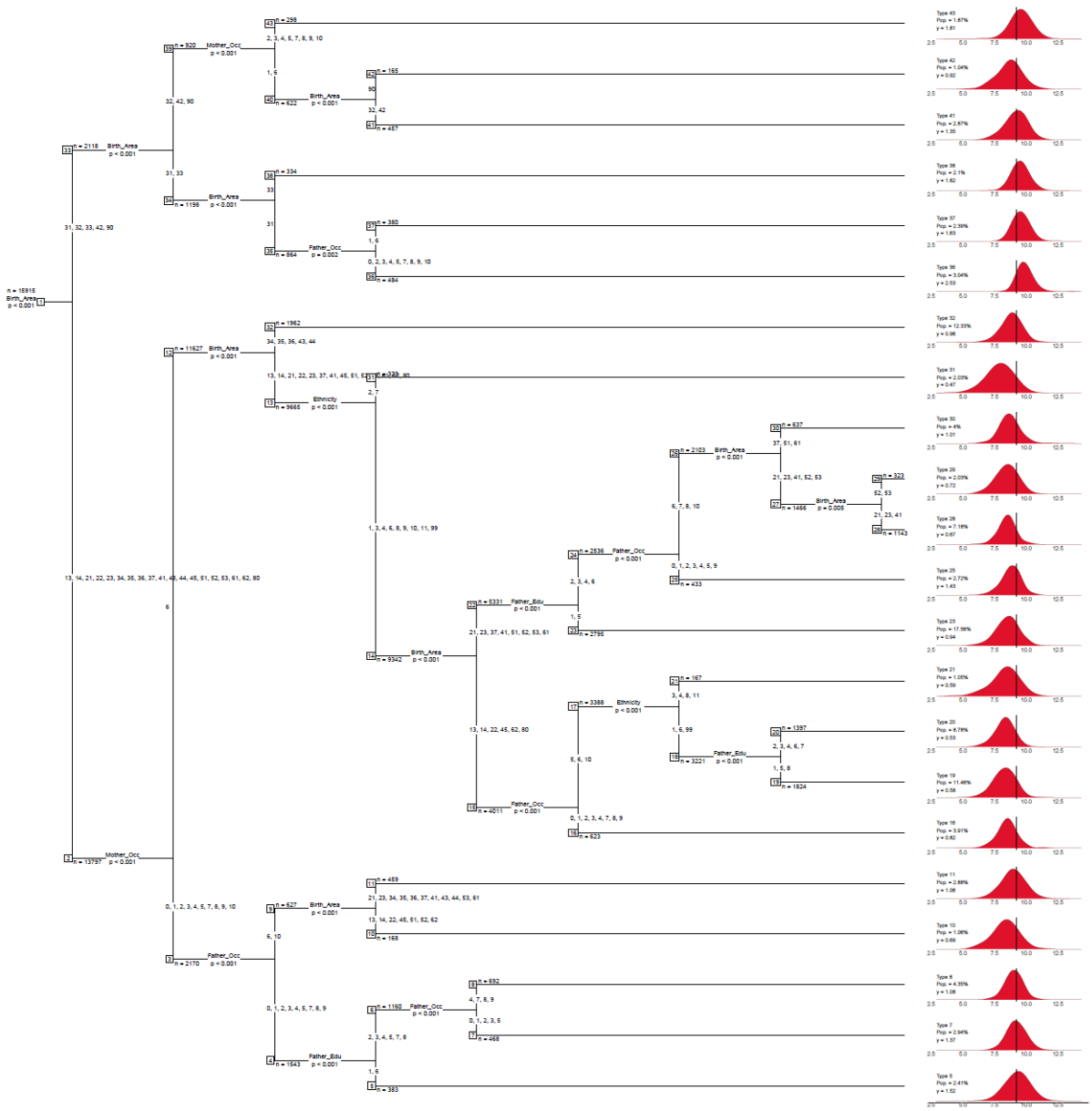
B: Additional figures

Figure B1: Sensitivity of ex-post IOp to the Bernstein polynomial order in South Africa



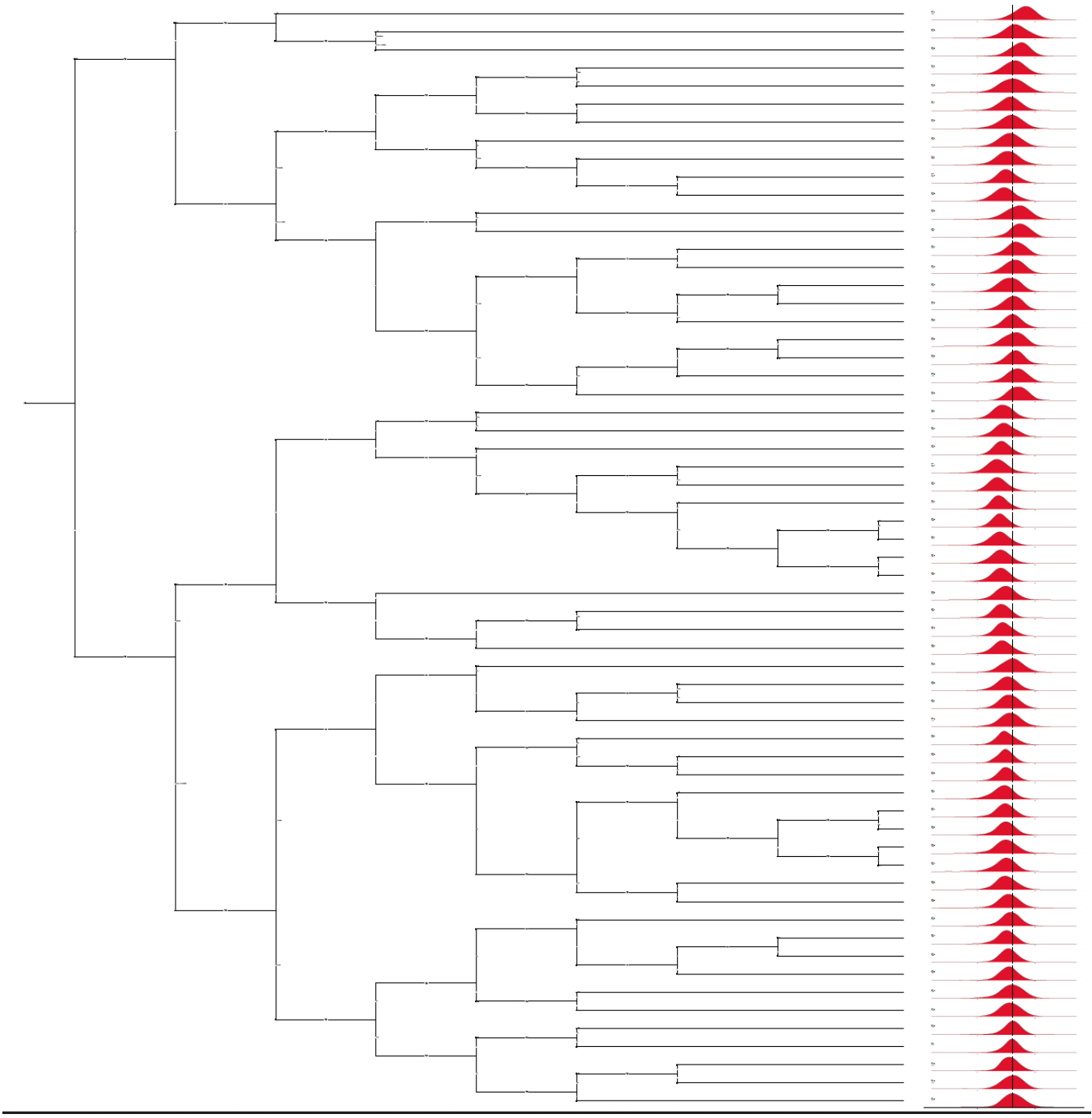
Note: The plot shows the log-likelihood value (blue line) associated with different Bernstein polynomial order values. We also show absolute IOp (red line) and relative IOp (green line) estimated with different Bernstein polynomial order values. Source: *NIDS (2017).*

Figure B2a: Full Transformation Tree for China (2018)



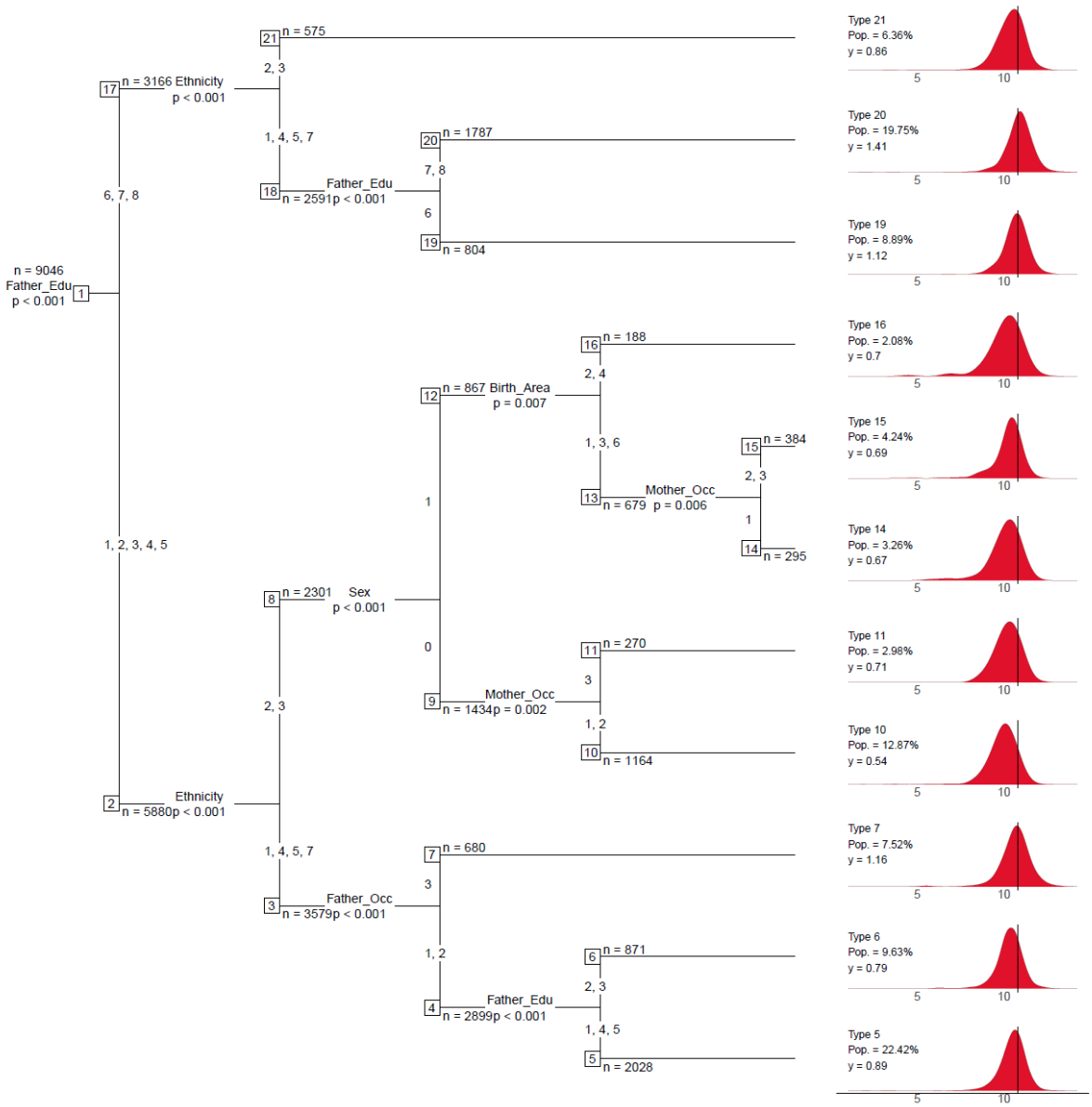
Note: Splitting nodes show their sample size and the p-value associated to the splitting. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (1 Han, 2 Mongol, 3 Hui, 4 Tibetan, 5 Miao, 7 Yi, 8 Zhuang, 9 Bouyei, 10 Korean, 11 Manchu, 99 Other), Birth Area (13 Hebei, 14 Shanxi, 21 Liaoning, 22 Jilin, 23 Heilongjiang, 31 Shanghai, 32 Jiangsu, 33 Zhejiang, 34 Anhui, 35 Fujian, 36 Jiangxi, 37 Shandong, 41 Henan, 42 Hubei, 43 Hunan, 44 Guangdong, 45 Guangxi Zhuang Autonomous Region, 51 Sichuan, 52 Guizhou, 53 Yunnan, 61 Shaanxi, 62 Gansu, 80 Not available, 90 Other), Parent's education (1 Illiterate/Semi-literate, 2 Primary school, 3 Junior high school, 4 Senior high school/secondary school/technical school/vocational senior school, 5 3-year college, 6 4-year college/Bachelor's degree, 7 Master's degree, 8 Doctoral degree), Parent's occupation (0 Armed forces, 1 Managers, 2 Professionals, 3 Technicians and Associate professionals, 4 Clerks, 5 Services and Sales workers, 6 Agricultural, Forestry and Fishery workers, 7 Craft and trade workers, 8 Plant and machine operators and assemblers, 9 Elementary occupations, 10 Unemployed). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: CFPS (2018).

Figure B2b: Full Transformation Tree for India (2012)



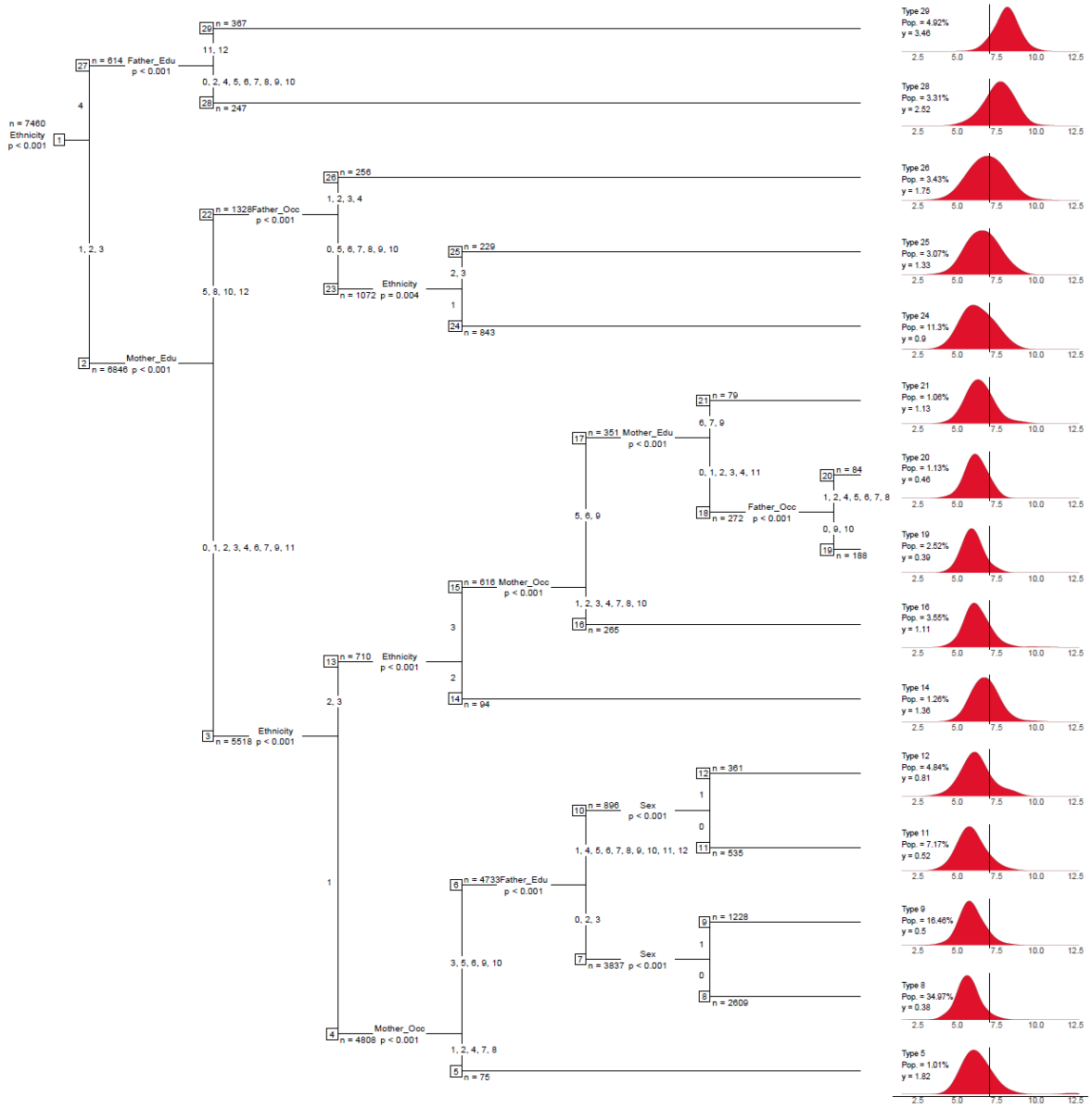
Note: Splitting nodes show their sample size and the p-value associated to the splitting. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (2 Forward caste, 3 Other Backward castes (OBC), 4 Dalit, 5 Adivasi, 6 Muslim, 7 Christian, Sikh, Jain), Parent's Education (0 None, 1 Incomplete primary, 2 Complete primary, 3 Incomplete secondary, 4 Complete secondary, 5 Higher secondary, 6 Post-secondary or higher), Birth Area (1 Jammu & Kashmir, 2 Himachal Pradesh, 3 Punjab, 4 Another State, 5 Uttarakhand, 6 Haryana, 7 Delhi, 8 Rajasthan, 9 Uttar Pradesh, 10 Bihar, 11 Overseas, 18 Northeast, 19 West Bengal, 20 Jharkhand, 21 Orissa, 22 Chhattisgarh, 23 Madhya Pradesh, 24 Gujarat, 27 Maharashtra, 28 Andhra Pradesh, 29 Karnataka, 32 Kerala, 33 Tamil Nadu). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: IHDS (2012).

Figure B2c: Full Ex-post/Transformation Tree for USA (2018)



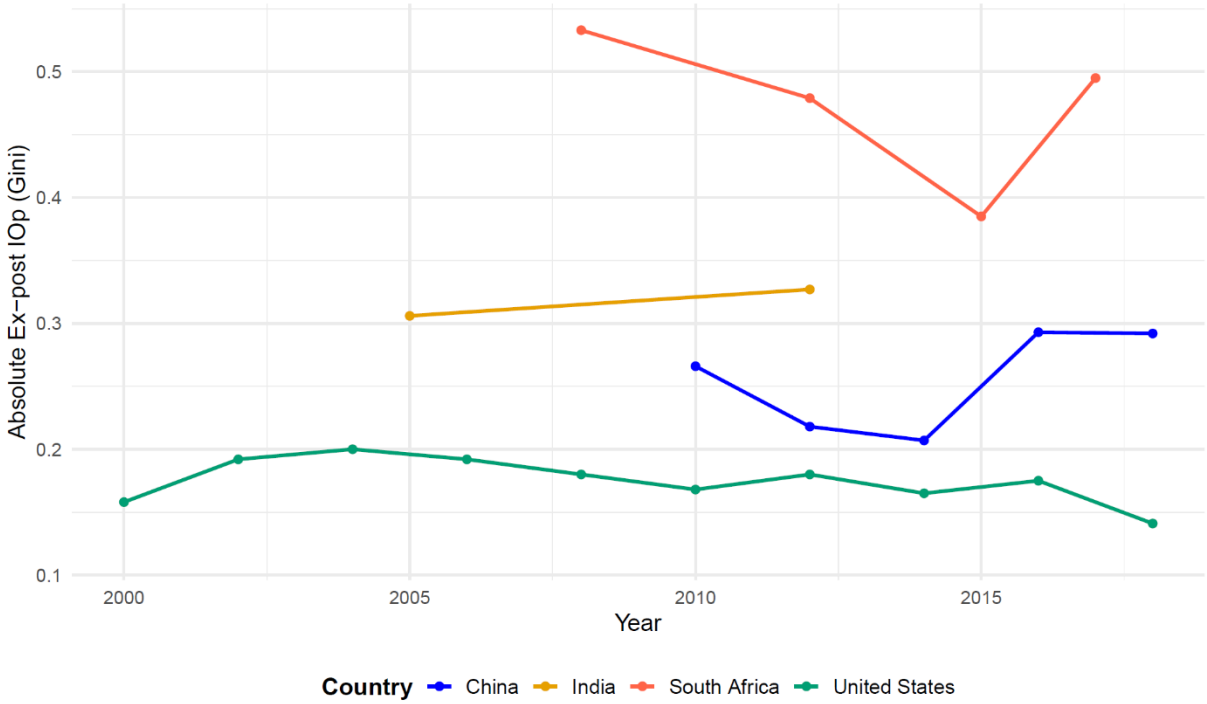
Note: Splitting nodes show their sample size and the p-value associated to the splitting. Circumstance categories are Gender (0 Male, 1 Female), Ethnicity (1 White, 2 Black, 3 American Indian/Aleut/Eskimo, 4 Asian/Pacific Islander, 5 Hispanic, 7 Other), Region of upbringing (1 Northeast, 2 North Central, 3 South, 4 West, 5 Alaska/Hawaii, 6 Foreign country), Parents' education (1 0–5 Grades, 2 6–8 Grades, 3 9–11 Grades, 4 High school, 5 12+ Grades + non-academic training, 6 Some college, 7 College degree, 8 Advanced college degree), and Parents' occupation (ISCO) (1 Basic, 2 Middle, 3 High). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: Own elaboration from the PSID (2018).

Figure B2d: Full Ex-post/Transformation Tree for South Africa (2017)



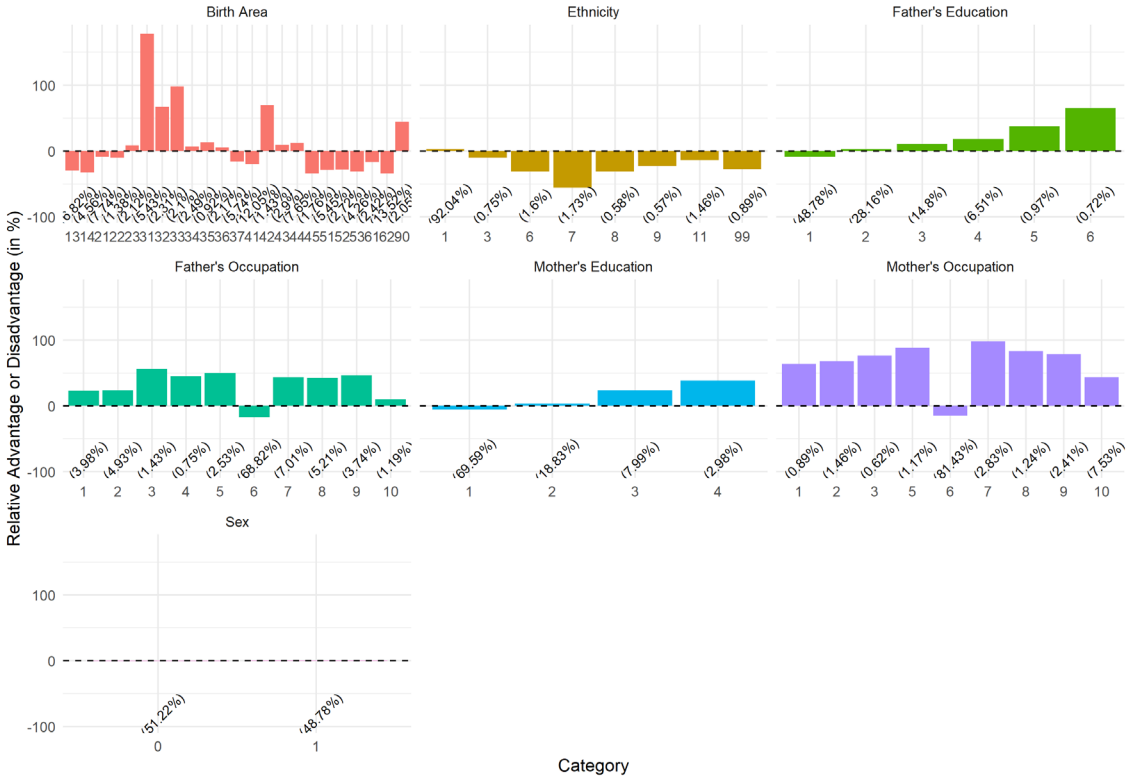
Note: Splitting nodes show their sample size and the p-value associated to the splitting. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (1 African, 2 Asian/Indian, 3 Coloured, 4 White), Parent's Education (0 Zero, 1 Grade 1, 2 Grade 2, 3 Grade 3, 4 Grade 4, 5 Grade 5, 6 Grade 6, 7 Grade 7, 8 Grade 8, 9 Grade 9, 10 Grade 10, 11 Grade 11, 12 Grade 12), Parent's Occupation (0 Military, 1 Managers, 2 Professionals, 3 Technicians and Professionals, 4 Clerical Support, 5 Service and sales, 6 Farm, Forest, Fishery, 7 Craft, 8 Operators, 9 Elementary, 10 Others). The panels on the right display the log-density of type-specific incomes. The labels indicate the share of the population each type represents (Pop.) and their average income relative to the overall sample mean (y = 1), which is also depicted as a vertical black line in the log-density plot. Source: NIDS (2017).

Figure B3: Time Trends of Inequality of Opportunity (Ex-post, TrT)



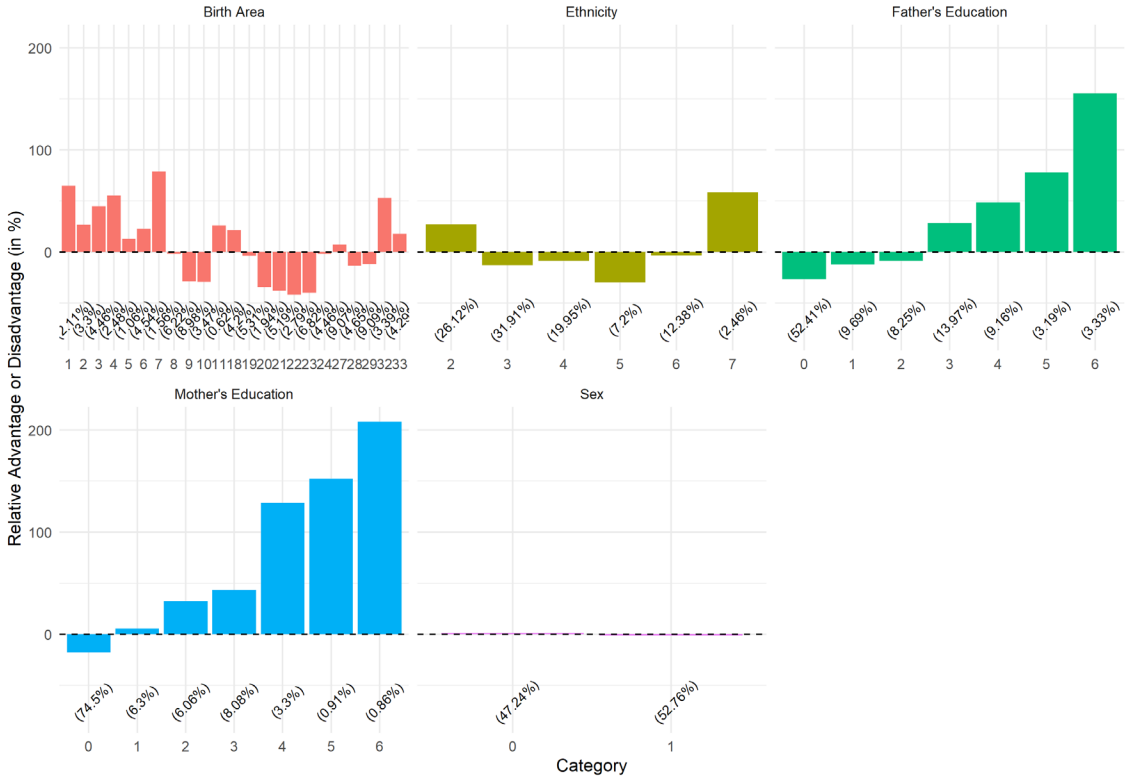
Source: CFPS, IHDS, PSID, and NIDS.

Figure B4: Marginal effect of circumstances on opportunities: China



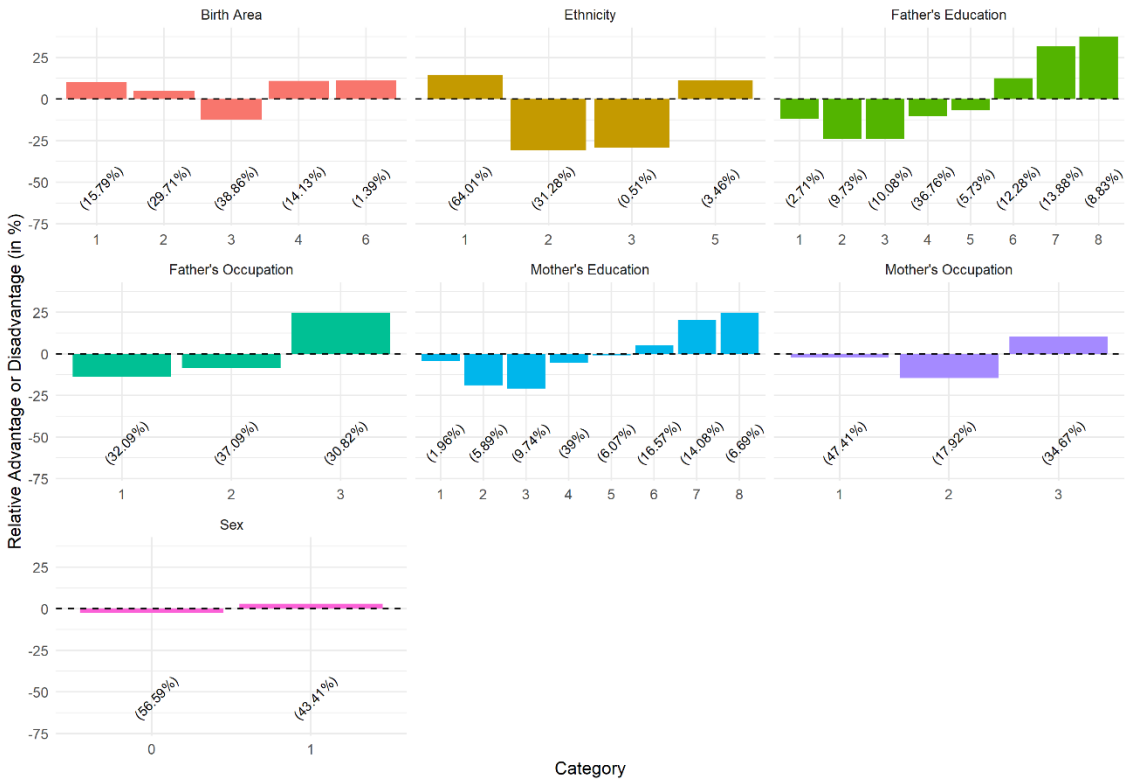
Note: Values on the y-axis represent the relative advantage or disadvantage associated with each category, computed as $100 \times \frac{\textit{average income (category)}}{\textit{average income (sample)}}$*. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (1 Han, 2 Mongol, 3 Hui, 4 Tibetan, 5 Miao, 7 Yi, 8 Zhuang, 9 Bouyei, 10 Korean, 11 Manchu, 99 Other), Birth Area (13 Hebei, 14 Shanxi, 21 Liaoning, 22 Jilin, 23 Heilongjiang, 31 Shanghai, 32 Jiangsu, 33 Zhejiang, 34 Anhui, 35 Fujian, 36 Jiangxi, 37 Shandong, 41 Henan, 42 Hubei, 43 Hunan, 44 Guangdong, 45 Guangxi Zhuang Autonomous Region, 51 Sichuan, 52 Guizhou, 53 Yunnan, 61 Shaanxi, 62 Gansu, 80 Not available, 90 Other), Parent's education (1 Illiterate/Semi-literate, 2 Primary school, 3 Junior high school, 4 Senior high school/secondary school/technical school/vocational senior school, 5 3-year college, 6 4-year college/Bachelor's degree, 7 Master's degree, 8 Doctoral degree), Parent's occupation (0 Armed forces, 1 Managers, 2 Professionals, 3 Technicians and Associate professionals, 4 Clerks, 5 Services and Sales workers, 6 Agricultural, Forestry and Fishery workers, 7 Craft and trade workers, 8 Plant and machine operators and assemblers, 9 Elementary occupations, 10 Unemployed). Numbers in parentheses denote population shares in each category. We are not showing categories populated by less than 0.5% of the sample size. Source: CFPS (2018).*

Figure B5: Marginal effect of circumstances on opportunities: India



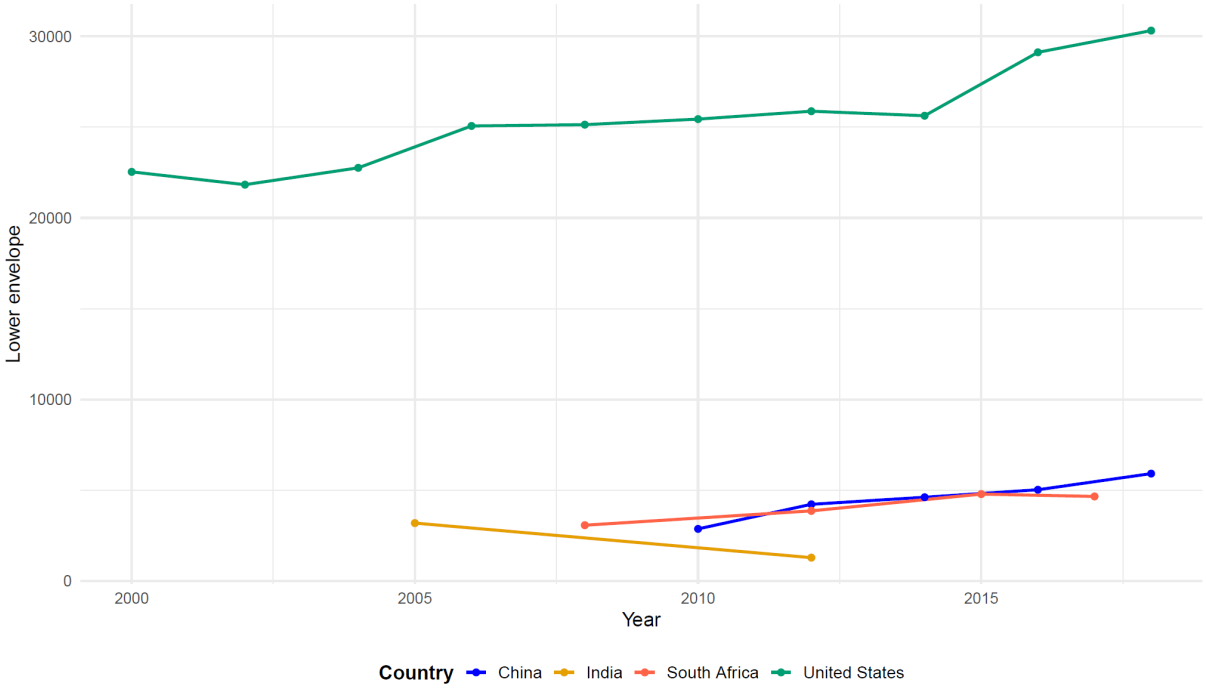
Note: Values on the y-axis represent the relative advantage or disadvantage associated with each category, computed as $100 \times \frac{\textit{average income (category)}}{\textit{average income (sample)}}$*. Circumstance categories are Gender (0 Female, 1 Male), Ethnicity (2 Forward caste, 3 Other Backward castes (OBC), 4 Dalit, 5 Adivasi, 6 Muslim, 7 Christian, Sikh, Jain), Parent's Education (0 None, 1 Incomplete primary, 2 Complete primary, 3 Incomplete secondary, 4 Complete secondary, 5 Higher secondary, 6 Post-secondary or higher), Birth Area (1 Jammu & Kashmir, 2 Himachal Pradesh, 3 Punjab, 4 Another State, 5 Uttarakhand, 6 Haryana, 7 Delhi, 8 Rajasthan, 9 Uttar Pradesh, 10 Bihar, 11 Overseas, 18 Northeast, 19 West Bengal, 20 Jharkhand, 21 Orissa, 22 Chhattisgarh, 23 Madhya Pradesh, 24 Gujarat, 27 Maharashtra, 28 Andhra Pradesh, 29 Karnataka, 32 Kerala, 33 Tamil Nadu). Numbers in parentheses denote population shares in each category. We are not showing categories populated by less than 0.5% of the sample size. Source: IHDS (2012).*

Figure B6: Marginal effect of circumstances on opportunities: United States



Note: Values on the y-axis represent the relative advantage or disadvantage associated with each category, computed as $100 \times \frac{\textit{average income (category)}}{\textit{average income (sample)}}$*. Circumstance categories are Gender (0 Male, 1 Female), Ethnicity (1 White, 2 Black, 3 American Indian/Aleut/Eskimo, 5 Hispanic), Region of upbringing (1 Northeast, 2 North Central, 3 South, 4 West, 6 Foreign country), Parents' education (1 0–5 Grades, 2 6–8 Grades, 3 9–11 Grades, 4 High school, 5 12+ Grades + non-academic training, 6 Some college, 7 College degree, 8 Advanced college degree), and Parents' occupation (ISCO) (1 Basic, 2 Middle, 3 High). Numbers in parentheses denote population shares in each category. We are not showing categories populated by less than 0.5% of the sample size. Source: PSID (2018)*

Figure B7: Lower envelope



Source: CFPS (2018), IHDS (2012), PSID (2018), and NIDS (2017). Monetary values in $2017.