LSE International Inequalities Institute

WORKING PAPER 107

Inherited inequality: a general framework and an application to South Africa

Paolo Brunori
International Inequalities Institute; Università di Firenze

Francisco H. G. Ferreira
International Inequalities Institute; IZA

Pedro Salas-Rojo
International Inequalities Institute; EQUALITAS

SEPTEMBER 2023

This paper is superseded by III Working Paper No. 160. Please use that version instead.

Paolo Brunori
International Inequalities Institute; Università di Firenze

Francisco H. G. Ferreira
International Inequalities Institute; IZA

Pedro Salas-Rojo
International Inequalities Institute; EQUALITAS

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

For further information on the work of the Institute, please contact the Institute Manager, Liza Ryan at e.ryan@lse.ac.uk

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
@LSEInequalities

Inherited inequality

A general framework and an application to South Africa

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

Abstract: Scholars have sought to quantify the extent of inequality which is inherited from past generations in many different ways, including a large body of work on intergenerational mobility and inequality of opportunity. This paper makes three contributions to that broad literature. First, we show that many of the most prominent approaches to measuring mobility or inequality of opportunity fit within a general framework which involves, as a first step, a calculation of the extent to which inherited circumstances can predict current incomes. The importance of prediction has led to recent applications of machine learning tools to solve the model selection challenge in the presence of competing upward and downward biases. Our second contribution is to apply transformation trees to the computation of inequality of opportunity. Because the algorithm is built on a likelihood maximization that involves splitting the sample into groups with the most salient differences between their conditional cumulative distributions, it is particularly well-suited to measuring ex-post inequality of opportunity, following Roemer (1998). Our third contribution is to apply the method to data from South Africa, arguably the world's most unequal country, and find that almost three-quarters of its current inequality is inherited from predetermined circumstances, with race playing the largest role, but parental background also making an important contribution.

Keywords: Inequality; opportunity; mobility; transformation trees; South Africa.

JEL Codes: D31, D63, J62

---

[1] All three authors are with the International Inequalities Institute at the London School of Economics. Brunori is also at Università di Firenze, Ferreira is also affiliated with IZA, and Salas-Rojo is also affiliated with EQUALITAS. Correspondence to Pedro Salas-Rojo: p.salas-rojo@lse.ac.uk. We are grateful to Torsten Hothorn and Achim Zeileis for very helpful advice, and to Pedro Torres for superb research assistance. We also thank seminar participants at the London School of Economics and the Universities of Bari, Bicocca, la Laguna, and Tilburg for useful comments. All errors are ours.

1 Introduction

People's educational and professional achievements, their incomes, and their wealth are generally not independent of their backgrounds. Various attributes that are determined at or before birth – such as biological sex; race or ethnicity; parental income and other aspects of family background - are powerful predictors of a person's own economic outcomes later in life.

Economists have typically considered this an important fact: a copious literature has sought to quantify the extent to which inherited or pre-determined characteristics shape people's life outcomes, and to compare results across societies or over time. There is very little in that literature that attempts to disentangle the multiple causal pathways or to estimate full structural models with behavioral parameters, because it was quickly understood that the identification problems are almost insurmountable.[2]

Although there are obviously multiple studies that seek to estimate the causal effects of *specific* characteristics, - say, race or gender – on specific outcomes – say, wages or job interviews – all of the dominant approaches used to quantify the *overall extent* to which the variation in, say, current incomes reflects the effects of inherited factors, have been descriptive. These approaches include the literatures on intergenerational mobility; inequality of opportunity; and sibling correlations.

This paper contributes to that broad literature in three ways. First, we note that all of these descriptive approaches rely on using observed inherited characteristics 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 incomes, and show that a wide range of measures in current use are special cases.

---

[2] Parental education, for example, will generally affect both the quantity and quality of the parent's time inputs into the child's development at home. It will also affect, or interact with, school choice and neighborhood location, each of which are likely to have their own separate effects. It may also affect the child's employment and marriage (or household formation more broadly) opportunities later in life. Some of these effects of parental education will operate through parental income, others will operate directly. They will potentially operate differently across sexes, races or castes. They will likely interact with family wealth, separately from parental income. They may be confounded with genetic endowments, which are also transmitted separately. And so on. See Haveman and Wolfe (1995) for a classic discussion of (some of) these multiple pathways.

Once the central role of prediction is recognized, it is natural to consider options among modern data-driven (or 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, conditional inference trees, random forests, and transformation trees are three highly promising approaches.

Since conditional inference trees and random forests have already been used in this context (see Brunori, Hufe, and Mahler, 2023), we focus on transformation trees, which have recently been developed by Hothorn and Zeileis (2021). Our second contribution is thus to show that, because this approach provides a powerful algorithm to predict not only means, but full conditional distributions for different population subgroups, it is particularly well-suited to inequality decompositions that depend on differences in higher moments of the income distribution between subgroups, e.g., "ex-post" inequality of opportunity (Ex-post IOp). It also provides an optimal solution – in a well-defined statistical sense – to a problem that has bedeviled the literature(s) so far, namely the choice of model specification.

To the best of our knowledge, ex-post IOp was first used empirically to estimate the share of inequality predicted by inherited circumstances by Checchi and Peragine (2010), with an application to Italy.[3] But it draws on a rich theoretical tradition in normative economics that argues that equal opportunities are achieved when all individuals who exert the same degree of effort or responsibility can ultimately achieve the same outcomes, regardless of inherited circumstances (see, e.g., Roemer, 1993, 1998; Fleurbaey, 1994, 2008). Under some assumptions, the theory suggests, the appropriate degree of effort, once cleansed of the effects of circumstances, can be proxied by the relative position – that is, the quantile – of an individual in the income distribution of those that have the same inherited circumstances as she does – her "type". (Roemer, 1998).

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 severely limited its use in practice. Our proposed approach can significantly alleviate these challenges. That said, the attractiveness of the approach does not require adherence to the specific normative views embodied in the theoretical literature. Our results can also be interpreted in the

[3] See also Lefranc, Pistolesi, and Trannoy (2009).

spirit of alternative inequality decompositions, in which the between-groups term is not independent of within-group inequality.[4] In our third contribution, we apply the method to South Africa, arguably the world's most unequal country. We present the full decomposition, including the type-specific conditional distributions. We find that more than 70% of the country's Gini coefficient of 0.6 is accounted for by inherited circumstances.

The paper proceeds as follows. The next section describes a general framework for the estimation of the importance 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 empirical challenges faced by these approaches, focusing on model selection. Section 4 then introduces our own approach to estimating ex-post IOp using transformation trees as another case, and describes its operation.

Section 5 describes our data and Section 6 presents results. These results include not only estimates of the share of current inequality in South Africa which are predicted by a set of inherited circumstances, but also (i) a schematic description of the population partition which generates the most salient cleavages in South African society (again, in a well-defined statistical sense); (ii) estimates of the conditional cumulative distributions by 'type' (or population sub-group); (iii) the implied decomposition of the density function into a mixture of these sub-group distributions; (iv) a Shapley-Shorrocks decomposition of the relative (predictive) importance of individual circumstances in the overall decomposition; and (v) an estimate of the lower-envelope of the decomposition, which corresponds to the maximand in Roemer's original policy objective. It also compares our ex-post IOp results to ex-ante estimates from conditional inference trees and forests, as in Brunori, Hufe and Mahler (2023). Taken together, this set of analytical and visualization methods represent complementary tools that enable a deeper understanding of inequality of opportunity. Section 7 concludes.

[4] 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.

2. Inherited inequality: a 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; and a set of inherited characteristics, which we call circumstances (following Roemer, 1998). For individual *i*, these are represented by a k-dimensional vector $\boldsymbol{c_i}$. In general, many people may share the same vector of circumstances, and each of those groups is called a "type". The population can then be exhaustively divided into a set of types, $C = \{\tau_1, \dots, \tau_m, \dots, \tau_M\}$, where $\tau_m := \{\forall i | \boldsymbol{c_i} = \boldsymbol{c_m}\}$, such that $\bigcup_1^M \tau_m = \mathcal{N}$ and $\bigcap_1^M \tau_m = \emptyset$. $C \in \mathbb{C}$, the set of all possible partitions.

A situation in which there is no inherited inequality is one in which the joint distribution {y, **c**} is characterized by $y \perp \boldsymbol{c}$. In that case, there is obviously no difference between the conditional income distributions obtained from that joint distribution:

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

If (1) does not hold, then the associations between the vector **c** and y across the population imply that the 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.

It is straightforward to see that most methods for estimating the intergenerational transmission of advantage currently in use revolve around estimating models of the general form (2), using different functions in the set of possible functions $\mathcal{F}$. In addition, in many cases the final estimates are summarized by a comparison of inequality in current-generation income, $I(y)$ and inequality in the distribution of the incomes predicted by the inherited circumstance: $\hat{y} = \hat{f}(c)$, $I(\hat{y})$. In other words, measures of mobility or inequality of opportunity are often of the form $O = g\big(I(\hat{y}), I(y)\big)$.

Intergenerational mobility

What generally distinguishes estimates of intergenerational mobility is the assumption that there is a single circumstance, namely the previous generation value of *y*, y_p.[5] Then $f(\boldsymbol{c}, \varepsilon)$ may for example take the form:

$$\mathrm{y} = f_M(c, \varepsilon) = e^{\alpha + \beta \log y_p + \varepsilon} \tag{3}$$

Taking logarithms, equation (3) becomes the standard Galtonian regression that has been the workhorse of intergenerational mobility estimates from Solon (1992) to Chetty et al. (2014). Predicted income is then:

$$\hat{y}_M = \hat{f}_M(c) = e^{\hat{\alpha} + \hat{\beta} \log y_p + \sigma^2/2} \tag{4}$$

Where, σ denotes the standard deviation of the residuals ε. Now, although the regression coefficient $\hat{\beta}$ – the intergenerational elasticity – is often used as a summary index of persistence or "inheritability" (the opposite of mobility), another commonly used measure (which has the advantage of being equally sensitive to both margins), is the correlation coefficient between $\log y$ and $\log y_p$. Since this coefficient is the square root of the R^2 of the Galtonian regression, it can be written as a specific case of $O = g\big(I(\hat{y}), I(y)\big)$, namely:

$$\hat{\rho} = \sqrt{\frac{I(\hat{y}_M)}{I(y)}} \quad \text{when} \quad I(x) = Var \log x \tag{5}$$

Noting that the rank of an observation x_i in a distribution $F(x)$ is simply the quantile $q_i = F(x_i)$, and that this cumulative distribution function is inversible, it is clear that there will also be a specific predictor $f_R(c, \varepsilon)$ for rank-rank regression or correlation coefficients.

Ex-ante inequality of opportunity

The literature on inequality of opportunity has usually considered a vector (*k >1*) of circumstance variables, rather than a scalar. When information on parental income or wealth is available, those variables can be elements in **c**. But they are complemented by others, such as ethnicity, sex,

[5] We say 'generally' because there are studies that include the incomes of more than one generation as circumstances (Olivetti, Paserman, and Salisbury, 2018). There are also studies that consider interactions with race. (e.g., Mazumder, 2014).

parental education or occupation, etc.[6] Frequently, however, this approach has been used for societies or periods for which reliable information on parental income is not readily available.

In that case too, scalar indices summarizing the extent of inheritability (here: inequality of opportunity) are often of the form $O = g\big(I(\hat{y}), I(y)\big)$. In the ex-ante parametric approach of Ferreira and Gignoux (2011) or Niehues and Peichl (2014), the logarithm of parental income in (3) is simply replaced by the vector of circumstances, and the prediction function is given by:

$$f_{EA}(c, \varepsilon) = e^{\alpha + C\gamma + \varepsilon} \tag{6}$$

This generates a vector of predicted incomes analogous to that in (4) – without the Blackburn (2007) correction – and the relative measure of inequality of opportunity is precisely:

$$IOR_{EA} = \frac{I(\hat{y}_{EA})}{I(y)} \tag{7}$$

A version of Equation (7) can also describe the ex-ante non-parametric estimator of inequality of opportunity (Checchi and Peragine, 2010; Ferreira and Gignoux, 2011) when the prediction function is changed from (6) to (8):

$$f(c, \varepsilon) = \int_0^1 y dF(y|\boldsymbol{c}) \tag{8}$$

Equation (8) simply yields the conditional means for all those who share the same vector of circumstances $\boldsymbol{c}$. So $I\big(\hat{y}_{EA(n)}\big)$ is simply computed over the smoothed distribution where individual incomes are replaced by the average incomes of all individuals who share the same vector of circumstances – that is, individuals in the same type. [7]

In fact, both the parametric and non-parametric prediction functions - (6) and (8) – are predicting type means, with the caveat that (6) imposes a linear functional form on the relationship between $\boldsymbol{c}$ and y. The reference situation of equality of opportunity is therefore:

$$\mu(y|\boldsymbol{c_l}) = \mu(y|\boldsymbol{c_m}), \forall \boldsymbol{c_l}, \boldsymbol{c_m} \in C \tag{9}$$

So, inequality of opportunity quantifies deviations from (9).

[6] See Bjorklund, Jäntti, and Roemer (2012) for an example of IOp using parental income as a circumstance.
[7] See Foster and Shneyerov (2000).

Ex-post inequality of opportunity

But equation (9) is clearly weaker than (1): it is implied by but does not imply (1). It is possible that two types have cumulative distribution functions (CDF) that are different but have the same mean. Since it is Equation (1) – full equality of the type-conditional distribution functions – that really implies and is implied by the orthogonality of income and circumstances, many authors have preferred empirical approaches that use estimates of the CDF, rather than just the mean, to either detect or measure inequality of opportunity. Lefranc, Pistolesi, and Trannoy (2009), for example, use stochastic dominance techniques to test for differences across type distribution functions, and thus the null hypothesis of equal opportunities.

For *measuring* IOp, Checchi and Peragine (2010) propose to aggregate income differences across the quantiles of the conditional distributions, while abstracting from level differences across types. Their prediction function is given by:

$$f_{EP}(c,\varepsilon) = \frac{\mu}{\mu_q} F^{-1}(q|\boldsymbol{c}) \tag{10}$$

Since $y_{qc} = F^{-1}(q|\boldsymbol{c})$,

$$I(\hat{y}_{EP}) = \int_{q=0}^{1} \frac{\mu}{\mu_q} I_q(y_{qc})\mathrm{dq} \tag{11}$$

And

$$IOR_{EP} = \frac{I(\hat{y}_{EP})}{I(y)} \tag{12}$$

Equation (12) is analogous to (7), but uses (11) to predict incomes, rather than (6) or (8). In words, Checchi and Peragine (2010) compute inequality in predicted incomes by computing some inequality measure across types for each decile; then multiplying that inequality by the ratio of the overall mean to the quantile mean (again, computed across types), and finally aggregating across quantiles. IOp is, once again, the ratio of inequality in predicted incomes to observed inequality.

The case for computing inequality of opportunity as horizontal gaps between cumulative distribution functions – as departures from the definition of equality of opportunity in (1) – can therefore be made with no reference to the notion of effort. There is no effort variable in equations

(10) – (12). Historically, this logic has appealed to theorists of equal opportunities because, under some assumptions, the relative degree of effort expended – or responsibility taken – by a person can be proxied by her relative position (quantile) in the income distribution of her type (see Roemer, 1998). Under those assumptions, Equation (1) does not simply denote the orthogonality of outcomes and predetermined circumstances. It also corresponds to a situation in which people who exert the same degree of effort achieve the same outcomes.[8] But the use of (12) as a meaningful measure of deviations from the ideal of incomes orthogonal to circumstances does not *require* adherence to the theory or its assumptions.

3 The central empirical challenge: model selection

Empirical applications of all three versions of the prediction problem described above 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 all suffer from a model selection problem, and this is the issue this section focuses on.

The intergenerational mobility literature makes most sense when interpreted as attempts to estimate, as accurately as possible, a descriptive measure of association between two variables. This may be a regression coefficient, a correlation coefficient, or some summary statistic from a transition matrix or copula. It is presumably understood that these parameter estimates do not represent – in any way, shape, or form – estimates of the causal effect of parental income on child income, since they are hopelessly biased by omitted variables with which parental incomes are bound to be correlated. So, they must clearly be interpreted as simple estimates of bivariate association.[9]

In the IOp literature, where the explicit intent is to quantify the extent to which today's inequality is inherited – that is, the extent to which inherited circumstances predict incomes today – authors

[8] These assumptions are: the degree of effort exerted is by definition orthogonal to circumstances; all circumstances are observable; the effect of luck cannot re-rank individuals in terms of income; and income is a monotonic function of effort (for a discussion see Roemer and Trannoy (2015)).

[9] Yet, as shown earlier, measures of association such as the correlation coefficient are very closely related to measures of the share of inequality predicted by the background variable – and are sometimes interpreted as such in the literature.

make use of additional background variables that might be available in the data. And as soon as one considers the use of additional background variables – which may be many and may consist of multiple categories – one faces the standard issue of model selection in the presence of two competing biases.

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 inevitably a strict subset of background circumstances. Omission of the unobserved circumstances tends to bias estimates of IOp downwards (Ferreira and Gignoux, 2011; Roemer and Trannoy, 2016).

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 same problem manifests as exploding sampling variation around cell means as cell sizes decline below a certain level. This problem introduces an upward biased in the estimation of explained variance (Chakravarty and Eichhorn 1994; Brunori, Peragine, and Serlenga, 2019).

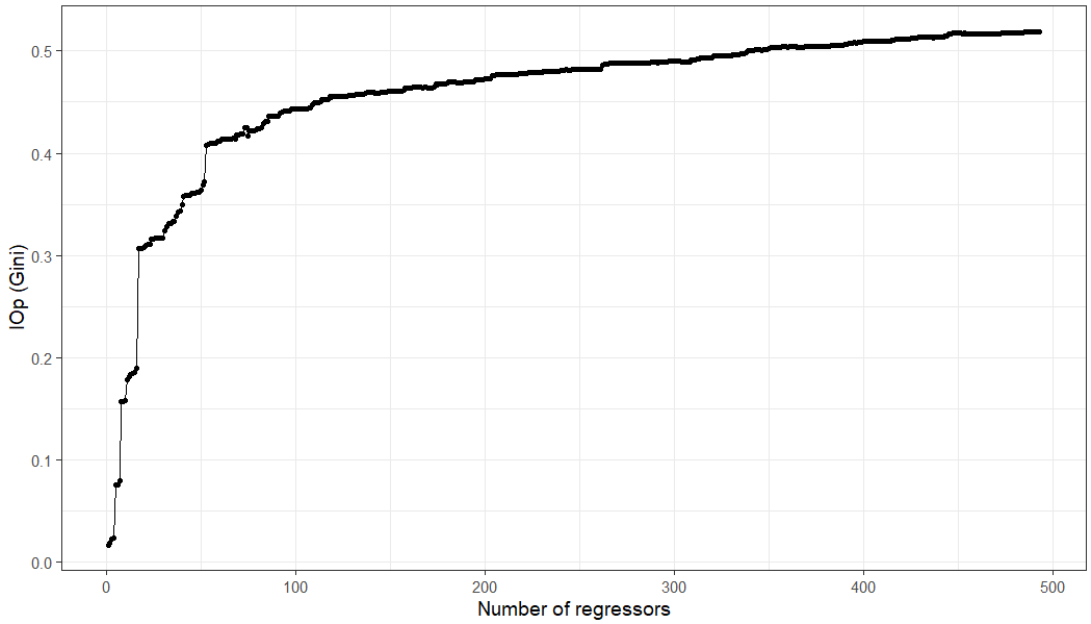
Although this problem was recognized from the outset, most of the early literature failed to address the trade-off between the two kinds of bias in a systematic way.[10] The early studies that proposed either parametric or non-parametric methods to estimate IOp relied on ad-hoc specifications, either of the regression model or of the type partition. Yet, changing the number of regressors in such a model can dramatically alter the final estimates of IOp

To illustrate the point, we show here the values for IO_{EA} that we obtain by specifying hundreds of regression models of increasing complexity. The illustration is based on the data that we will

[10] 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.

subsequently use for estimating inherited inequality in South Africa. (See Section 5.) Figure 1, which plots IOp estimates (using the Gini coefficient as $I(x)$) against the number of regressors included in a linear regression using our own data, illustrates this variation. It reports results from a standard ex-ante parametric approach, as models rise in complexity by adding regressors. In constructing this figure, we used all circumstance variables we use in the remainder of the paper. Furthermore, we restricted interactions to pairwise interactions, thereby dampening the potential growth in the IOp estimates. Moreover, given that all regressors are categorical and the inclusion of all interactions leads to a large number of sparsely populated categories, we consider "only" the 493 regressors that describe at least 10 observations in the sample (e.g. we exclude the interaction "Father education == 1 and Mother education==10" which concerns no observation in the sample). Still the number of possible interaction terms is huge (approximately 1.948 x 10^{296}). Therefore, for each possible number of regressors we select the most appropriate specification by backward stepwise selection (Lumley, 2022). Even with these restrictions, ex-ante IOp Gini estimates from our dataset with models of increasing complexity range from 0.016 to 0.52 (from 2.5% to 86% of total inequality).

Figure 1: Ex-ante parametric IOp by backward stepwise selection



Source: Author's calculation on NIDS 5

It should be clear from Figure 1 that, in the presence of these two biases working in opposite directions, obtaining a meaningful estimate of $I(\hat{y})/I(y)$ depends crucially on selecting the 'right' model for the prediction function $y = f(\boldsymbol{c}, \varepsilon)$. But 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. When the model is being used for prediction, however, 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, some machine learning methods have recently been applied to the measurement of inequality of opportunity, in attempts to let the data determine the right prediction model. Li Donni, Rodriguez, and Dias (2015) for example, suggest the use of finite mixture model to define types. But these models are extremely costly in terms of parameters and tend to produce rather parsimonious partitions, leading to very conservative IOp estimates.[11]

In a similar spirit, Brunori, Hufe, and Mahler (2023) use conditional inference trees and random forests (CITF), which were introduced by Hothorn, Hornik, and Zeileis (2006). 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 differences across means in the two resulting cells. Since the ex-ante approach to IOp involves computing inequality among type means, such an algorithm is the conceptually right approach to selecting the partition and estimating Equation (7), albeit with a different functional form $f \in \mathcal{F}$ than those in (6) or (8).

But precisely because conditional inference trees focus on differences between means, not full distribution functions, those who subscribe to the stricter criterion of equal CDFs for equality of

[11] These models have been extensively used in the health economics literature. The typical partition obtained is made of an unrealistically low number of types. Li Donni, Rodriguez, and Dias (2015) use a five-type partition to model IOp in health a sample of 17,000 individuals, representative of the cohort of individuals born in UK in the third week of March 1958. The partition used by Carrieri, Davillas, and Jones (2020) is even more parsimonious. Using a subsample of the Understanding Society: The UK Household Longitudinal Study made of 5,800 respondents they define a partition in three types. Brunori, Trannoy, and Guidi (2021) suggested the use of cross-validation to obtain a more realistic number of nodes, which nevertheless remains constrained by the large number of parameters necessary to estimate latent classes.

opportunity – including those who follow Roemer (1998) in interpreting the quantiles of those CDF's as relative measures of individual effort – will need an alternative data-driven approach. In what follows, we propose the use of one such approach, namely transformation trees. Transformation trees are supervised machine learning algorithms recently introduced by Hothorn and Zeileis (2021). In the next section we explain how the algorithm works and how it represents an exact empirical implementation of Roemer's approach to inequality of opportunity.

4 Estimating IOp using Transformation Trees

As noted in Section 2, an ex-post approach to inequality of opportunity essentially consists of measuring inequality for each quantile, across the types' conditional distributions functions, as in Eq (11) above, and then appropriately aggregating across quantiles. The key ingredient for the approach, therefore, is to estimate the income level at quantile *q* in type *c*, that is: the conditional quantile function $y_{qc} = F^{-1}(q|C = c)$. 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 defined, 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 rather arbitrarily, and 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 $F\left(y_{qc}|C = c\right)$, 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.

The 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 $F\big(y_{qc}|C=c\big)$. In the limit, they assume that there exist parameters $\theta \in \Theta$ such that:

$$F\big(y_{qc}|C=c\big) = F\left(\tilde{y}_{qc}, \theta(c)\right), \theta: \mathbb{C} \to \Theta \tag{13}$$

$\theta(c)$ is known as the conditional parameter function, which maps from the set of all possible type partitions on to the set of possible distributional parameters. Under this assumption, the problem of estimating the conditional distributions across 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:

$$\hat{\theta}^N(c) = \arg\max_{\theta \epsilon \Theta} \sum_{i=1}^{N} w_i(c)\ell_i(\theta) \tag{14}$$

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:

$$w_i(c) = \sum_{b=1}^{B} I(c \in \mathcal{B}_b \wedge c_i \in \mathcal{B}_b) \tag{15}$$

The indicator function takes the value 1 when observation i is sufficiently "close" to c, so the weights in (14) 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 (14-15) 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 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:

$$\hat{y}_T = \hat{f}_T(c) = \frac{\mu}{\mu_q}\tilde{y}_{qc} \quad \text{where} \quad \tilde{y}_{qc} = F^{-1}\left(q, \hat{\theta}^N(c)\right) \tag{16}$$

The Transformation Tree estimate of ex-post inequality of opportunity is simply:

$$IOR_T = \frac{I(\hat{y}_T)}{I(y)} \tag{17}$$

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

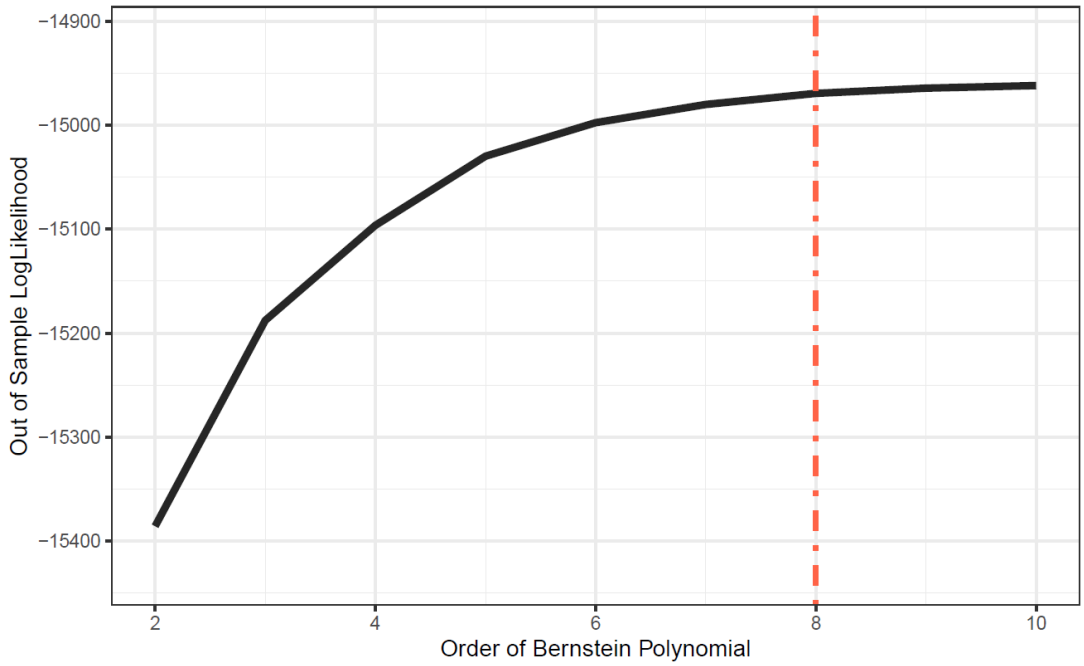
1. set a confidence level (α);
2. set a polynomial order (M);
3. estimate the unconditional distribution with the Bernstein polynomial of order M;
4. test the null hypothesis of polynomial parameters stability for all possible partitions based on each x and store $p - values$.
5. if $\forall x$ and each possible partition the Bonferroni-adjusted $p - value > \alpha$, stop 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.

In our application below, we follow statistical convention and set α to 0.01. Then, we set M, the order of the Bernstein Polynomial. The selection of M is not as simple as that of α, because how well a certain order interpolates the distribution is intrinsically data-dependent. An order too small might result in a poor approximation of the distribution, while a too elevated order would translate into a loss of degrees of freedom and high computational costs.

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 (ranged between 2 and 10). We select the lowest order in which the relative improvement of the log-likelihood that would be obtained by estimating an additional parameter is smaller than 0.1%. In our application, this procedure – summarized in Figure 2 below – yields a Bernstein polynomial of order 8.

In step 3, an unconditional CDF for our sample is 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. To intuitively illustrate this key test, Appendix 2 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).

Figure 2: Out of Sample Log-Likelihood by orders of Bernstein Polynomial



Source: Authors' elaboration from NIDS 5.

After following steps 4-7 we obtain an estimated Transformation Tree for South Africa and, from that tree, a number of outputs that are described in Section 6. But before presenting those results, we briefly describe our dataset in Section 5.

5. Data

We apply this method to the latest wave of the National Income Dynamics Study (NIDS 5) survey, carried out by the Southern Africa Labour and Development Research Unit (SALDRU) for the year 2017 (Brophy et al., 2018). NIDS is a longitudinal survey, with previous waves collected in 2008, 2010/11, 2012, and 2014/5. 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.[12] Moreover, IOp has already been analyzed in South Africa (see Piraino, 2015, and Brunori, Ferreira, and Peragine, 2021), so our results can be readily benchmarked against alternative methods.

Before any filters, the NIDS 2017 contains 20,461 individuals. The reason we use only the 2017 wave of the survey is that, in that year, SALDRU oversamples rich households, allowing for more precise inequality estimates due to the inclusion of more detailed information from the top of the income distribution (Branson, 2019). This was done in earlier waves. Our main results are obtained from this complete sample. However, SALDRU also provides appropriate weights to exclude wealthy households oversampled in 2017 and we report statistics on both samples in this section for comparability. We refer to the sample without oversampling of the rich as 2017b.

As our outcome variable we use monthly age-adjusted equivalized disposable household income, in 2015 rands. It includes all regular incomes received by households, including imputed rental income from owner-occupied housing, net of taxes. To account for scale economies in consumption, the square-root equivalence scale is used (Buhmann et al., 1988; OECD, 2013). The age adjustment – applied to account, at least in part, for life-cycle dynamics – consists of regressing our income variable (as defined so far) on age and age squared, and using the sum of constant and residual as the adjusted variable (see, e.g., Palomino et al., 2022).

The circumstances available in the NIDS 2017 dataset are: sex (male and female), ethnicity (African, Asian/Indian, coloured, and white), fathers' and mothers' education (13 levels, ranging from "Non-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)[13]. Item non-response is a serious issue in these data, particularly for information on respondent's parents. We are able to alleviate the problem

[12] See Mahler and Baur (2023) for recent estimates.

[13] Note that the question refers to current occupation of the parents or the last occupation. We exploit the panel structure of NIDS and look at information about circumstances reported by the same individuals in previous waves and a) For those with missing circumstances, we will with the oldest value available, and b) we use the first value of the parental occupation reported in the data.

somewhat by matching individuals across waves of the NIDS, which is a longitudinal dataset with previous waves collected in 2008, 2010/11, 2012, and 2014/5. Table 1 reports the shares of observations with missing information by circumstance variable, before and after this cross-wave matching. Although substantial progress is made in the parental occupation variables, this is less the case with mother's education.

Table 1. Missing circumstances before and after cross-wave matching

Matching	Ethnicity	F.Occ	M.Occ	F.Edu	M.Edu	Sex
Before	5.07	37.67	35.95	44.21	44.72	5.03
After	0.01	5.03	5.04	30.83	43.42	0.00

Source: Own elaboration from NIDS 5. F.Occ stands for Father Occupation, M.Occ stands for Mother Occupation, F.Edu stands for Father Education, M.Edu stands for Mother Education

We then apply two filters to the sample: we restrict the analysis sample to adults aged between 18 and 80; and omit all observations with any missing information in either income or circumstances. This leaves us with 7,297 observations for the analysis. There is clearly a risk of sample selection if observations are not missing at random. We cannot completely address that problem, which plagues most studies of intergenerational mobility or inequality of opportunity in developing countries. However, we do at least use cross sectional weights calibrated to province, sex, race, and age group totals. As proposed in (Brunori, Salas-Rojo, and Verme 2022), we correct these weights for item non-response by applying the reweighting method proposed in (Korinek, Mistiaen, and Ravallion 2006) and implemented by Munoz and Morelli (2021). The reader is referred to those papers for details.

Table 2 shows some basic descriptive income statistics for both the 2017 and 2017b analysis samples, including Gini coefficients which, in both cases, are just over 0.6 – despite the equivalence scale and age adjustments.[14]

[14] These two adjustments are likely to reduce inequality, relative to per capita income unadjusted for age.

Table 2: Descriptive Income Statistics

Sample	N	Mean	Sd	Gini	MLD
2017	7297	6474.20	11173.20	0.605	0.678
2017b	6730	6418.23	11470.35	0.599	0.664

Source: Own elaboration from NIDS 5. N stands for the analysis sample size. Sd stands for Standard Deviation, MLD stands for Mean Logarithmic Deviation. Incomes in rands (2015).

Table 3 contains summary descriptive statistics for the circumstance variables, as proportions of the weighted analysis sample.

Table 3: Descriptive Circumstance Statistics

Ethnicity	Figure labels	2017		2017b	
African	1	78.27		83.58	
Asian/Indian	2	1.9		1.1	
Coloured	3	11.55		12.12	
White	4	8.28		3.19	
Sex	Figure labels	2017		2017b	
Female	1	37.14		36.12	
Male	0	62.86		63.88	
Education	Figure labels	Mother (2017)	Mother (2017b)	Father (2017)	Father (2017b)
Non-Educated	0	54.73	58.59	57.38	61.71
Grade 1	1	0.59	0.64	0.67	0.73
Grade 2	2	1.47	1.56	1.52	1.63
Grade 2	3	2.38	2.54	2.32	2.39
Grade 4	4	3.38	3.54	2.54	2.66
Grade 5	5	2.71	2.82	2.37	2.44
Grade 6	6	3.03	3.12	2.63	2.64
Grade 7	7	4.32	4.37	3.29	3.19
Grade 8	8	7.83	7.37	7.35	6.91
Grade 9	9	1.9	1.87	1.82	1.72
Grade 10	10	4.77	3.92	4.4	3.49
Grade 11	11	2.03	2.05	1.6	1.53

Grade 12 and more	12	10.85	7.61	12.11	8.95
Occupation	Figure labels	Mother (2017)	Mother (2017b)	Father (2017)	Father (2017b)
Army	0	0.01	0.01	0.58	0.55
Managers	1	0.58	0.42	2.64	1.78
Professionals	2	5.8	4.1	4.15	3.05
Technicians	3	1.4	0.91	2.1	1.6
Clerks	4	1.69	0.85	0.9	0.7
Service	5	3.1	2.75	6.43	6.37
Skilled	6	0.22	0.24	0.88	0.85
Craft	7	1.34	1.26	10.65	9.81
Operators	8	0.26	0.21	12.16	12.38
Elementary	9	24.06	25.13	20.58	21.25
Other/not in the labour force	10	61.55	64.13	38.93	41.66

Source: Own elaboration from NIDS 5. All values are shares (%) of the analysis sample,

6. Results: Inequality of Opportunity in South Africa.

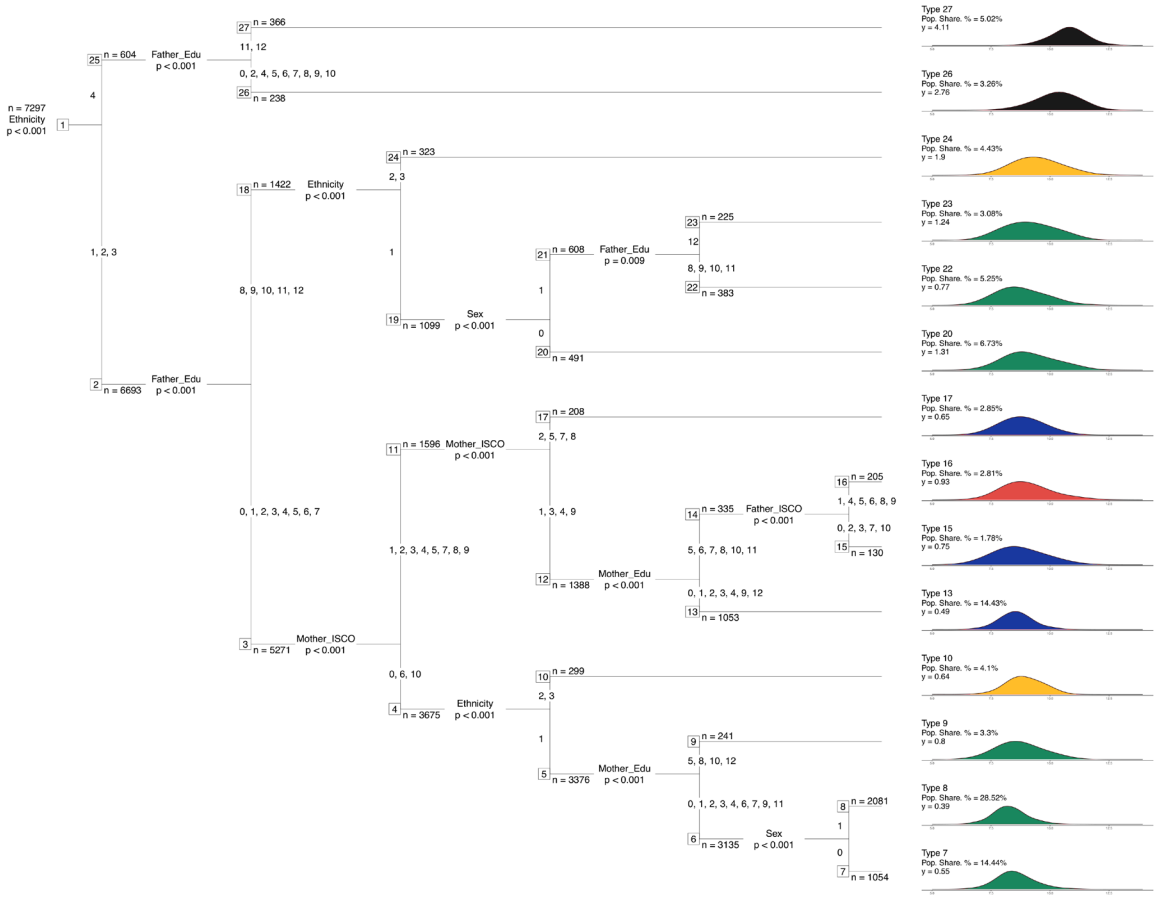
Applying the algorithm outlined in Section 4 to estimate Equation (16), yields the transformation tree shown in Figure 3. The splitting process generated by the algorithm should be read from left to right. The first split divides the population between the White population (ethnicity = 4; above) and all others (ethnicity = 1, 2, 3). As we move to the right, other circumstances subsequently partition the population following the algorithm, until the final nodes – types – are reached. There are fourteen types in this optimal partition, and the Figure shows the parametrically estimated density function for each of them, as well as indicating the population share accounted for by each type, and its mean income as a multiple or share of the overall mean.[15]

In terms of the model selection challenge illustrated in Figure 1, the algorithm partitioned the population into these fourteen groups (and fit CDF's to them) so as to maximize the likelihood of fitting the data, under the restrictions $f \in \mathcal{F}$, with $\mathcal{F}$ being the class of recursive binary TrT

[15] For clarity, given the high right-skewness of South African income distribution, and although we use income in levels to compute all our measures, we plot the density of log income.

estimators. The partition corresponds to the products (or interactions) of various dummy variables defined over the circumstances. Type 27, for example, which is the richest type at the top of the Figure, corresponds to the product of dummy variables $x_1 = \mathbf{1}_{race=white} \times \mathbf{1}_{father\ education=11\ or\ 12}$. Type 8, which is the poorest type and second from the bottom of the Figure, corresponds to $x_{13} = \mathbf{1}_{race=black} \times \mathbf{1}_{father\ education\in\{0-7\}} \times \mathbf{1}_{mother\ occup.\ \in\{0,6,10\}} \times \mathbf{1}_{sex=female} \times \mathbf{1}_{mother\ education\ \in\{0-4,6,7,9,11\}}$. And so on.

Figure 3: Transformation Tree for South Africa, NIDS 2017.



Source: Authors' elaboration from NIDS 5.

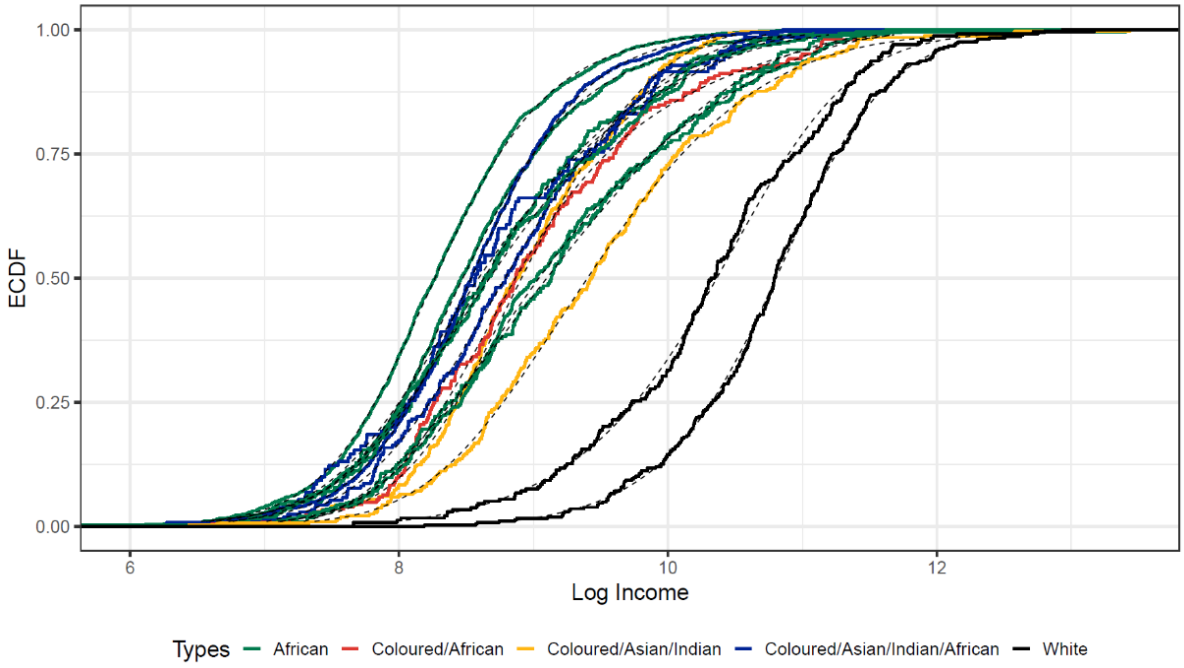
The ability to identify these specific patterns in data does have some cost in terms of model variance. This type of tree is not immune to the problem of sensitivity to the estimated model, which is common to regression and classification trees, and therefore, we caution against

overinterpreting the obtained partition and complementing the analysis with resampling-based tools that we introduce in the following pages. Moreover, when interpreting trees, it is important to keep in mind that in some cases a split can be misleading. When the algorithm uses a certain circumstance to divide the sample, it must place all individuals from the node that originates the split either in one subgroup or the other. If there are very few respondents who have a specific value for the characteristic in question, the assignment to the group can be almost random. To address this issue, it is possible to complement the analysis of the tree structure with tabulations that show the share of observations in each type and category by circumstance like the ones presented in Appendix 4. Take, for example, the composition of type 8, the first row of table A.6 regarding mother's occupation. Type 8 includes both respondents with non-working mothers and mothers in skilled manual occupations. However, the relative composition is extremely different, with the first group consisting of over two thousand respondents, while only six respondents report a mother in a skilled manual job.

Figure 4 shows the estimated cumulative distribution functions (ECDF) for all fourteen types. The different colors denote types characterized by a certain ethnic group or mix of groups. The polarization of South African society by race is clearly visible, with the two richest types, 26 and 27, (a) being exclusively white and (b) comprising all of white people in the sample. There are no white people in the other twelve types in our sample. Together, they represent 8.3% of the sample. At the same time, although the whites are isolated at one end of the distribution of opportunities in this country, they are not homogeneous. The tree has split those with the most highly educated fathers (completed secondary or tertiary) from the rest. The difference between their average incomes is 135% of overall the sample mean. At the other extreme, the poorest type consists exclusively of black females with generally less educated parents and mothers in certain low-skilled occupations[16]. This is a large group, accounting for over 28% of the population and earning less than 40% of the overall mean. In between, the socially intermediate position of South Africans of Indian and Asian origin (alongside some of the so-called ‘coloured’) is evident in Types 10 and 24, pictured in yellow.

[16] To allow for maximum flexibility in the estimation, both parental occupation and parental education are treated as categorical, rather than ordinal, variables. Nevertheless, with few exceptions, the sample is split consistently with the order of the variables.

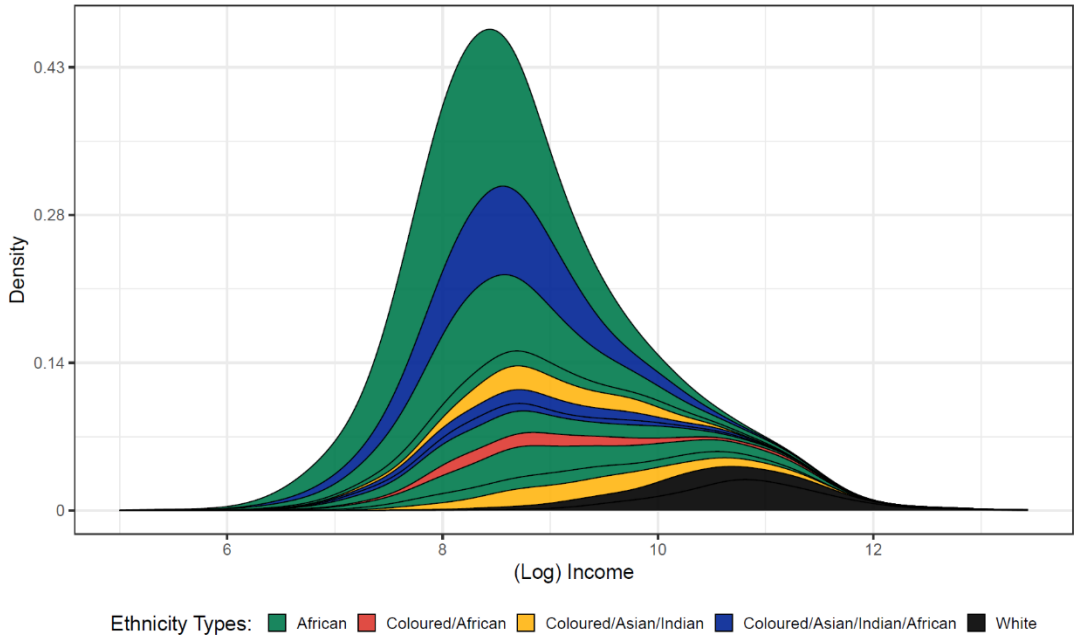
Figure 4: Conditional Income Distributions by Type.



Source: Authors' elaboration from NIDS 5.

The same information can be represented in yet another potentially useful way, by showing the country's overall population density function (of log incomes) as a mixture of the distributions of the fourteen types, as shown in Figure 5 below. Vertical slices of this kernel density estimate would then yield the racial composition of each income range corresponding to the logarithmic scale on the horizontal axis.

Figure 5: Density function as a mixture of type distributions



Source: Authors' elaboration from NIDS 5.

The dashed line accompanying each ECDF in Figure 4 is the outcome predicted by the Bernstein polynomial: $F\left(\tilde{y}_{qc}, \hat{\theta}^{N}(c)\right)$ for each type. The estimated incomes at each quantile, $\tilde{y}_{qc}$, are used to compute IOp through Equations (16-17). We use two different inequality measures $I(x)$ for that computation, namely the Gini coefficient and the mean log deviation (MLD). The mean log deviation was used extensively in the early IOp literature, given its ideal decomposability properties (see Foster and Shneyerov, 2000 and Ferreira and Gignoux, 2011). As it became increasingly clear that standard decomposability was not, in fact, required for the measurement of IOp – unless one wishes to interpret within-group inequality as being entirely driven by effort – the Gini has been used more frequently. It has the advantage, as noted by Brunori, Palmisano, and Peragine (2019), that it 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 and, although we report both measures in Table 4 below, we focus the discussion on the Gini estimates. The upper part contains results for the main sample for 2017 (which oversamples the rich) and the bottom part reports results for the alternative sample (2017b), as discussed in Section 3.

Table 4: Inequality of Opportunity Results

Sample 2017	Gini	Abs. Gini IOp	Rel. Gini IOp	MLD	Abs. MLD. IOp	Rel. MLD. IOp	Types
TRT	0.605	0.445	73.58	0.678	0.330	48.75	14
CIT	0.605	0.408	67.44	0.678	0.274	40.41	12
CIRF	0.605	0.430	71.07	0.678	0.299	44.10	
Sample 2017b	Gini	Abs. Gini IOp	Rel. Gini IOp	MLD	Abs. MLD. IOp	Rel. MLD. 4.IOp	Types
TRT	0.599	0.418	69.77	0.664	0.288	43.44	14
CIT	0.599	0.401	66.96	0.664	0.264	39.82	9
CIRF	0.599	0.379	63.24	0.664	0.229	34.42	

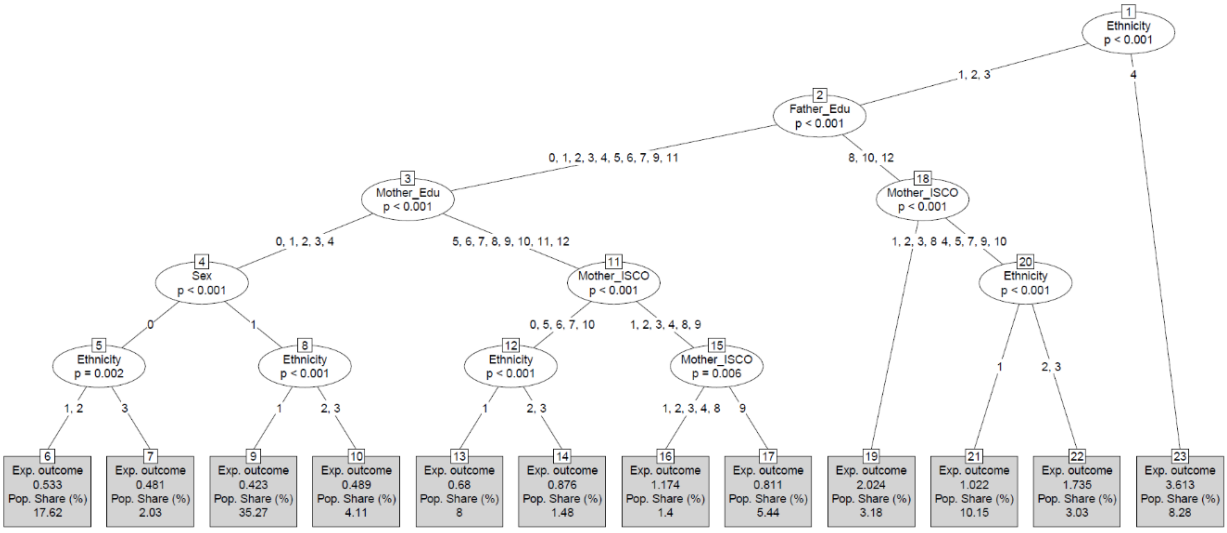
Source: Own elaboration from NIDS 5. TrT stands for Transformation Tree, CIT for Conditional Inference Tree, CIRF for Conditional Inference Random Forest, Abs. for Absolute, Rel for Relative.

The headline results are in the first row of the table. The Gini coefficient calculated on the vector $\widehat{\boldsymbol{y}}_{\boldsymbol{T}}$ (see Eq.16) obtained from the ECDFs of the fourteen types in our ex-post partition is 0.44, or 73% of the overall Gini coefficient of 0.61 for South Africa. This is a remarkable number: the "opportunity Gini" for South Africa is higher than the overall income Gini coefficient of the United States (0.41 as reported by the World Bank for the same year).[17] Not only that, but inherited inequalities account for almost three-quarters of the (extremely high) inequality in current incomes in the country. While this is perhaps not entirely surprising, given the history of Apartheid, it is certainly the case that previous methods had not found similarly high opportunity ratios. Piraino (2015), for example, employs the ex-ante approach and two possible econometric methods to estimate inequality of opportunity on gross employment earnings (using up to 54 Roemerian types). Depending on the set of circumstances considered he finds a level of IOp ranging between 17% and 24% of total inequality measured with mean logarithmic deviation (MLD) – which compares to our MLD estimate of 48%. This difference is in part due to the oversampling of richer households, but it persists when using comparable samples where the relative IOp in MLD is 36%.

[17] See https://data.worldbank.org/indicator/SI.POV.GINI?locations=US

The second and third rows in each part of Table 4 contain benchmark estimates from applying ex-ante approaches to our data. Figure 6 shows the ex-ante tree obtained with same data, replicating (exactly) the approach of Brunori, Hufe, and Mahler (2023) to construct conditional inference trees and random forests. As noted in Section 3, these two (closely related) machine-learning estimators have recently been applied to the computation of IOp (in Europe) and we include their estimates here for comparison and benchmarking only.[18] Note that although the structure of the ex-ante tree is similar, with a preponderant role of race in the definition of the tree structure, some differences emerge (e.g., white respondents are no longer split in two types).

Figure 6: Conditional Inference Tree for South Africa, 2017



Source: Own elaboration from NIDS 5.

In terms of the IOp summary statistics in Table 4, the ex-ante (CITF) tree estimates are a little lower than for the ex-post (TrT): an opportunity Gini of 0.41 in the ex-ante case, versus 0.44 in the ex-post case. It is tempting to conclude that this might be because, by looking only at type means, the ex-ante approach misses additional differences along the ECDFs. But one should be cautious with this

[18] They are not our focus in this paper and readers are referred to Brunori, Hufe, and Mahler (2023) for definitions and methodological descriptions.

interpretation. The random forest ex-ante estimate in the third row (0.43), which is known to be more robust than that of a single tree, is very close to the ex-post tree result. We interpret the broad similarity in the estimates across the three different methods – particularly the the TrT and the random forest – for both the Gini coefficient and the MLD – as an indication of the robustness of the data-driven approach to the assessment of inherited inequality.

We also interpret the fact that these methods tend to find higher shares of inherited inequality in overall dispersion than earlier approaches as a reflection of the ability of the algorithms to identify the most salient inequalities across subgroups. With fourteen "variables" or sets of interactions between dummy variables, our transformation tree finds an inequality in predicted incomes roughly similar to that of 200 regressors in the backward stepwise selection procedure depicted in Figure 1. Furthermore, adding another 300 regressors in that exercise yielded another six Gini points, likely by overfitting the data. This reflects the ability of the trees and forests to identify the "right" subgroups to focus on, by the very design of the algorithms.

Yet, although the ex-ante and ex-post methods presented here yield similar headline measures of IOp, they do identify different partitions – as one would expect from the fact that CITFs (and forests) are designed to find the most statistically significant differences between averages, and the TrT are looking for more general differences across CDFs, including in higher moments. Since both partitions (into 14 ex-post types and 12 ex-ante types) are of the same sample, we can map which ex-ante and ex-post type each individual in the sample belongs to. The mapping is shown in the Sankey plot in Appendix 3. Although space limitations preclude a detailed analysis of the plot, we note that movements between ex-ante and ex-post types are commonest when the ECDFs in Figure 4 are not far apart and cross one another. Examples include ex-post types 7 and 13, as well as 10 and 16. Indeed it can be seen that most members of ex-post type 13 are merged with either type 8 or 10 in the ex-ante case. These different allocations are the result of allowing differences in higher moments of the type distributions to affect splitting decisions in the tree.

While it may not always be possible to provide an intuitive explanation for the differences between the two partitions, there are cases in which it is possible to understand which characteristics distinguish respondents who are in two different types in the two partitions. Ex-ante type 17, for example, consists of non-white respondents whose mothers were employed in elementary occupations. Within this group, there are several subtypes in the ex post partition, but the majority of observations, over 80%, are concentrated in two ex-post types: 50% of the observations belong

to ex-post type 15, constituting 95% of this type, while 31% are categorized as ex-post type 16, making up 97% of this type. The distinguishing circumstance between the two groups is the occupation of their fathers. Individuals in ex-post type 15 tend to have fathers in craft occupations or unspecified occupational statuses which include also fathers outside the labor force, while the majority of individuals in type 16 have fathers whose occupation falls under the category of operators and elementary workers. The latter group, represented by respondents in ex-post type 16, have higher average income, although the difference is not significant enough to allow a split in the ex-ante tree. But also display a different cumulative distribution function with a substantially higher income variance.

Ex-ante and ax-post trees are therefore complementary tools to understand inequality of opportunity. It should be noted, however, that both CIT and TrT – as well as trees in general – are well known for their ability to detect complex interaction effects (low bias), but also to be highly sensitive to the exact sample observed (high variance). For this reason, the structure of a single tree should never be interpreted beyond its statistical meaning: the most likely partition in types in the observed sample, among the many probable partitions that could be obtained from other samples equally representative of the population of interest.

In the remainder of this section, we briefly present two additional sets of results that can also improve the robustness of our understanding of the phenomenon and can be easily obtained from this approach to inherited inequality: (i) a descriptive decomposition of the role of each individual circumstance variable, and (ii) an estimate of the objective function for a Rawlsian opportunity-egalitarian.

The role of individual circumstances

The prediction function in equation (16) is highly non-linear in circumstances, so that 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 similar cases in inequality analysis, the decomposition method most suitable to our application is the Shapley-Shorrocks decomposition (Shapley, 1953; Shorrocks, 2013). This decomposition computes the total contribution of a particular circumstance variable c_k to predicted inequality as the reduction in the latter when c_k is omitted from the prediction, averaged across all possible combinations of circumstances that omit

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;[19]
B) Estimate IOp in this subsample, as described in Section 4, but setting $\alpha = 1$;
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 its difference with respect to IOp;
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;
H) Repeat the algorithm for each c_k, $k \in \{1, \dots, K\}$.

Analogously to the common approach used in estimating random forests, we construct trees on a subsample 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 certain iterations, making the assessment of the relative contribution of each circumstance robust to the typical problem of variance of estimates based on a single tree.

Table 5 presents the results of the Shapley-Shorrocks decomposition across the six circumstance variables available in our data set. Results are presented as percentage shares of the ex-post opportunity Gini coefficients reported in Table 5, for both the main 2017 sample and the secondary sample 2017b.

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

Table 5: Ex-post tree Shapley value Decomposition (as % of Gini IOp)

Year	Ethnicity	F.Occ	M.Occ	F.Edu	M.Edu	Sex
2017	30.59	14.16	16.07	17.63	17.23	4.33
2017p	44.64	10.18	12.29	14.57	13.01	5.3

Source: Own elaboration from NIDS 5. F.Occ stands for Father Occupation, M.Occ stands for Mother Occupation, F.Edu stands for Father Education, M.Edu stands for Mother Education.

The importance of the race or ethnicity variable, which was already evident from the tree in Figure 3, is confirmed here: it contributes 31% of IOp in the sample that oversamples the rich, and as much as 45% of the other sample. The difference reflects the fact that much of the "added" inequality among the rich is inequality among whites. Fathers' and mothers' educational levels come next in importance, with about 17% each, followed closely by their occupational categories, where the mother's occupation appears to contribute just a little more than the father's. Naturally, it should go without saying that, in keeping with the measurement-using-prediction spirit of our analysis, these decompositions are purely descriptive.

The lower envelope of quantile functions

Although the analysis of inherited inequality, in any of the forms described in Section 2, is inherently descriptive, it often raises normative questions about what the policy objectives should be with regard to intergenerational persistence, or inequality of opportunity. As with inequality in general, one must contend, in particular, with the *leveling-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 value. Policies might be arranged in such a way that there was no inherited inequality, but everyone lived in poverty.

The standard normative response to this philosophical objection is Rawls's argument that inequalities should be tolerated only insofar as they are to the benefit of the worst-off (Rawls, 1971). This gives rise to Rawlsian maximin objective functions, familiar to economists. And indeed, various versions of maximin objectives have been proposed in the context of inequality of

opportunity.[20] One version is to arrange society and choose policies so as to maximize the (average of the) lowest incomes at each quantile of the conditional distribution functions, across 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:

$$L(q) = \min_{\tau_m} F^{-1}(q, \boldsymbol{c_m}) \tag{18}$$

And choose policies so as to:

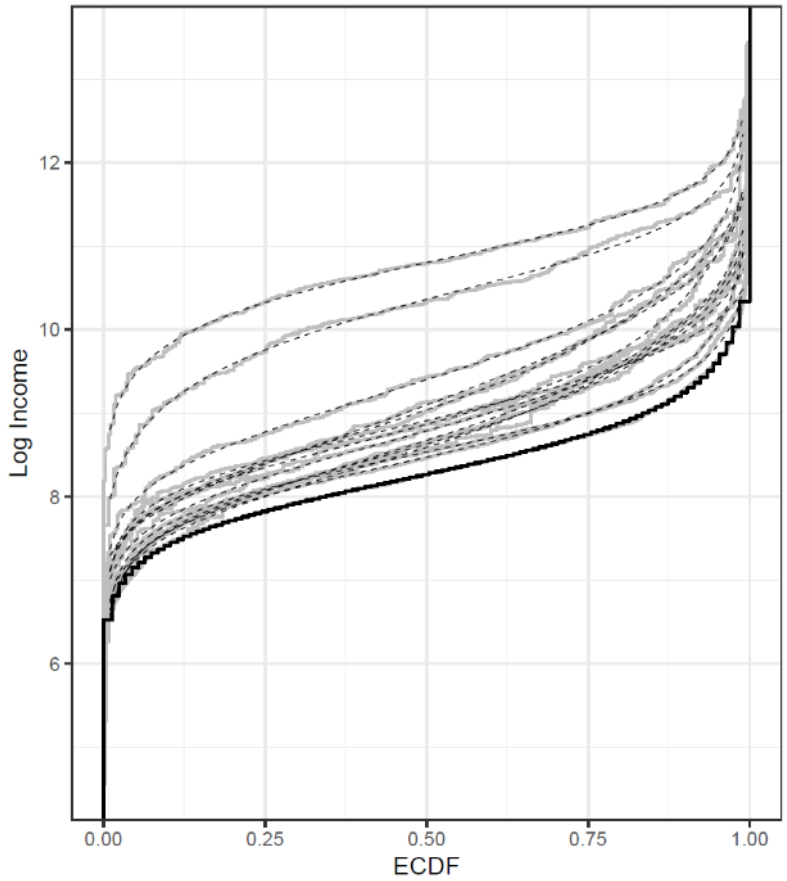
$$Max \int_0^1 L(q)dq \tag{19}$$

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).

Equation (18) defines the lower envelope of the set of quantile functions (inverse functions of the distribution function). Graphically, if one inverts the conditional CDFs in Figure 4, one obtains the type quantile functions, as in Figure 7 below. $L(q)$ defines the lowest points in the graph at each quantile. If the poorest type were first-order stochastically dominated by all other types, then this would simply be its quantile function, and Equation (19) 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 (19) 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 $\int_0^1 L(q)dq$ would be its average income. Therefore, the value of the maximand in (19) is informative per se, as a measure of shared income in a society, and is interestingly read in relative terms, as a measure of how close the shared income is to the average income, $\frac{\int_0^1 L(q)dq}{\int_0^1 F(q)dq}$.

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

Figure 7: Type quantile functions and the lower envelope



Source: Authors' elaboration from NIDS 5

In practice, a literal computation of $\int_0^1 L(q)dq$ might be over-sensitive to small (and possibly unstable) types detected in a particular sample. We therefore propose a robust version of the lower envelope which consists, in each quantile, of the average of the worst-off types adding up to at least 10% of the population. In the present application, however, the robust version is almost identical to the strict definition in (18), because South Africa's worst-off type (Type 8 in Figure 3) is dominated to a large extent by all other types and is also very large in terms of population. The area below

South Africa's lower envelope in 2017 is 2,203 Rand, or 34% of the overall mean of 6,474.20 rand, as shown in Table 2.[21]

7. Conclusions

The extent to which inequality 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 biological sex, race, parental income, or other indicators of family background. We have shown that an array of commonly used measures of intergenerational mobility and inequality of opportunity can be written down as functions of the ratio of inequality in these predicted incomes to inequality in current-generation incomes. What varies between them is the number and nature of the variables used for prediction, and of the prediction function itself. But they can all be expressed as a two-step procedure, in which incomes are first predicted by parental incomes or other inherited characteristics, and then inequality in those predictions is compared to observed inequality.

Such prediction problems inherently involve a statistical trade-off between a downward bias arising from omitting certain variables and interaction terms, and an upward bias from including too many such variables and overfitting the model. Data-driven, machine learning techniques, which are designed to perform well out of sample and avoid overfitting by regularization were developed to solve this class of prediction problems. In particular, we have proposed the use of transformation trees (Hothorn and Zeileis, 2021) to estimate ex-post inequality of opportunity, which involves computing horizontal distances across the conditional distribution functions of suitably defined population subgroups (types) and aggregating them across quantiles.

Transformation trees are particularly well-suited to the ex-post IOp approach because they predict incomes by simultaneously partitioning the sample and fitting flexible parametric esimates of these conditional distribution functions, so as to solve a well-defined local adaptive maximum likelihood

[21] The strict (non-robust) lower envelope is 2,168 rand in the 2017 sample. This declines to 1,941 rand in the 2017b sample (2,125 rand in the robust version).

problem. They should be of interest to those whose normative view of equal opportunities follow Roemer (1993, 1998), in which conditional quantiles are associated with relative degrees of responsibility or effort. But we argue that the method is of more general appeal: if one thinks of equal opportunity – or the absence of inherited inequality – as a situation in which predetermined and parental characteristics are orthogonal to – have no predictive power over – present-generation outcomes, then Equation (1) is the critical condition for it to hold. Equality of group means, which is tested by other algorithms such as linear regressions, tranditional non-parametric inequality decompositions, or conditional inference trees, is necessary but not sufficient. Transformation trees compute detect and quantify differences along the full conditional distribution functions.

We applied this method to South Africa, arguably the world's most unequal country, and found an opportunity Gini coefficient – our preferred measure of inequality in predicted incomes – of 0.44, corresponding to almost three-quarters of overall South African income inequality. When using an alternative measure like the mean log deviation, our estimate of inequality of opportunity the predicted share of inequality was at least twice as high in our estimate than in the previous literature.

Another advantage of this approach is that it generates a number of byproducts which are descriptively informative of the structure of inequality in South Africa. These include the transfomation tree itself, graphical depictions of the conditonal distributions, a Shapley decomposition of the relative contributions of individual circumstance, and an estimate of lower envelope of the set of quantile functions, an average of which is a meaningful measure of opportunity deprivation and an estimate of the policy maximand proposed by Roemer (1998).

That said, all estimation methods have advantages and disadvantages, and data-driven learning algorithms are no exception. Among the limitations of regression trees is the relatively high variance in the identified structure. As a result, researchers should not report only trees and forests, but also incorporate relative importance decomposition through bagging, and potentially integrating other standard econometric models, as supplementary tools. Employing these approaches collectively is most likely to lead to a thorough and robust understanding of inherited inequalities.

References

Bjorklund, Anders, Markus Jantti, and John E. Roemer. 2012. “Equality of opportunity and the distribution of long-run income in Sweden.” Social Choice and Welfare, 39, 675-696.

Blackburn, McKinley. 2007. “Estimating wage differentials without logarithms.” Labour Economics, 14, 73-98.

Branson, Nicola. 2019. “Adding a Top-up Sample to the National Income Dynamics Study in South Africa.” NIDS Technical Paper number 8.

Bourguignon, François, Francisco HG Ferreira, and Marta Menéndez. 2007. “Inequality of Opportunity in Brazil.” Review of Income and Wealth, 53, 585-618.

Brophy, Timothy, Nicola Branson, Reza C. Daniels, Murray Leibbrandt, Cecil Mlatsheni, and Ingrid Woolard. 2018. “National Income Dynamics Study Panel User Manual.” NIDS Technical Note Release.

Brunori, Paolo, Francisco HG 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.

Brunori, Paolo, Pedro Salas-Rojo, and Paolo Verme. 2022. “Estimating Inequality with Missing Incomes.” International Inequalities Institute Working Papers number 82.

Brunori, Paolo, Alain Trannoy, and Caterina Francesca Guidi. 2021. “Ranking populations in terms of inequality of health opportunity: a flexible latent type approach.” Health Economics, 30, 358-383.

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.

Carrieri, Vincenzo, Apostolos Davillas, and Andrew M. Jones. 2020. “A Latent Class Approach to Inequity in Health Using Biomarker Data.” Health Economics, 29, 808-826.

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.” The American Economic Review, 104, 141-147.

Checchi, Daniele, and Vito Peragine. 2010. “Inequality of Opportunity in Italy.” The Journal of Economic Inequality, 8, 429-450.

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

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.

Haveman, Robert, and Barbara Wolfe. 1995. “The determinants of children’s attainments: a review of methods and findings.” Journal of Economic Literature, 33, 1829-1878.

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.

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.

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion. 2006. “Survey Nonresponse and the Distribution of Income.” The Journal of Economic Inequality, 4, 33-55.

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.

LiDonni, 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.

Lumley, Thomas. 2022. R package “leaps”. Available at https://CRAN.R-project.org/package=leaps. Version 3.1, accessed on the August 31st, 2023.

Mazumder, Bhashkar. 2014. “Black–white differences in intergenerational economic mobility in the United States.” Economic Perspectives, 38, 1-18.

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

Munoz, Ercio, and Salvatore Morelli. 2021. “Kmr: A Command to Correct Survey Weights for Unit Nonresponse Using Groups’ Response Rates.” The Stata Journal, 21, 206-219.

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.

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.” In 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.” The Journal of Economic Inequality, 11, 99-126.

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

Van De Gaer, Dirk. 1995. “Equality of Opportunity and Investment in Human Capital.” Ph.D. Dissertation, Katholieke Universiteit Leuven.

Appendix 1: 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 (5) 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:

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

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: $\mathfrak{I} = \cup_{b=1,\dots B}\mathcal{B}_b$. 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 $\hat{y}_{qc} = F^{-1}\left(q,\ \hat{\theta}(c)\right)$, from which a measure of ex-post inequality of opportunity can be computed as $\widehat{IOp} = \int_{q=0}^{1} w_q I_q(\hat{y}_{qc})$.

Appendix 2: An illustration of the M-fluctuation test using made-up data

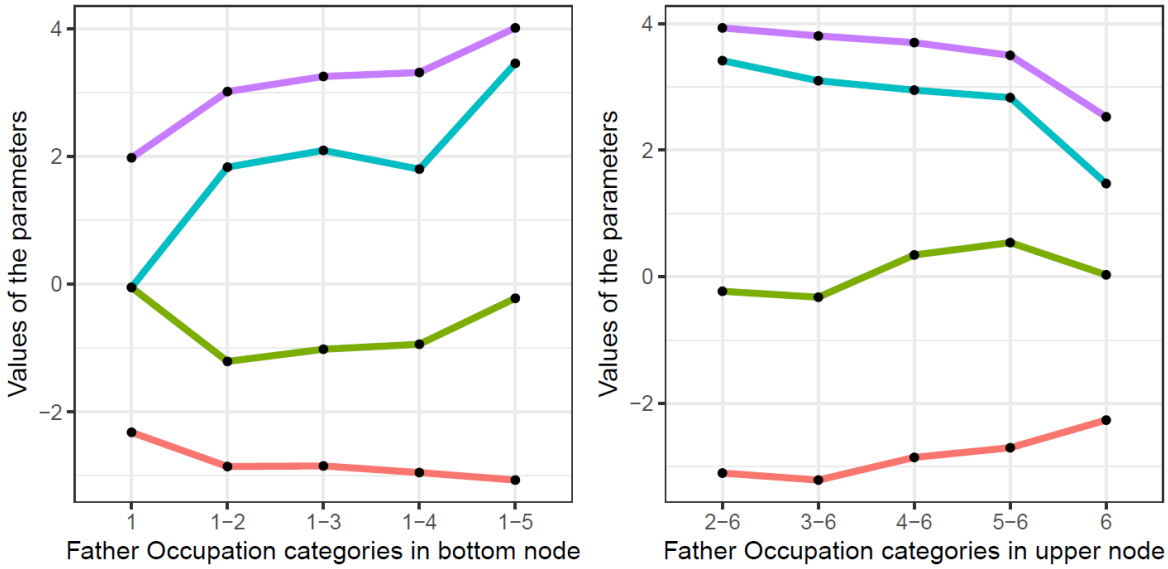
The algorithm employs an M-fluctuation test on parameters stability to allow the number of Roemerian types proliferate. Purely as an example, we show how the algorithm performs the partition in types in a simplified hypothetical case in which father's occupation is the only circumstance and the logarithm of income is the outcome of interest.[22] 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 categories ordered from smaller associated expected income to higher associated expected income.

In Figure A.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, move 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.

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

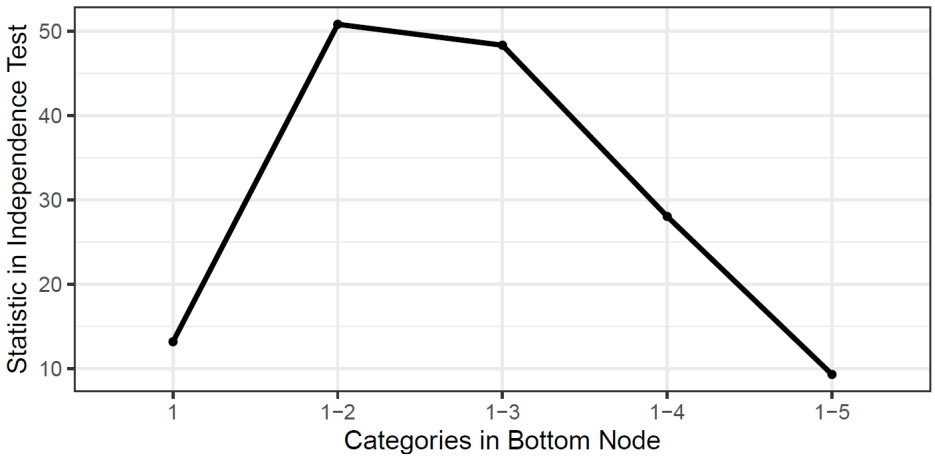
Figure A1. Values for the Parameters of the Bernstein Polynomial in each node



Source: Own Elaboration on NIDS 5

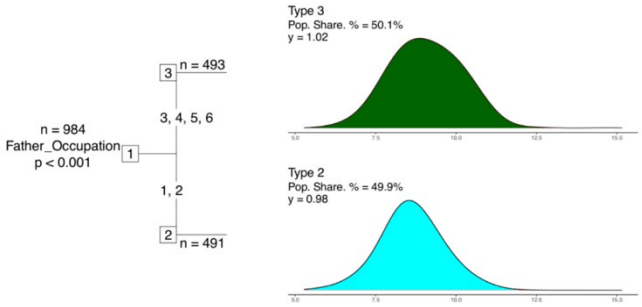
That selection is guided by the M-fluctuation test. Figure A.2 shows the value of the statistics for the tests described in step 4. The higher value (associated with a smaller p-value) is achieved when the bottom node has categories 1 and 2. That is the splitting point, as confirmed in Figure A.2. The 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

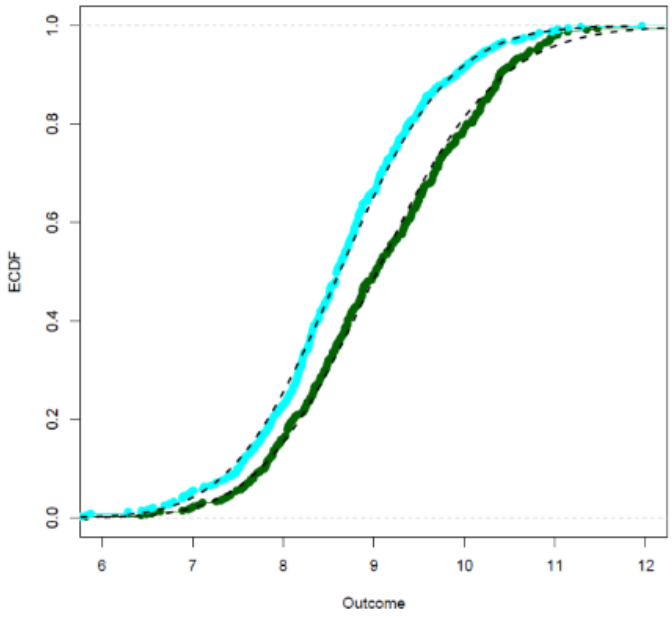
Figure A3. Transformation Tree (example)



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.

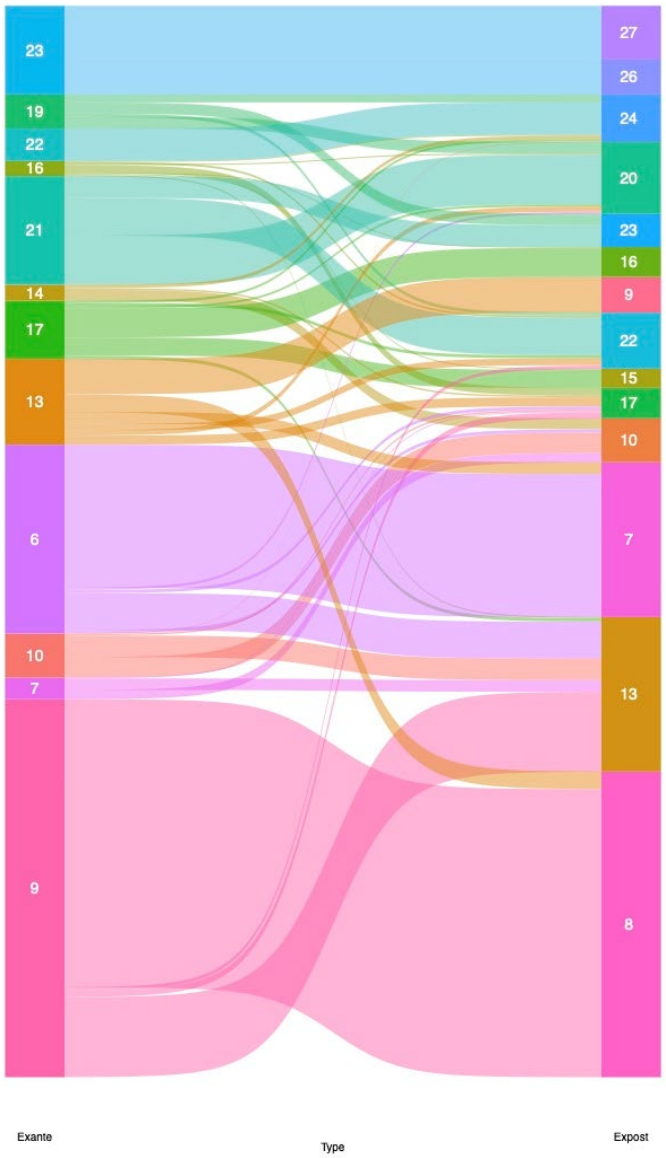
Figure A4. ECDFs (example)



Source: Own Elaboration

Appendix 3: The Sankey Plot

The Sankey plot below, also known as an alluvial diagram, connects the ex-ante and ex-post types to which each individual in the sample belongs. Ex-ante types are on the left-hand column, and ex-post types are the right. In both columns, types are ordered from higher income (top) to lower income (bottom). While for the white population the only difference between the two approaches is that the single ex-ante type is split into two by the ex-post TrT algorithm, much more movement is observed among poorer types.



Source: Own elaboration from NIDS 5.

Appendix 4: Type composition by circumstances.

Table A1: Ethnicity by ex-post types

	Circum	1	2	3	4	Type Sh.
Types	Mean	0.67	1.82	1.07	3.61	
8	0.39	28.52	0	0	0	28.52
13	0.49	11.28	0.07	3.08	0	14.43
7	0.55	14.44	0	0	0	14.44
10	0.64	0	0.9	3.19	0	4.09
17	0.65	2.29	0.1	0.47	0	2.86
15	0.75	1.36	0.03	0.4	0	1.79
22	0.77	5.25	0	0	0	5.25
9	0.8	3.3	0	0	0	3.3
16	0.93	2.01	0	0.79	0	2.81
23	1.24	3.08	0	0	0	3.08
20	1.31	6.73	0	0	0	6.73
24	1.9	0	0.81	3.62	0	4.43
26	2.76	0	0	0	3.26	3.26
27	4.11	0	0	0	5.02	5.02
	Circ Share	78.27	1.90	11.55	8.28	100

Source: Own elaboration from NIDS 5. Circumstance categories are 1: African; 2: Asian/Indian; 3: Coloured; 4: White. Circum. Stands for Circumstance Categories, Type Sh. stands for Type Shares, Circ. Sh. stands for Circumstance Shares.

Table A.2: Sex by ex-post types

	Circum	0	1	Type Sh.
Types	Mean	1.15	0.88	
8	0.39	0	28.52	28.52
13	0.49	4.71	9.72	14.43
7	0.55	14.44	0	14.44
10	0.64	1.55	2.55	4.09
17	0.65	1.1	1.75	2.86
15	0.75	0.63	1.15	1.79
22	0.77	0	5.25	5.25
9	0.8	1.25	2.06	3.3
16	0.93	0.99	1.82	2.81
23	1.24	0	3.08	3.08
20	1.31	6.73	0	6.73
24	1.9	1.88	2.55	4.43
26	2.76	1.49	1.77	3.26
27	4.11	2.37	2.64	5.02
	Circ Share	37.14	62.86	100

Source: Own elaboration from NIDS 5. Circumstance categories are 0: Female; 1: Male. Circum. Stands for Circumstance Categories, Type Sh. stands for Type Shares, Circ. Sh. stands for Circumstance Shares.

Table A.3: Father Education by ex-post types

		0	1	2	3	4	5	6	7	8	9	10	11	12	Type Sh.
Types	Mean	0.49	0.43	0.66	0.63	0.72	0.68	0.68	0.8	1.32	0.89	1.85	1.33	2.55	
8	0.39	25.35	0.23	0.37	0.55	0.64	0.47	0.44	0.47	0	0	0	0	0	28.52
13	0.49	11.46	0.15	0.37	0.45	0.59	0.4	0.38	0.63	0	0	0	0	0	14.43
7	0.55	12.29	0.12	0.19	0.33	0.44	0.34	0.45	0.27	0	0	0	0	0	14.44
10	0.64	2.7	0	0.08	0.29	0.19	0.18	0.22	0.44	0	0	0	0	0	4.09
17	0.65	1.58	0.04	0.12	0.29	0.14	0.23	0.16	0.29	0	0	0	0	0	2.86
15	0.75	0.96	0.01	0.03	0.12	0.1	0.14	0.16	0.26	0	0	0	0	0	1.79
22	0.77	0	0	0	0	0	0	0	0	2.78	0.74	1.01	0.71	0	5.25
9	0.8	1.75	0.04	0.19	0.1	0.15	0.36	0.38	0.33	0	0	0	0	0	3.3
16	0.93	1.16	0.07	0.15	0.19	0.27	0.18	0.33	0.45	0	0	0	0	0	2.81
23	1.24	0	0	0	0	0	0	0	0	0	0	0	0	3.08	3.08
20	1.31	0	0	0	0	0	0	0	0	1.93	0.53	0.89	0.48	2.89	6.73
24	1.9	0	0	0	0	0	0	0	0	1.47	0.41	1.01	0.25	1.29	4.43
26	2.76	0.12	0	0.01	0	0.01	0.08	0.1	0.15	1.16	0.14	1.48	0	0	3.26
27	4.11	0	0	0	0	0	0	0	0	0	0	0	0.16	4.85	5.02
	Circ Share	57.38	0.67	1.52	2.32	2.53	2.37	2.63	3.29	7.35	1.82	4.4	1.6	12.11	100

Source: Own elaboration from NIDS 5. Circumstance categories are 0: Non-Educated, Then the remaining values correspond to Grades from 1 to 12 (or more). Circum. Stands for Circumstance Categories, Type Sh. stands for Type Shares, Circ. Sh. stands for Circumstance Shares.

Table A.4: Mother Education by ex-post types

	Circum	0	1	2	3	4	5	6	7	8	9	10	11	12	Type Sh.
Types	Mean	0.48	0.53	0.58	0.56	0.74	1.02	0.69	1.01	1.34	1.19	1.85	1.05	2.51	
8	0.39	24.49	0.19	0.44	0.66	1.11	0	0.58	0.6	0	0.25	0	0.21	0	28.52
13	0.49	11.55	0.25	0.53	0.63	1.04	0	0	0	0	0.21	0	0	0.22	14.43
7	0.55	12.37	0.08	0.11	0.36	0.42	0	0.33	0.47	0	0.16	0	0.14	0	14.44
10	0.64	2.82	0.03	0.04	0.1	0.1	0.25	0.12	0.29	0.21	0	0.1	0	0.05	4.09
17	0.65	0.77	0.01	0.07	0.19	0.08	0.05	0.22	0.21	0.44	0.11	0.19	0.14	0.37	2.86
15	0.75	0	0	0	0	0	0.33	0.29	0.47	0.47	0	0.12	0.11	0	1.79
22	0.77	1.01	0.03	0.08	0.18	0.23	0.19	0.42	0.38	1.26	0.27	0.45	0.3	0.42	5.25
9	0.8	0	0	0	0	0	0.89	0	0	1.44	0	0.51	0	0.47	3.3
16	0.93	0	0	0	0	0	0.55	0.44	0.73	0.78	0	0.26	0.05	0	2.81
23	1.24	0.23	0	0.03	0.04	0.05	0.04	0.05	0.14	0.18	0.08	0.27	0.33	1.63	3.08
20	1.31	0.95	0	0.07	0.07	0.12	0.22	0.27	0.42	0.99	0.37	0.67	0.48	2.1	6.73
24	1.9	0.37	0	0.1	0.16	0.21	0.18	0.22	0.44	1.03	0.26	0.49	0.12	0.85	4.43
26	2.76	0.11	0	0	0	0	0.01	0.08	0.15	0.85	0.12	1.06	0.05	0.82	3.26
27	4.11	0.05	0	0	0	0.01	0	0	0.03	0.19	0.07	0.64	0.1	3.92	5.02
	Circ Share	54.73	0.59	1.47	2.38	3.38	2.71	3.03	4.32	7.83	1.9	4.77	2.03	10.85	100

Source: Own elaboration from NIDS 5. Circumstance categories are 0: Non-Educated, Then the remaining values correspond to Grades from 1 to 12 (or more) . Circum. Stands for Circumstance Categories, Type Sh. stands for Type Shares, Circ. Sh. stands for Circumstance Shares.

Table A.5: Father Occupation by ex-post types

	Circum	0	1	2	3	4	5	6	7	8	9	10	Type Sh.
Types	Mean	1.05	2.59	2.37	2.09	2.37	1.17	0.82	1.33	0.7	0.86	0.59	
8	0.39	0.12	0.14	0.14	0.22	0.03	1.04	0.12	1.27	2.55	4.48	18.4	28.52
13	0.49	0.04	0.11	0.08	0.11	0.08	0.74	0.16	1.74	2.01	5.99	3.36	14.43
7	0.55	0.12	0.14	0.07	0.15	0.04	0.77	0.11	0.69	1.66	2.56	8.14	14.44
10	0.64	0	0.03	0.04	0.01	0	0.18	0.05	0.78	0.34	1.36	1.3	4.09
17	0.65	0.01	0.03	0.08	0.1	0.01	0.33	0.04	0.48	0.51	0.71	0.55	2.86
15	0.75	0.03	0	0.01	0.08	0	0	0	0.69	0	0	0.97	1.79
22	0.77	0.03	0.1	0.19	0.11	0.08	0.63	0	0.7	1.18	0.77	1.47	5.25
9	0.8	0.01	0.05	0.01	0.03	0.01	0.16	0.05	0.32	0.53	0.42	1.69	3.3
16	0.93	0	0.05	0	0	0.03	0.32	0.03	0	0.9	1.48	0	2.81
23	1.24	0.04	0.23	0.74	0.1	0.1	0.48	0.01	0.19	0.42	0.19	0.58	3.08
20	1.31	0.07	0.45	0.88	0.32	0.08	0.89	0.01	0.92	0.96	0.74	1.41	6.73
24	1.9	0.04	0.19	0.42	0.16	0.14	0.37	0.08	1.15	0.48	0.71	0.67	4.43
26	2.76	0.03	0.26	0.08	0.18	0.07	0.27	0.1	0.97	0.4	0.77	0.14	3.26
27	4.11	0.03	0.86	1.4	0.53	0.23	0.25	0.1	0.75	0.21	0.4	0.26	5.02
	Circ Share	0.58	2.64	4.15	2.10	0.9	6.43	0.88	10.65	12.16	20.58	38.93	100

Source: Own elaboration from NIDS 5. Circumstance categories are 0: Army; 1: Managers; 2: Professionals; 3: Technicians; 4: Clerks; 5: Service; 6: Skilled; 7: Craft; 8: Operators; 9: Elementary; 10 Others. Circum. Stands for Circumstance Categories, Type Sh. stands for Type Shares, Circ. Sh. stands for Circumstance Shares.

Table A.6: Mother Occupation by ex-post types

	Circum	0	1	2	3	4	5	6	7	8	9	10	Type Sh.
Types	Mean	0.24	2.07	2.69	2.9	2.98	1.4	0.68	1.06	1.81	0.72	0.72	
8	0.39	0	0	0	0	0	0	0.08	0	0	0	28.44	28.52
13	0.49	0	0.04	0	0.18	0.07	0	0	0	0	14.14	0	14.43
7	0.55	0	0	0	0	0	0	0.05	0	0	0	14.39	14.44
10	0.64	0	0	0	0	0	0	0.05	0	0	0	4.04	4.09
17	0.65	0	0	0.82	0	0	1.34	0	0.6	0.08	0	0	2.86
15	0.75	0	0.01	0	0.04	0.04	0	0	0	0	1.69	0	1.79
22	0.77	0	0	0.37	0.07	0.07	0.22	0	0.11	0.04	1.51	2.86	5.25
9	0.8	0.01	0	0	0	0	0	0.01	0	0	0	3.28	3.3
16	0.93	0	0	0	0.05	0.03	0	0	0	0	2.73	0	2.81
23	1.24	0	0.07	0.86	0.1	0.11	0.18	0	0.1	0.01	0.42	1.23	3.08
20	1.31	0	0.07	0.97	0.11	0.12	0.38	0.01	0.15	0.05	1.73	3.12	6.73
24	1.9	0	0.08	0.52	0.14	0.11	0.37	0	0.18	0.07	1.25	1.71	4.43
26	2.76	0	0.1	0.52	0.18	0.48	0.3	0	0.12	0	0.32	1.25	3.26
27	4.11	0	0.21	1.73	0.53	0.66	0.3	0	0.08	0	0.29	1.22	5.02
	Circ Share	0.01	0.58	5.8	1.4	1.69	3.10	0.22	1.34	0.26	24.06	61.55	100

Source: Own elaboration from NIDS 5. Circumstance categories are 0: Army; 1: Managers; 2: Professionals; 3: Technicians; 4: Clerks; 5: Service; 6: Skilled; 7: Craft; 8: Operators; 9: Elementary; 10 Do not work. Circum. Stands for Circumstance Categories, Type Sh. stands for Type Shares, Circ. Sh. stands for Circumstance Shares