LSE International Inequalities Institute

WORKING PAPER 158

Global Estimates of Opportunity and Mobility: A Database

Francisco H.G. Ferreira; Vito Peragine; Paolo Brunori; Pedro Salas-Rojo; Domenico Moramarco; Luis Barajas; Teresa Barbieri; Nancy Daza-Báez; Gaurav Datt; Vito de Sandi; Fabio Farella; Arturo Martinez Jr.; John Nguyen; Albert Park; Enza Simeone; Louis Sirugue; Pedro Torres López; Giorgia Zotti

JANUARY 2026

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

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

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

International
Inequalities Institute

Global Estimates of Opportunity and Mobility: A Database

Francisco H.G. Ferreira, Vito Peragine[1], Paolo Brunori, Pedro Salas-Rojo, Domenico Moramarco, Luis Barajas, Teresa Barbieri, Nancy Daza-Báez, Gaurav Datt, Vito de Sandi, Fabio Farella, Arturo Martinez Jr., John Nguyen, Albert Park, Enza Simeone, Louis Sirugue, Pedro Torres López, Giorgia Zotti.

11 January 2026

Abstract: This paper describes a new public-access online database containing internationally comparable estimates of inequality of opportunity for seventy-two countries, covering two-thirds of the world's population. The estimates were computed directly from the unit-record microdata for 196 household surveys, using a suite of machine-learning tools selected to minimize the omitted variable and overfitting biases discussed in the literature. Overall, differences in opportunities account for substantial shares of total income inequality (with the mean of our preferred estimate being 40.9%), but there is substantial variation across countries, with estimates ranging from 18.9% in Denmark (2011) to 76.7% in South Africa (2017). The latest US estimate of 41.6% places it among the most opportunity unequal high-income countries. We also find strong support for the existence of a positive association between income inequality and relative inequality of opportunity, analogous to the "Great Gatsby Curve" for mobility and inequality. Similarly, there is evidence of an inverted-U "Opportunity Kuznets curve". The database is available at www.geom.ecineq.org.

Keywords: Inequality of opportunity; mobility; machine learning.

JEL Codes: D31, D63, I39

[1] Correspondence to vitorocco.peragine@uniba.it. The authors are grateful to the VélezReyes+ Plataforma Filantrópica for generous financial support, as well as for technical support with website design. We are grateful to Elisa Mansur and her team at VélezReyes+ for many useful discussions. All errors are our own.

1. Introduction

This paper introduces and describes a new public-access dataset containing summary estimates of inequality of opportunity and ancillary information for seventy-two countries over the last two to four decades. In line with the recent literature, inequality of opportunity (IOp) is defined as either the amount (absolute) or the share (relative) of the dispersion in the distribution of an outcome (such as household incomes) that can be predicted by differences in people's "circumstances" – factors beyond their control which nonetheless shape and constrain their choice sets.

This definition of inequality of opportunity draws on theoretical work by Roemer (1993, 1998), van de Gaer (1993), and Fleurbaey (1994), and has been used in an empirical literature that seeks to measure the extent of inequality of opportunity in different countries over the last two decades or so.[2] It is a literature that draws explicitly on normative principles about distributional fairness, but it is also closely related to the literature on intergenerational persistence in individual outcomes, such as education, income, or wealth, and to the sociological analysis of social stratification.

Björklund and Jäntti (2020), for example, include inequality of opportunity as one of four established approaches "to the study of how individuals' income and education during adulthood are related to their family background" (p.1), with the other three being intergenerational mobility (IGM); causal intergenerational effects; and sibling correlations. Brunori et al. (2013) and Corak (2013) also highlight the connections between intergenerational (im)mobility and inequality of opportunity and illustrate how the two are correlated in practice.

A key difference between these two approaches – IGM and IOp – is that, whereas mobility estimates are basically about quantifying or describing the bivariate associations between a parent's outcome and that of their children, IOp takes a broader, multivariate approach: parental income is an excellent candidate circumstance variable, but so are parental education and occupation, place of birth, race, gender, and the neighborhood in which a person grows up, to mention only a few.[3] IOp is then assessed as the level (or share) of inequality that can be predicted by all these circumstances together. In a recent study using administrative data for Sweden, Adermon, Brandén and Nybom (2025) find that, although IOp and IGM estimates correlate strongly, "the share of total inequality that can be attributed to family background factors is substantially higher for the sibling correlation and the IOp indices than what is implied by intergenerational estimates" (p.18).

[2] See Bourguignon et al. (2007) and Checchi and Peragine (2010) for some of the first empirical studies of inequality of opportunity.

[3] In that sense, it is closer to the sibling correlation approach, which estimates the share of dispersion in adult incomes that can be 'explained' by all factors shared by siblings. However, as Björklund and Jäntti (2020) note, IOp could in principle also include circumstances that differ among siblings, such as birth-order, different pre-schools attended, or the siblings' ages when certain exogenous shocks hit the family.

It is increasingly recognized that this kind of inequality matters both intrinsically and instrumentally. Evidence from both opinion surveys and behavioral experiments suggests that people object most strongly to inequality they see as arising from factors independent of people's choices and efforts (e.g. Konow, 2000; Almås et al., 2010; Cappelen et al., 2010, 2013; Pew Research Center, 2012). Instrumentally, very different kinds of evidence indicate that unequal opportunities may reduce economic efficiency and growth (e.g., Hsieh et al., 2019; Marrero and Rodríguez, 2013).

Yet, although there are now a few cross-country datasets that allow for international comparisons of intergenerational mobility estimates, particularly for educational attainment – e.g., Neidhöfer et al. (2018) and van der Weide et al. (2024) – we are not aware of any large comparable datasets for inequality of opportunity across countries.[4] This is the gap the Global Estimates of Opportunity and Mobility Database (GEOM: www.geom.ecineq.org) seeks to fill. Drawing on the unit-record level data from 196 representative household surveys we estimate inequality of opportunity for 72 countries around the world, representing just over two-thirds of the world's population, over a period spanning up to 44 years.

For each of our 196 country-year data points we provide estimates for two different IOp concepts – known as ex-ante and ex-post IOp and defined in Section 2 below – in each case using both the Gini coefficient and the mean logarithmic deviation (MLD) as scalar measures of inequality. While we report all these estimates below, we focus on the ex-ante results in this paper for brevity. For these ex-ante estimates and across the entire database, we find Gini coefficients ranging from 0.05 (for Denmark in 2011) to 0.48 (for South Africa in 2008). If we consider only the latest available year for each country, the range is from 0.07 in Denmark 2019 to 0.47 in South Africa 2017. Relative to the country's own income inequality, IOp accounts for as little as 18.9% of total inequality in Denmark (2011), to as much as 76.7% in South Africa (2017). There is considerable variation across regions, with Latin America substantially overrepresented at the top of the range, while Europe is overrepresented at the bottom. There is much more dispersion across Asia, and it is difficult to compare Africa to the rest of the world, since most African countries use consumption, rather than income, as the main measure of economic advantage. The latest estimate for the United States is 41.6% of total inequality, slightly above the mean value of 40.9% and quite a bit higher than the median (38.4%)

We also find considerable variation in trends over time with, for example, a substantial increase in absolute IOp in the United States from 1978 to 2002 (followed by a slight decline), contrasting with a decline in Peru between 2007 and 2015. We are also able to confirm the existence of an opportunity "Great Gatsby Curve" – a positive association between IOp and cross-sectional inequality – much as found by Brunori et al. (2013) for a much smaller and less comparable earlier sample, and analogously to the original Gatsby Curve for intergenerational

[4] There is also less internationally comparable information on IGM in incomes than in education, although Muñoz and van der Weide (2025) begin to close that gap.

mobility (Corak, 2013). Similarly, we explore whether an “Opportunity Kuznets Curve” may be discerned in the cross-sectional data.

In addition to comparable headline IOp estimates, our estimation methods also allow us to report the relative descriptive importance of individual circumstance variables, such as mother’s education or ethnicity; and the population partitions selected for each country-year by our data-driven prediction algorithms – including the richest and poorest social groups as defined by circumstances. While there is no room in this paper to do justice to all these byproducts of the estimation strategy, individual country results are available from the online database.[5]

The remainder of the paper is organized as follows. Section 2 briefly reviews the theoretical framework underpinning the IOp literature. Section 3 describes the data sources used for the analysis, as well as our treatment of the samples. It defines the income and circumstance variables employed and provides some broad summary statistics across the dataset. Section 4 describes the methods used to predict current-generation incomes from circumstance variables, both from an ex-ante and from an ex-post perspective. Section 5 presents an overview of results, focusing on international comparisons. Section 6 concludes.

2. Theoretical framework

The canonical model used in the literature to measure IOp can be described as follows.[6] Consider a distribution of an outcome x in a given population and suppose that all determinants of x can be classified into either a set of *circumstances* C that lie beyond individual control, or as responsibility characteristics, summarized by a variable e, denoting *effort*, belonging to the set $\mathbf{E}$. Circumstances belong to a finite set $\mathbf{C}$. The outcome of interest is then generated by a function $g\colon \mathbf{C} \times \mathbf{E} \to \mathbb{R}$, such that:

$$x = g(C, e) \tag{1}$$

In this framework, each individual in the population is fully characterized by the triple (x, C, e). The population can then be exhaustively partitioned in two ways: into types $\{T_1^C, T_2^C, \ldots T_n^C\}$, which are groups of individuals that share the same circumstances, or into tranches $\{T_1^e, T_2^e, \ldots T_m^e\}$, which are groups within which everyone shares the same degree of effort.

[5] The interested reader can reproduce all graphs shown in the results section by downloading the data or interacting with the platform. Information about the team and institutions involved in the GEOM database, as well as a glossary and a documentation section are also available on the website, aimed at providing readers with the necessary tools to explore the complete results.

[6] Different variants of this model were proposed in theoretical contributions by Fleurbaey (1994), Roemer (1993), Van de Gaer (1993), Peragine (2002) and used in the first empirical analyses of inequality of opportunity: see Bourguignon et al. (2007), Checchi and Peragine (2010) and Ferreira and Gignoux (2011). See Ferreira and Peragine (2016) and Roemer and Trannoy (2016) for reviews of the literature.

This is a reduced-form model in which neither opportunities themselves, nor the structural process by which outcomes are determined, are explicitly modelled. The idea is to infer the differences in opportunities available to individuals by observing differences in the distributions of the outcome variable conditional on different combinations of circumstances – that is across type-specific outcome distributions. If circumstances ought not to influence outcomes – either directly or through their influence on efforts – any differences across conditional distributions are *prima-facie* evidence of inequality of opportunity (Ferreira and Peragine, 2016).

The normative foundations of this opportunity egalitarian theory rest on two distinct and independent principles: the *Compensation Principle*, according to which all outcome inequalities due to circumstances are unfair and should be compensated by society; and the *Reward Principle*, which is concerned with the apportionment of outcomes to effort and, in some of its formulations, requires that outcome inequalities due to effort be respected. Two main versions of the compensation principle have been proposed, each yielding a different approach to the measurement of inequality of opportunity: the *ex ante* and the *ex post* approaches.

According to the ex-ante approach, there is equality of opportunity if the *values* of the sets of opportunities available to individuals are the same for everyone, regardless of circumstances (*ex-ante compensation*). In the model introduced above, the support of a type's (T_i^C) outcome distribution, which is the outcome distribution conditional on circumstances C_i, is interpreted as the opportunity set of all individuals with circumstances C_i. There are obviously many ways in which such a set could be valued – one of which is to take its expected value. Hence the focus is on the inequality between the types. This approach is ex ante with respect to the revelation of effort (van de Gaer, 1993).

On the other hand, in the ex-post approach, there is equality of opportunity if and only if all those who exert the same degree of effort end up with the same outcome. Because effort is difficult to observe and because its absolute level is likely to be influenced by circumstances, Roemer's identification assumption is commonly adopted. This assumption identifies the relative degree of an individual's effort by the person's *rank* in the type-specific outcome distribution. In this case, tranches (e.g., T_j^e) are defined as sets of individuals who belong to the same quantiles in their respective type distributions, and the ex-post principle of compensation requires reducing outcome inequality within tranches (*ex-post compensation*). This means that inequality of opportunity within this approach is measured as inequality within tranches (Roemer, 1998).

As far as the *reward* principle is concerned, different versions of the principle have been proposed in the literature, expressing different attitudes to the outcome inequality observed within types, that is: among individuals endowed with the same circumstances. Among the many interpretations, the most common in empirical applications is *utilitarian reward*, according to which one should focus only on the sum (or the average) of the achievements obtained by each group of individuals sharing the same circumstances, and to remain neutral with respect to the way differences in effort are remunerated within these groups. Alternative formulations have been proposed, from inequality-averse reward (Ramos and Van de Gaer, 2016; Fleurbaey et al. 2024) which incorporates some aversion to inequality even within types, to intermediate and agnostic positions (Peragine, 2002; Peragine and Serlenga, 2008; and Fleurbaey and Peragine 2013).

Measuring IOp

Once a version of the compensation and a version of the reward principle are adopted, the derivation of a scalar measure of inequality of opportunity follows a two-step procedure: first, the actual distribution, call it $\boldsymbol{x}$, is transformed into a counterfactual distribution $\tilde{\boldsymbol{x}}$**,** which reflects only and fully the unfair inequality in $\boldsymbol{x}$, while all the fair inequality is removed. In the second step, a measure of inequality is applied to $\tilde{\boldsymbol{x}}$. The first step is where the choice of compensation and reward principles matter. In fact, different versions of the counterfactual distribution, and hence different measures of inequality of opportunity, which are either consistent with the ex-ante or the ex-post compensation and with different versions of reward, have been proposed in the literature: they express different and sometimes conflicting views on equality of opportunity and the distributional rankings they generate may be different. See Ferreira and Peragine (2016) for a discussion.

One measure extensively used in the literature is *Between-Types* inequality, which arises from the combination of ex-ante compensation with utilitarian reward. Taken together, these two versions of the principles imply valuing opportunity sets using their mean or expected value, and computing inequality on a counterfactual distribution $\tilde{\mathbf{x}}_{BT}$ that is obtained by replacing each individual outcome by the average outcome of the type the individual belongs to. This smoothing transformation is intended to remove all inequality within types, and different applications were implemented by Bourguignon et al. (2007), Checchi and Peragine (2010), Ferreira and Gignoux (2011), and others.[7]

An alternative, ex-post measure, inspired by Roemer (1993) and implemented by Checchi and Peragine (2010) and Aaberge et al (2011), is based on the *Within-Tranches* counterfactual distribution ($\tilde{\boldsymbol{x}}_{WTR}$). This distribution is obtained by replacing each individual outcome in each tranche with the ratio between that outcome and the average outcome of the tranche. This

[7] Although the utilitarian reward principle implies the between-types approach directly, because of the use of average incomes to value type opportunity sets, the approach is also consistent with other reward principles, such as *liberal reward*, for example.

normalization procedure is intended to remove all inequalities between tranches and to leave unchanged the inequality within tranches.

In both approaches, the fact that estimation typically occurs in samples, rather than entire populations, has important practical implications. Even with a relatively narrow set of circumstances – such as the one described in the next section – the interactions between the different categories across all variables routinely reach into the thousands, leading to the possibility of severely overfitted models. This gives rise to a tradeoff between two different kinds of bias in selecting a prediction model $\tilde{x}_i = f(C_i)$ to construct the counterfactual distribution $\widetilde{\boldsymbol{x}}$: Include too few variables and interactions, and the model will suffer from (downward) omitted variable bias. Include too many, and the model will suffer from (upward) overfitting bias.[8]

In the past, different parametric and non-parametric prediction methods were used but, until recently, these generally relied on arbitrary or *ad hoc* specifications. We address this challenge by using a suite of data-driven supervised machine learning techniques to select optimal prediction models, in the sense that the algorithms select partitions to maximize out-of-sample predictive power (following Brunori, Hufe and Mahler, 2023, and Brunori, Ferreira and Salas-Rojo, 2024). We use these approaches to partition country samples into empirical types and discuss them in more detail in Section 4.

Once the counterfactual distribution has been obtained, either in the ex-ante or in the ex-post versions, the second step of the measurement procedure can take place. Here, a specific inequality index $I(.)$ is applied to the counterfactual matrix to obtain an estimate of inequality of opportunity. The Gini coefficient and the Mean Logarithmic Deviation (MLD) are commonly used in the literature and the GEOM database aligns with this practice.

Two closely related versions of the IOp index are reported below. The first one is the absolute or level estimate of inequality of opportunity ($\mathrm{IO_A}$), given simply by the inequality measure computed over $\widetilde{\boldsymbol{x}}$, i.e. by $I(\widetilde{\boldsymbol{x}})$. The second measure is the ratio of $\mathrm{IO_A}$ to overall inequality in the relevant outcome variable (e.g. income), which yields the relative measure, $\mathrm{IO_R}$:

$$\mathrm{IO_R} = \frac{\mathrm{I}(\tilde{\mathrm{x}})}{\mathrm{I}(\mathrm{x})} \tag{2}$$

[8] Because of this trade-off, the choice of the empirical prediction model involves not only selecting a functional form and specification that adequately capture the desired principles of compensation and reward, but also selecting the partition into empirical types, which is generally coarser than the partition into theoretical types. As Ferreira and Brunori (2024) note: “The choice of the empirical partition $\|\hat{T}_i\|$ is an important component of the model selection problem, and it involves a trade-off between two different kinds of bias that work in opposite directions. The first is an omitted variable bias: selecting a partition $\|\hat{T}_i\|$ with too few empirical types (a low $n(\hat{T}_i)$) leads to an underestimate of IOp or inherited inequality, relative to the true theoretical partition (Ferreira and Gignoux, 2011). On the other hand, overfitting the sample data and choosing too large a $n(\hat{T}_i)$ can lead to an upward bias in estimates of IOp (Brunori, Peragine and Serlenga, 2018)” (p.16).

$\mathrm{IO_R}$ can be interpreted as the share of total inequality that can be predicted by circumstances or inherited characteristics which people cannot be held responsible for.

3. Data sources

The sample of countries

As noted in the Introduction, GEOM includes data from 72 countries, drawn from 196 household surveys. These surveys were selected with a view to balancing two key desiderata, namely (i) broad country coverage and (ii) as much data and methodological comparability as possible. We therefore restrict our attention to household surveys containing basically the same set of circumstance variables, define the outcome variable in the same way, and treat the original samples identically regarding issues such as missing information, income outliers, age ranges, equivalence scales and so on.

This pursuit of reasonable comparability across a wide range of countries inevitably has a cost in terms of the types of data and the range of circumstance variables that can be used in each country. In most developing countries, for example, large administrative datasets that combine social security records across generations and match parents to their children do not exist. Even in the rare instances where they do exist, they exclude, by definition, the typically large informal sectors of those economies. Even among surveys, the extent of information on circumstances available in detailed panel surveys, such as the Panel Study of Income Dynamics (PSID) in the US or the German Socioeconomic Panel (G-SOEP), is extremely rare among developing countries. Yet, as we will see, developing countries tend to have high levels of inequality of opportunity – higher than most developed countries. To exclude them based on data limitations would be a serious case of 'looking where the light is, rather than where the problem is'.

Table A1 in the Appendix lists all countries, time periods, and surveys included in GEOM, as well as the source from which we obtained access to the microdata.[9] The time coverage varies significantly: for some countries, such as Armenia (2016), Colombia (2010), or Mali (2019), we only have data for a single point in time, while for others, like the USA, Peru or Australia, we have data for eight years or more. Since we use EU-SILC data for 2005, 2011, and 2019, most European countries have data for three points in time.[10]

Table 1 summarizes the coverage information from Table A1 in more synthetic form. In Panel A we use the World Bank's geographical region classification and display the total number of countries in each region, the number of countries included in GEOM, the share as a

[9] Some microdata sets were obtained from or cleaned by institutional partners, such as the Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS) at the University of La Plata, the Centro de Estudios Espinosa Yglesias (CEEY), Monash University and the Asian Development Bank (ADB). The list in Table A1 in the Appendix corresponds to the countries and time periods available in GEOM Version 1 (June 2024). Future updates are expected to include estimates for additional countries and time periods.

[10] A few European countries have fewer time points due to sample size or data limitations, such as Sweden (2019) and Malta (2011, 2019).

percentage of the total, and the share of the region's population covered.[11] GEOM includes estimates of inequality of opportunity for 72 out of 217 countries globally, representing 66.9% of the world's population in 2022. While regions such as North America, South Asia, Latin America and Caribbean, and East Asia and Pacific have a broad coverage, encompassing more than 75% of their population, others, like the Middle East and North Africa (0.1%), have a much lower rate. In Panel B, countries are grouped by the World Bank's income range classification. The high-income and upper-middle-income categories are well-covered in terms of population, with coverage rates reaching 78.4% and 82.3%, respectively. In contrast, GEOM covers only 22.9% of the population in low-income countries, highlighting the need for increased representation in these areas.

Table 1: GEOM Coverage

Panel A

Macro Region (Geographical)	Number of Countries (WB)	Number of Countries (GEOM)	Share of Countries (%)	Share of Population (%)
East Asia and Pacific	37	6	16.22	75.26
Europe and Central Asia	58	36	62.07	65.96
Latin America and Caribbean	42	10	23.81	82.9
Middle East and North Africa	21	1	4.76	0.11
North America[12]	3	1	33.33	89.53
South Asia	8	2	25	75.43
Sub-Saharan Africa	48	16	33.33	49.34
Total	*217*	*72*	*33.18*	*66.89*

Panel B

Macro Region (Economic)	Number of Countries (WB)	Number of Countries (GEOM)	Share of Countries (%)	Share of Population (%)
High Income	82	35	42.68	78.4
Upper-middle Income	54	14	25.93	82.26
Lower-middle Income	54	14	25.93	59.36
Low Income	26	9	34.62	22.93

[11] The classification of countries was retrieved from the World Bank Country and Lending Groups on the 1st of September 2023: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

[12] Mexico is included in Latin America and Caribbean. North America includes Canada, the United States and Bermuda.

Total	*217*	*72*	*33.18*	*66.89*

Source: Own elaboration. The classification of countries was retrieved from the World Bank Country and Lending Groups. The population data refers to 2022 and comes from the World Bank Open Data Repository (Variable SP.POP.TOTL).

Country and survey selection for inclusion in GEOM followed four main criteria: First, the sample had to be representative of the entire country.[13] Second, it must contain information on a money-metric wellbeing indicator, such as income or consumption expenditures, at the individual or household level.[14] Third, it also had to contain individual-level information on at least five of the following seven "circumstance", or inherited characteristic, variables:

- Area or Region of Birth.
- Sex.
- Ethnicity. [15]
- Occupation of the Father and/or Mother.
- Education of the Father and/or Mother.

The fourth is a minimum sample size criterion: to be included in GEOM a sample must have a minimum of 1,500 individual observations with complete information and strictly positive outcomes. This sample size is based on the share of observations used in the resampling process for random forests and Shapley value decompositions, which is set at 0.632 (see the next section). To ensure that we retain at least 1,000 observations after each resampling, we calculate the required number of observations as 1,000/0.632 ≈ 1,500.

Sample definitions within countries and key variables

That is how our sample *of* countries was selected. Now we turn to how final analysis samples (FAS) were defined and constructed *within* countries. First, we restrict the sample to individuals aged 18 or older.[16] Second, to be included in the FAS, observations must have complete information for the selected circumstances and the outcome. To simplify the analysis, categorical variables are limited to a maximum of 25 values, so if a circumstance variable has 26 or more categories, they are merged until 25 different values are reached. This merging process is country-specific and depends on the nature of the variable. For example, the Area of birth in China (2018) was initially provided in 32 categories, but we recoded

[13] The only exception to this is Argentina, where the 2014 ENES is representative only of the country's urban areas. In 2022, the urban share of Argentina's population was 92% (World Bank).

[14] This rules out, for example, the use of the Demographic and Health Surveys (DHS).

[15] In a few cases the "ethnicity" circumstance is proxied by religion, the language spoken at home, or similar definitions. These proxies are used when explicit ethnicity data is not available but other variables are sufficiently correlated with ethnic or cultural identity.

[16] This decision is based on the age of consent or responsibility, which is typically 18 in many countries and is associated with the legal right to vote and be tried as an adult. The dominant view in the literature is that any inequality observed among children younger than the age of consent is inequality of opportunity. See, e.g. Hufe et al. 2017.

provinces with fewer than 300 observations into an "*Others*" category. For each country and year, GEOM provides a dedicated file within the "Country Profile" section containing comprehensive data documentation. This includes information about the survey, sample characteristics, weights, outcome and circumstances definitions (including any merging of categories), main descriptive statistics, and a missing data analysis.

The dependent (money-metric wellbeing) variable is defined to measure current monetary well-being and calculated as total household disposable income, or total household consumption expenditure, divided by the square root of household size to account for scale equivalence. The resulting *equivalized household disposable income* (income, henceforth) measure is preferred, but *equivalized household consumption expenditure* is used when appropriate income data is unavailable.[17] The unit of analysis is the individual, so our analysis is of a distribution of equivalized household disposable income per individual. One implication of this is that any intrahousehold inequality is ignored, with evident implications for the importance of the sex circumstance variable.

Before estimation we adjust equivalized household income (or consumption) to control for systematic correlations between life cycle and the outcome distribution (Solon 1992). We use a regression approach where, for each observation *i*, we subtract from the dependent variable y_i the predictions obtained from the regression of (log) dependent variable on age and age squared (Brunori et al., 2023), as follows.[18]

$$Ln(y_j) = \alpha + \beta age_j + \gamma age_j^2 + \varepsilon_j \tag{4}$$

$$y_i^{adj} = ex\,p\big(Ln(y_i) - \hat{\beta} age_i - \hat{\gamma} age_i^2\big) \tag{5}$$

All inequality measures we report are scale-invariant, but the database does contain scale-sensitive information, such as the mean incomes of different types. To enable cross-country and temporal comparisons in these variables, all monetary values are expressed in 2017 US dollars after adjusting by Purchasing Power Parity (PPP) and the Consumer Price Index (CPI). We use Stata to download the PIP (Poverty and Inequality Platform at the World Bank) series for CPI and PPP.[19] If the CPI value from PIP is missing, we use the CPI series provided by the World Bank (*Consumer Price Index (2010 = 100)*, downloaded on September 5th, 2023), after

[17] We use Equivalized Household Consumption Expenditure as a dependent variable for Benin, Burkina Faso, Ivory Coast, Ghana, Guinea-Bissau, Indonesia, Mali, Malawi, Niger, Nigeria, Senegal, Sierra Leone, Togo, Timor-Leste, Tanzania, and Uganda.

[18] To avoid including young adults who may earn only a small share of the household income and thereby introduce biases in the adjustment, the age-adjustment regression is run on household heads, indexed by *j* in Equation (4). After running the regression and adjusting incomes for all individuals *i*, we rescale the adjusted incomes to match the sample mean. We use the household head as reported in the survey; if this information is unavailable, we treat the respondent as the household head.

[19] We use the PIP Stata command ("ssc install pip") and execute "pip tables, table(ppp) clear" and "pip tables, table(cpi) clear" to obtain the PIP series for PPP and CPI (base 2017=100). The version used in GEOM was downloaded on September 5th, 2023. We thank Daniel Gerszon Mahler for his help.

modifying the base year to 2017. The PPP values come directly from PIP with no further modifications. The national currencies are adjusted in this manner:

$$\frac{National\ Currency}{PPP_{2017} * \left(\frac{CPI_{year}}{100}\right)} = 2017\ USD \tag{3}$$

As regards the circumstance variables, all are categorical in nature and the definition of specific categories is inevitably specific to each country. For instance, in Brazil (2014), the original *Birth Area* variable has 27 possible values, corresponding to the 26 states plus a "foreign" category, while in Belgium (2019), it is defined by the three values available in EU-SILC data (Local, born in the European Union, and Other). A similar variation is found in the definition of parental occupation or education. In Senegal (2018), *Mother's Education* is classified into five levels, whereas in Ecuador it is defined by the number of years of education attained, ranging from 0 to 15. The variable denoting the individual's *Ethnicity* is also defined to capture country idiosyncrasies. For example, in South Africa, it takes four values based on self-reported ethnicity (African, Asian, Coloured, and White), while in Peru it includes 5 categories: White, Indigenous, Afro-descendent, mixed-race, and others.

Any threat to comparability posed by these country idiosyncrasies is substantially mitigated by the data-driven nature of the algorithms we use to partition the sample. All three approaches discussed in the next section rely on (different versions of) recursive binary partitioning, where identical statistical criteria are used to divide the sample based on circumstance variables. This allows us to avoid ad-hoc partitions while combining respect for country specificities with methodological comparability.

Finally, for certain countries, our different time periods come from panel data, where the same individuals or households are interviewed in multiple waves (e.g., Australia, South Africa and South Korea). In these cases, inconsistencies in responses regarding retrospective circumstances (such as a parent's educational attainment) can occasionally arise. To resolve these inconsistencies, we set the value to that reported in the first available wave. For example, if an individual reports in 2012 that her mother was illiterate, but in 2018 reports that her mother attended primary school, we assume the information from 2012 is correct, as it is (i) less prone to recall bias, and (ii) more likely to have held when the respondent was a young child. Additionally, when missing observations are encountered for time-invariant circumstances where information is available for the same individual from other waves of the panel, we use this available information, always prioritizing the oldest information (i.e., from earlier waves).

4. Estimation methods

As noted in Section 2, we follow Brunori, Hufe and Mahler (2023) and Brunori, Ferreira and Salas-Rojo (2024) in using data-driven supervised machine learning techniques to select our prediction models $\tilde{x}_i = f(C_i)$, which are then used to generate the counterfactual distributions $\widetilde{\boldsymbol{x}}$**,** on which inequality of opportunity is calculated.

Specifically, we employ regression trees and random forests, which have several advantages for the estimation of IOp from survey data. First, tree-based methods generate predictions by partitioning the regressor space into non-overlapping regions. This implies that individuals are assigned to mutually exclusive groups defined by the interaction of their circumstances. The use of trees is therefore quite consistent with the idea that the interaction of circumstances partitions the population into societal types that have access to different sets of opportunities (as discussed in Section 2).

A second advantage stems from the flexibility of tree-based algorithms which can, by construction, handle many predictors without the risk of overfitting. The tuning of the algorithm prevents the model from becoming too complex, which would otherwise result in noisy predictions and upward-biased estimates (Chakravarty and Eichhorn, 1994; Brunori et al., 2018). At the same time, regression trees are grown to select the partition that maximizes the ability of observable circumstances to predict the variation in income *out of sample*. This approach minimizes the risk of the downward bias frequently highlighted by researchers (e.g., Ferreira and Gignoux, 2011).

Among the many possible regression trees, we use conditional inference regression trees (Ctrees) and transformation trees (Trafotrees), introduced by Hothorn et al. (2006) and Hothorn and Zeileis (2021) respectively. The use of Ctrees and Trafotrees offers additional advantages. First, they address the bias inherent in standard recursive partitioning algorithms, which tend to overuse variables with many distinct values as splitting variables (Varian, 2014). Secondly, because they are based on a sequence of statistical tests, the resulting tree structures are more easily interpretable than standard trees and provide a formal test for the null hypothesis of equal opportunity in a population or subpopulations.

The Ctree algorithm searches for the partition that maximizes the statistical significance of differences between the *means* of the two resulting subsamples. It is therefore especially well-suited for the ex-ante approach to IOp – and the between-types version in particular – which, as discussed above, uses type means to construct the counterfactual distribution $\tilde{\mathbf{x}}_{BT}$. Conversely, the Trafotree algorithm recursively partitions the population into subsamples that differ most in terms of their *full conditional distributions*. It is therefore especially well-suited for the ex-post approach to IOp – and the within-tranches version in particular – which relies on estimates of each quantile of the conditional distributions to measure inequality within tranches.

We briefly summarize our use of the ex-ante (Ctree) algorithm below, although the reader is referred to Brunori, Hufe and Mahler (2023) for details. An analogous summary of the ex-post (Trafotree) algorithm is provided in Appendix B1, and the reader is referred to Brunori, Ferreira and Salas-Rojo (2024) for details.

Ex-ante IOp estimates in GEOM

The ex-ante Inequality of Opportunity (IOp) statistical approach adopted in GEOM was proposed by Brunori, Hufe and Mahler (2023). It employs the Conditional Inference regression Trees (Ctrees) and Conditional Inference Random Forests (CForest) developed by Hothorn, Hornik and Zeileis (2006). A Ctree is a supervised machine learning algorithm aimed at partitioning a regressor space to predict the variation of a dependent variable. The regression space is partitioned delivering a set of terminal nodes or leaves obtained by recursive binary splitting of the sample. The algorithm can be summarised as follows:

1. Set a confidence level (1- α).

2. Test the correlation between the dependent variable (outcome) and each regressor (circumstance). If for all observed regressors, the Bonferroni-adjusted *p-value* of the correlation test is higher than the chosen critical value α, exit. Otherwise, go to step 3.

3. Among all regressors in which the null hypothesis of independence is rejected, select the variable whose correlation with the outcome has the smallest *p-value* as splitting variable [c].

4. Consider how circumstance [c] can be used to partition the sample into two subsamples $[s_i, s_{-i}]$. Let S_c denote the set of all possible binary splits of the sample based on [c]. For each possible binary partition, compute the *p*-value for the null hypothesis that the mean in two sub-samples is the same ($\mathrm{p}^{[s_i, s_{-i}]}$).

5. Choose $[s_i, s_{-i}]^* = argmin_{S_c}\ \mathrm{p}^{[s_j, s_{-j}]}, \forall j$ as the most appropriate partition.

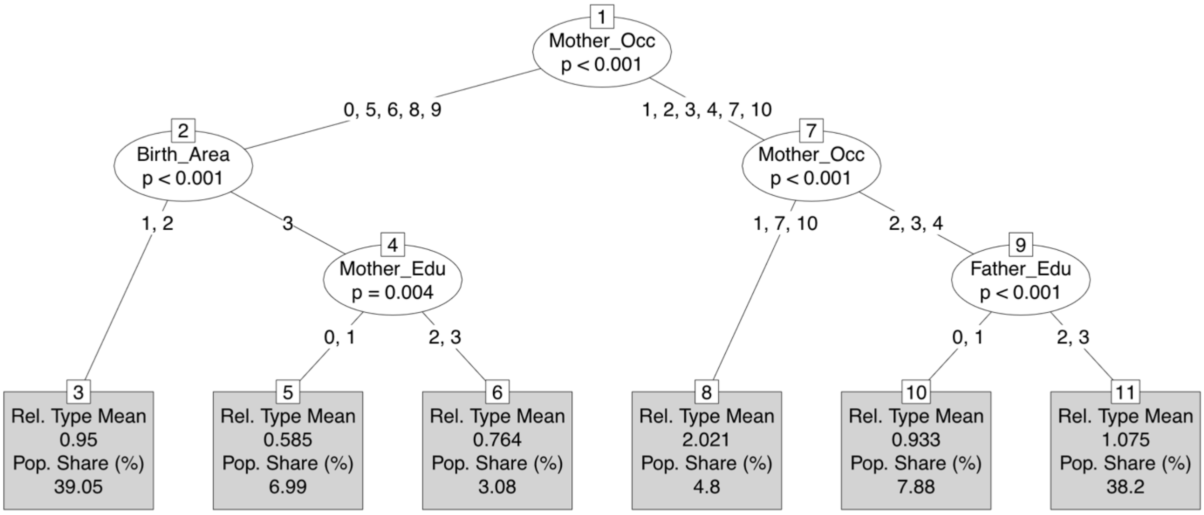
6. Repeat steps 2 – 5 for each resulting node (sub-sample) until the null hypothesis of step 2 cannot be rejected in any resulting sub-sample.

The output of this algorithm consists of an exhaustive partition of the sample into mutually exclusive groups. We treat these terminal nodes as types $t = \{1, \dots, T\}$, for which we compute the respective population weighted share $\widehat{w}_t$ and the weighted mean $\hat{\mu}_t$. Ex-ante IOp is estimated as $\widehat{IO}_a = I(\tilde{y})$, where $\tilde{y}$ is a counterfactual distribution obtained by replacing the outcome of individual *i* belonging to type *t* with $\tilde{y}_{i,t} = \hat{\mu}_t$. We select the Gini coefficient as the reference inequality measure, but we also report mean logarithmic deviation (MLD) estimates.[20]

[20] We set $\alpha = 0.01$. We impose an additional requirement, namely that each terminal node must have a minimum of 1% of the observations in the sample (or 50 if the sample size is smaller than 5000). This country-specific minimum is set to minimize the effect of different sample sizes on the depth of the tree. See Brunori, Hufe and Mahler (2023) for a discussion of the effect of sample size on IOp estimation. All remaining parameters are the default values in the "ctree" R function in the package "partykit" (Hothorn, Seibold and Zeileis, 2023). We do not use weights to determine splits. Including sampling weights expands the sample size, such that individual observations turn into hundreds or thousands of identical values. As a result, the tree becomes very deep, as null hypothesis are easily rejected. Weights are used to calculate the values of the counterfactual distribution and to estimate IOp.

In addition to producing an estimate of ex-ante IOp, the Ctree approach has the significant added benefit that the partitioning process itself contains interesting information on the structure of inequality of opportunity within a particular observed population. Ctrees in the database are displayed as in Figure 1, which is an example from Sweden, 2019. For simplicity, we normalize the expected outcome in each node dividing it by the sample mean, so a value higher than 1 can be interpreted as an expected outcome higher than the expected outcome in the entire population (e.g., type 8 in Figure 1 has an expected outcome approximately twice as large as the sample average).

Figure 1: Ctree example (Sweden, 2019).



Source: GEOM. Data from EUSILC (2019)

Like any other tree-based method, Ctrees are low-bias but high-variance learners, and therefore an aggregation procedure can improve the reliability of estimates. For this reason, we provide an alternative estimate of ex-ante IOp for each sample in the database, by aggregating 200 Ctrees into a random forest (Breiman, 2001; Hothorn, Hornik and Zeileis, 2006).[21] In the machine learning literature, when dealing with high variance learners, it is standard practice to use resampling methods. A random forest draws different subsamples of the original data and computes a tree on each one. Under the appropriate aggregation procedures, this process smooths sample dependency and generates robust IOp estimates.

[21] Following these authors, we use some default tuning parameters. In particular, we set alpha to 1 (mincriterion, $1 - \alpha = 0$), such that each tree is free to grow as much as it can. We use the default 0.632 share of each subsample drawn in every iteration. The minimum number of observations that we allow in each terminal node is 0.1% of the sample size, with the aim of maximizing comparability across surveys with different sample size (or 10, if the sample size was smaller than 1000). All remaining tuning parameters are set to the default values in the "*cforest*" R function in "partykit" (Hothorn, Seibold and Zeileis, 2023).

The GEOM database therefore reports three different estimates of IOp for each country-year: an ex-ante tree (Ctree) measure; an ex-post tree (Trafotree) measure; and an ex-ante forest measure.[22] Although these indices are quite strongly correlated (as we show in Section 5), they provide complementary information. Readers who are partial to the ex-post Principle of Compensation will naturally prefer our ex-post tree estimates. Among the ex-ante results, we recommend focusing on the random forest estimates, which are the most robust, and using the tree-based results as sources of complementary information on the structure of opportunity in each sample. The online database contains full pictures of both ex-ante and ex-post trees, analogous to Figure 1 above and to Figure B1 in Appendix B, for each country-year in the sample.

The Role of Individual Circumstances

In addition to the summary measures of IOp and the corresponding trees described above; the database also contains results for two different ways of assessing the relative predictive importance of individual circumstances in contributing to the total IOp estimate. The first way, which relies on Shapley-Shorrocks decompositions, provides an estimate of the *average* relative importance of each circumstance (and is therefore sensitive to its prevalence in the population), whereas the second relies on Partial Dependency Plots (PDPs) and provides an estimate of the *marginal* importance of each circumstance (to the person that has it, regardless of population prevalence). Neither estimate can be interpreted causally, of course: As with estimates of intergenerational mobility, omitted variables prevent any such interpretation. Nonetheless, quantifying the differences in the contributions of circumstances remains descriptively useful.

The Shapley value method calculates each variable's contribution to predicting variation in the outcome by assessing the average decline in explained outcome variability when the variable is excluded. The procedure involves drawing sub-samples, estimating IOp using a deep Ctree/Trafotree, and then re-estimating IOp after systematically removing circumstances by replacing their values with a vector ***1***. These drops are assessed by considering the case in which only the variable of interest is neutralized, as well as all cases where each possible combination of variables including the variable of interest, is neutralized. A weighted average of these drops provides the Shapley value (Shapley, 1952; Shorrocks, 2013). To account for sample dependency, this process is repeated 100 times, and the results are averaged across iterations.

Because the relative *average* importance of a control variable depends on its prevalence in the population (e.g., the relative importance of immigration background is inherently limited if there are few immigrants in the sample), we complement the analysis by plotting Partial Dependence Plots (PDPs) for each circumstance.[23] PDPs, originally introduced by Friedman

[22] We are not aware of suitable methods to produce forest analogues for Trafotrees. See Appendix B3 and Brunori, Ferreira and Salas-Rojo (2024) for a discussion.

[23] This is only done for ex-ante random forest IOp measures, which are our preferred estimates.

(2001), are visual tools designed to aid in interpreting machine learning outputs. They show how changes in a specific predictor variable affect the predicted outcome while holding all other variables constant. For instance, the partial dependence function for a particular feature, say "having a mother in a high-skilled occupation," represents the average prediction if we were to force all data points to assume that feature value: What would the average outcome be if all students had a mother in a high-skilled occupation?

This counterfactual exercise is implemented while keeping the distribution of all other features constant. It offers a valuable complement to Shapley values, as it focuses on the marginal importance of each characteristic, independent of their marginal distribution. Details of estimation algorithms for both Shapley values and PDPs may be found in Appendix B2.

5. Overview of the GEOM database

We can now provide an overview of the statistics and descriptive tools contained in the GEOM database. Space limitations prevent us from presenting all the evidence available online, and we invite the reader to visit and browse the site for themselves. Here we first present our six summary measures of IOp – absolute and relative estimates from ex-ante trees, random forests and ex-post trees – and explore how they co-vary. Then we focus more narrowly on our preferred estimates, namely those from ex-ante random forests, and showcase different ways to visualize both levels and trends across countries. Next, we investigate some empirical regularities in the relationships between IOp, on the one hand, and overall inequality and per capita GDP on the other. Finally, we describe some comparative results for the relative importance of individual characteristics.

5.1. Summary measures of IOp

Table 2 presents both absolute ($\mathrm{IO_A}$) and relative ($\mathrm{IO_R}$) estimates of inequality of opportunity for the latest available year for each of the 72 countries in GEOM, based on the Gini coefficient. Table A2 in Appendix A replicates this table using the mean logarithmic deviation (MLD). Column 3 indicates whether the estimates are based on equivalized household income or consumption, and column 4 presents total inequality in that variable. Columns 5-7 contain the absolute estimates, while columns 8-10 display the relative indices (in percentage terms).

Table 2: GEOM main results from latest wave (Gini)

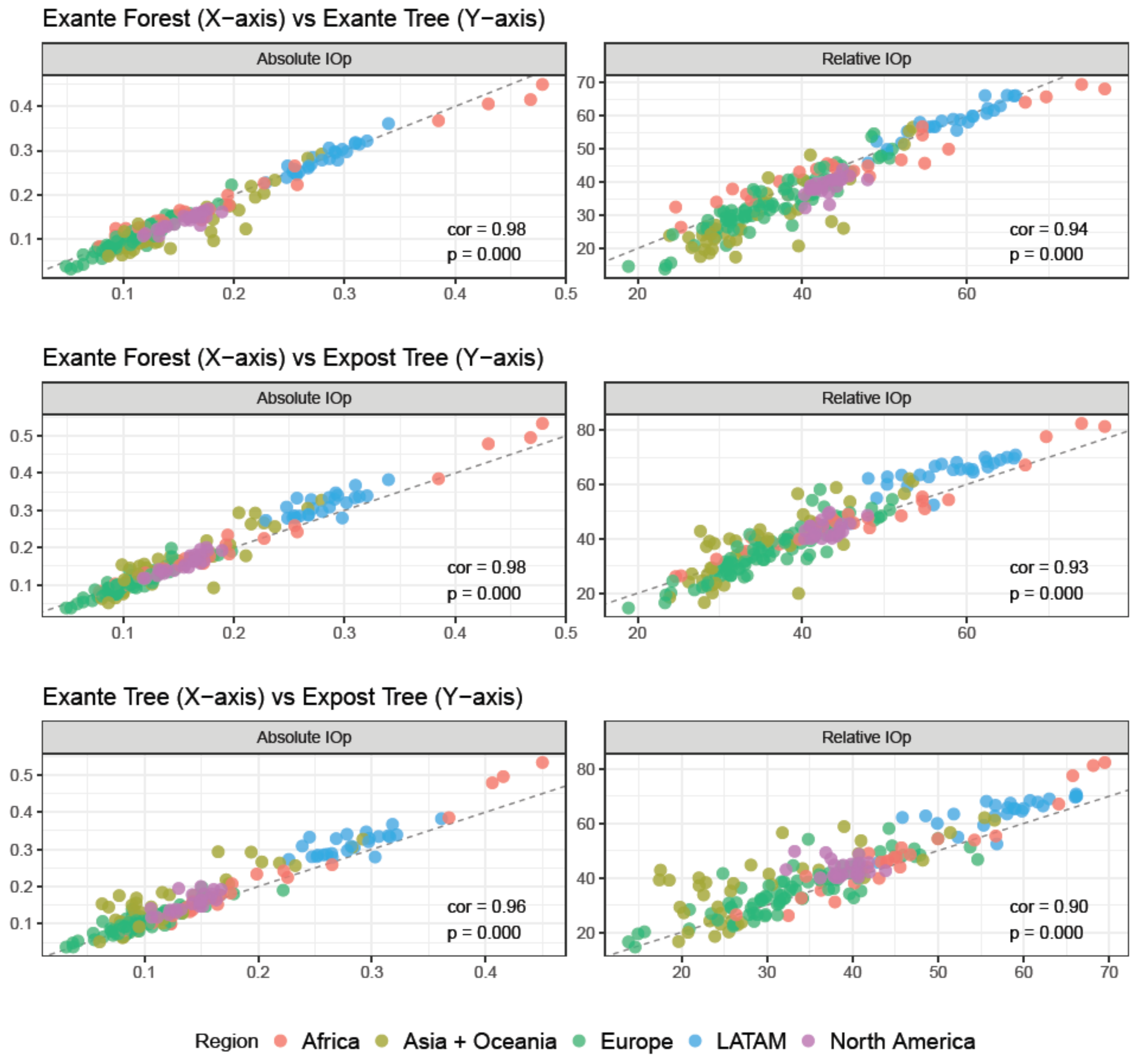
Country	Year	Variable	Sample Gini	Tree (ex-ante)	Random Forest (ex-ante)	Tree (ex-post)	Tree (ex-ante) Relative (%)	Random Forest (ex-ante) Relative (%)	Tree (ex-post) Relative (%)
Argentina	2014	Income	0.388	0.167	0.179	0.177	42.997	45.984	45.649
Armenia	2016	Income	0.412	0.116	0.179	0.184	28.114	43.554	44.720
Australia	2019	Income	0.355	0.094	0.101	0.081	26.584	28.358	22.782
Austria	2019	Income	0.280	0.096	0.099	0.091	34.130	35.235	32.382
Belgium	2019	Income	0.243	0.097	0.104	0.107	39.786	42.875	44.110
Benin	2018	Consumption	0.348	0.140	0.130	0.133	40.195	37.209	38.042
Bolivia	2008	Income	0.500	0.278	0.294	0.341	55.631	58.772	68.114
Brazil	2014	Income	0.488	0.322	0.320	0.339	66.010	65.703	69.600
Bulgaria	2019	Income	0.407	0.222	0.198	0.190	54.613	48.708	46.740
Burkina Faso	2018	Consumption	0.379	0.123	0.093	0.099	32.489	24.631	26.134
Chile	2015	Income	0.492	0.239	0.248	0.309	48.537	50.346	62.764
China	2018	Income	0.497	0.194	0.219	0.293	38.998	44.127	58.829
Colombia	2010	Income	0.535	0.245	0.257	0.333	45.815	48.019	62.182
Croatia	2011	Income	0.306	0.097	0.107	0.118	31.698	35.100	38.600
Cyprus	2019	Income	0.315	0.157	0.160	0.171	49.984	50.937	54.465
Czech Rep.	2019	Income	0.239	0.075	0.079	0.073	31.425	33.055	30.547
Denmark	2019	Income	0.268	0.056	0.072	0.057	20.902	26.900	21.237
Ecuador	2014	Income	0.455	0.227	0.229	0.273	49.890	50.308	60.000
Estonia	2019	Income	0.280	0.072	0.083	0.081	25.820	29.636	29.030
Finland	2019	Income	0.287	0.092	0.109	0.094	32.089	38.158	32.612
France	2019	Income	0.287	0.111	0.123	0.119	38.883	42.897	41.361
Gambia	2015	Income	0.576	0.199	0.195	0.234	34.544	33.797	40.573
Georgia	2016	Income	0.469	0.122	0.211	0.178	26.062	45.016	37.951
Germany	2019	Income	0.279	0.080	0.087	0.074	28.823	31.335	26.490

Country	Year	Variable	Sample Gini	Tree (ex-ante)	Random Forest (ex-ante)	Tree (ex-post)	Tree (ex-ante) Relative (%)	Random Forest (ex-ante) Relative (%)	Tree (ex-post) Relative (%)
Ghana	2017	Consumption	0.420	0.152	0.161	0.172	36.251	38.249	40.794
Greece	2019	Income	0.306	0.110	0.117	0.126	36.078	38.268	41.209
Guatemala	2011	Income	0.526	0.298	0.291	0.330	56.629	55.374	62.735
Guinea Bissau	2018	Consumption	0.312	0.142	0.134	0.137	45.548	43.017	43.946
Hungary	2019	Income	0.275	0.071	0.089	0.080	25.744	32.353	29.085
Iceland	2005	Income	0.263	0.055	0.081	0.075	20.935	30.775	28.343
India	2012	Income	0.527	0.292	0.279	0.327	55.416	53.007	62.094
Indonesia	2014	Consumption	0.428	0.116	0.126	0.101	27.043	29.472	23.681
Ireland	2019	Income	0.281	0.105	0.125	0.117	37.269	44.275	41.714
Italy	2019	Income	0.315	0.117	0.108	0.125	37.032	34.242	39.727
Ivory Coast	2018	Consumption	0.325	0.123	0.103	0.101	37.927	31.529	31.129
Kazakhstan	2016	Income	0.339	0.081	0.081	0.063	23.900	23.900	18.493
Kyrgyzstan	2016	Income	0.448	0.078	0.143	0.176	17.408	31.911	39.240
Latvia	2019	Income	0.337	0.086	0.107	0.104	25.541	31.652	30.703
Lithuania	2019	Income	0.341	0.107	0.104	0.103	31.204	30.472	30.179
Luxembourg	2019	Income	0.322	0.132	0.142	0.134	40.981	43.993	41.726
Malawi	2020	Consumption	0.357	0.161	0.156	0.170	45.169	43.629	47.578
Mali	2019	Consumption	0.344	0.125	0.114	0.122	36.303	33.217	35.429
Malta	2019	Income	0.266	0.083	0.091	0.091	31.149	34.124	34.388
Mexico	2017	Income	0.532	0.303	0.298	0.279	56.853	55.952	52.459
Mongolia	2016	Income	0.471	0.144	0.181	0.176	30.563	38.533	37.492
Nepal	2011	Income	0.538	0.219	0.216	0.263	40.692	40.115	48.931
Netherlands	2019	Income	0.255	0.063	0.086	0.106	24.834	33.608	41.598
Niger	2018	Consumption	0.311	0.082	0.079	0.082	26.397	25.273	26.429

Country	Year	Variable	Sample Gini	Tree (ex-ante)	Random Forest (ex-ante)	Tree (ex-post)	Tree (ex-ante) Relative (%)	Random Forest (ex-ante) Relative (%)	Tree (ex-post) Relative (%)
Nigeria	2019	Consumption	0.288	0.127	0.120	0.132	43.908	41.722	45.748
Norway	2019	Income	0.273	0.099	0.112	0.107	36.177	41.011	39.290
Panama	2003	Income	0.527	0.306	0.286	0.335	58.046	54.288	63.472
Peru	2015	Income	0.423	0.260	0.267	0.287	61.633	63.266	67.811
Poland	2019	Income	0.282	0.084	0.088	0.093	29.950	31.299	32.825
Portugal	2019	Income	0.306	0.118	0.135	0.135	38.705	44.034	44.263
Romania	2019	Income	0.341	0.151	0.144	0.198	44.245	42.161	58.133
Senegal	2018	Consumption	0.314	0.107	0.093	0.102	34.001	29.580	32.538
Sierra Leone	2018	Consumption	0.311	0.137	0.137	0.144	44.108	44.076	46.233
Slovakia	2019	Income	0.232	0.070	0.075	0.086	30.207	32.100	37.091
Slovenia	2019	Income	0.249	0.082	0.087	0.092	32.892	34.739	36.747
South Africa	2017	Income	0.610	0.415	0.468	0.496	68.121	76.746	81.256
South Korea	2019	Income	0.351	0.106	0.121	0.145	30.217	34.578	41.363
Spain	2019	Income	0.329	0.151	0.145	0.159	45.897	44.164	48.328
Sweden	2019	Income	0.276	0.104	0.094	0.099	37.786	34.047	36.080
Switzerland	2019	Income	0.283	0.084	0.093	0.074	29.767	32.945	25.953
Tajikistan	2016	Income	0.309	0.061	0.087	0.051	19.657	28.109	16.613
Tanzania	2013	Consumption	0.373	0.162	0.173	0.171	43.313	46.261	45.805
Timor Leste	2014	Consumption	0.282	0.117	0.101	0.113	41.341	35.877	40.099
Togo	2018	Consumption	0.382	0.164	0.151	0.152	43.093	39.659	39.843
Uganda	2014	Consumption	0.371	0.167	0.178	0.177	44.881	47.980	47.656
United Kingdom	2011	Income	0.324	0.076	0.096	0.087	23.479	29.503	26.846
United States of America	2014	Income	0.395	0.150	0.164	0.165	38.045	41.565	41.667
Uzbekistan	2016	Income	0.460	0.095	0.182	0.092	20.753	39.548	19.948

Figure 2 shows the correlations across the summary indices reported in the columns of Table 2, although now using all years (the pooled cross-section), rather than just the latest year. The panels on the left show the association between absolute IOp estimates obtained with ex-ante trees and forests (upper panel), ex-ante forest and ex-post trees (middle panel), and ex-ante and ex-post trees (bottom panel). The panels on the right show the association obtained with the same methods, now for relative IOp. At the bottom-right of each plot we report the correlation coefficients and the associated p-values.

Figure 2 – Correlations across IOp estimates



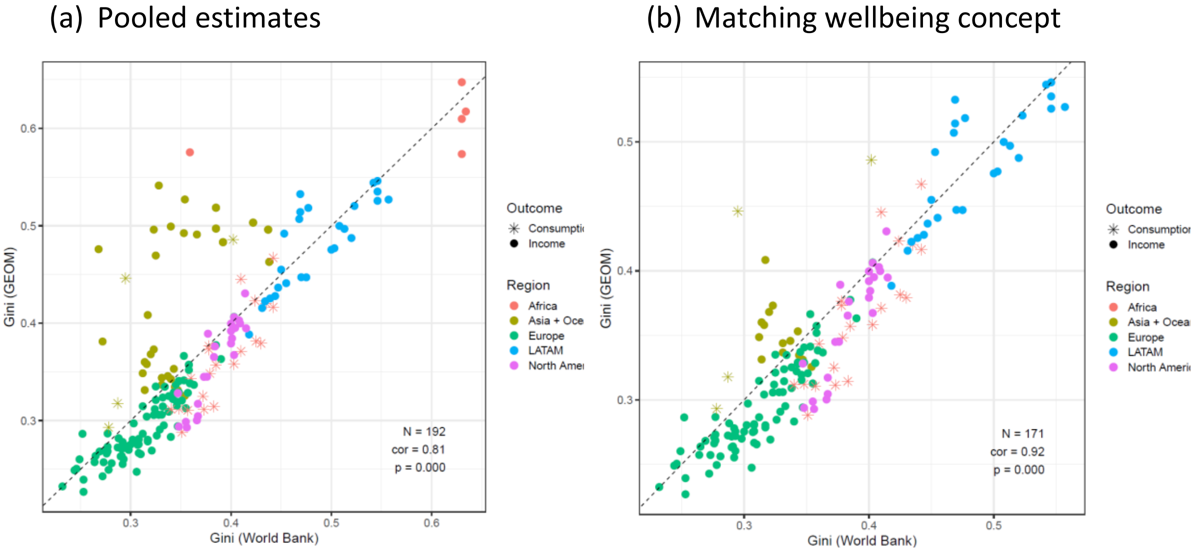
Note: Pooled cross section data. Elaboration based on GEOM data. The correlation coefficient (c) and the associated p-value (p) are shown at the bottom-right.

The literature (e.g., Fleurbaey and Peragine, 2013) has shown that the concepts of ex-ante and ex-post IOp represent distinct - and often incompatible - approaches to study unfairness. Yet, our estimates of ex-ante and ex-post IOp are strongly correlated. Since absolute IOp estimates are mechanically correlated with total inequality, it is less surprising to observe a stronger correlation in the left panels. However, the high association between relative ex-ante and relative ex-post is particularly noteworthy. It suggests that, despite their theoretical differences, the two approaches yield highly consistent descriptions of inherited inequalities.

Before turning to different ways in which these results can be viewed, it is useful to ask whether our sample definitions and treatment of the income and consumption data lead to large differences in the levels and ranks of income and consumption inequality relative to other publicly available compilations. There are two main sources of potential differences, namely: (i) the use of a $\sqrt{n}$ equivalence scale to adjust all household incomes; and (ii) changes in the composition of the sample due to the exclusion of observations with missing information.[24] Figure 3 plots our estimates against the corresponding figures published by the World Bank (World Development Indicators) – in Panel (a) for all countries and in Panel (b) only for those based on the same underlying concept of money-metric wellbeing – income or consumption.

The main deviations from the 45-degree line in Panel (a) are mostly Asian countries - Armenia, China, Georgia, Indonesia, India, Kazakhstan, Kyrgyzstan, Mongolia, Nepal, Uzbekistan – and are due to the use of income as an outcome for those countries (in the GEOM database), as opposed to consumption (in the World Bank database). In Panel (b), differences should be mostly due to the equivalence scale and within-country sample composition. While there are clear differences – and, as expected, inequality in equivalized incomes is lower than in per capita incomes (see Coulter et al.,1992) – it is reassuring that the total inequality estimates in GEOM are closely correlated with those reported by the World Bank ($\rho = 0.92$, p-value=0.00).

Figure 3 - Total inequality estimates: GEOM vs World Bank


(a) Pooled estimates

(b) Matching wellbeing concept

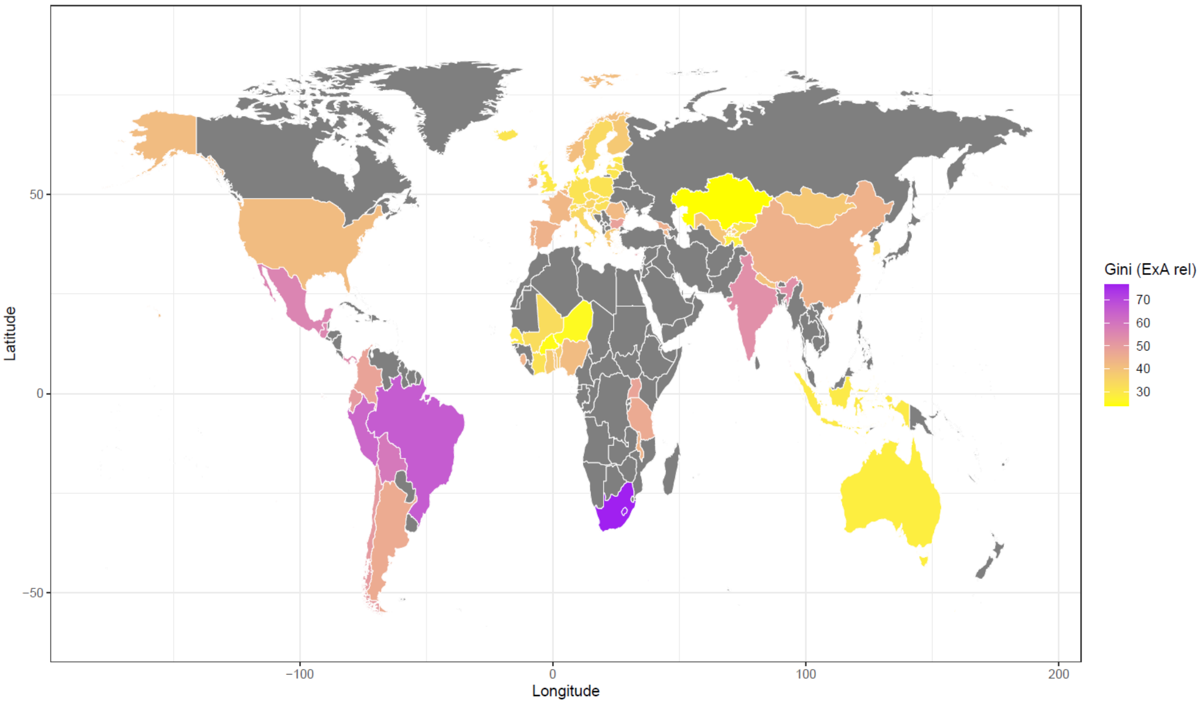
Note: Pooled cross-section data. Elaboration based on GEOM and World Bank data. The distinction between Income and Consumption as outcome refers to GEOM only.

[24] The summary statistics reported here are before the age adjustment; see Section 3.

5.2. Visualizing levels and trends of (ex-ante) IOp around the world

While Table 2 contains all our (Gini-based) summary IOp measures for the latest year available for each country, the database also allows for more intuitive ways to view results. Focusing on our preferred, random forest ex-ante Gini coefficient estimates, Figure 4 illustrates the kind of map available to GEOM website users. It reports estimates of relative IOp for the latest survey available, with the legend using a color scale where lower values are colored in yellow and higher values are colored in purple. Similar maps for total inequality or other definitions of IOp can be easily accessed via the interface in the website (https://geom.ecineq.org/world-view/), which allows the user to change the inequality measure (Gini vs MLD), the IOp perspective (ex-ante vs. ex-post), the approach (absolute vs relative) and the dependent variable (income, consumption, or both).

Figure 4 - Map of the Relative Ex Ante IOp estimates from GEOM (using Gini coefficients).

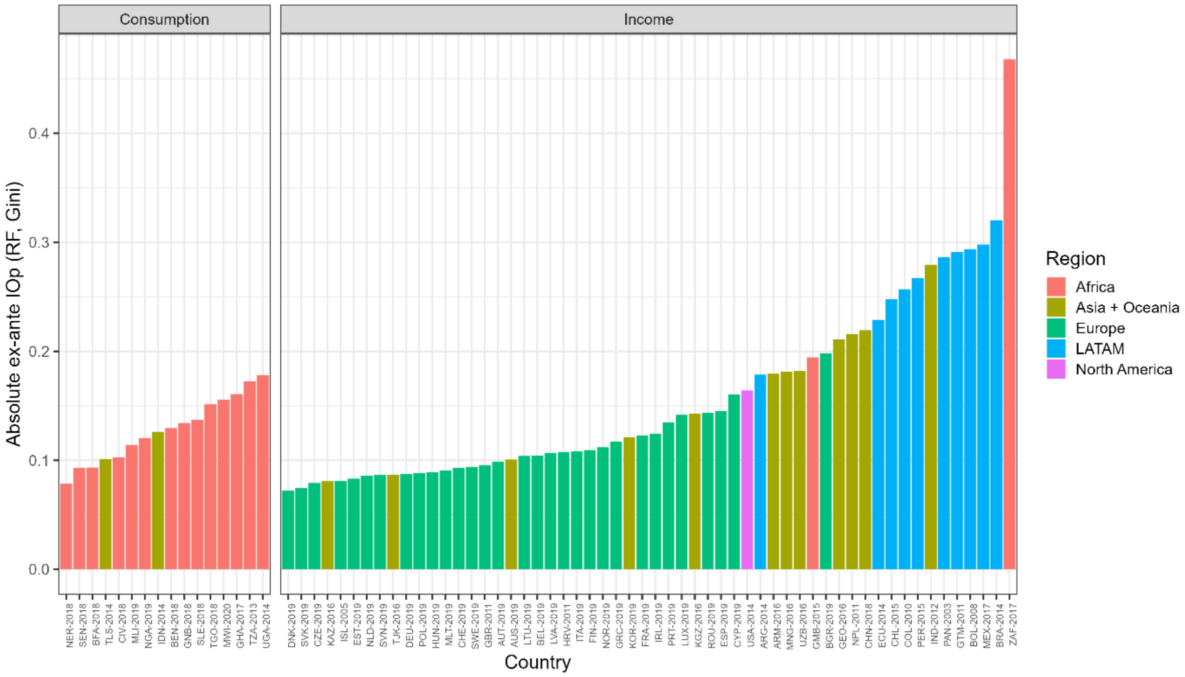


Note: latest available estimate for each country. Source: GEOM (https://geom.ecineq.org/world-view/).

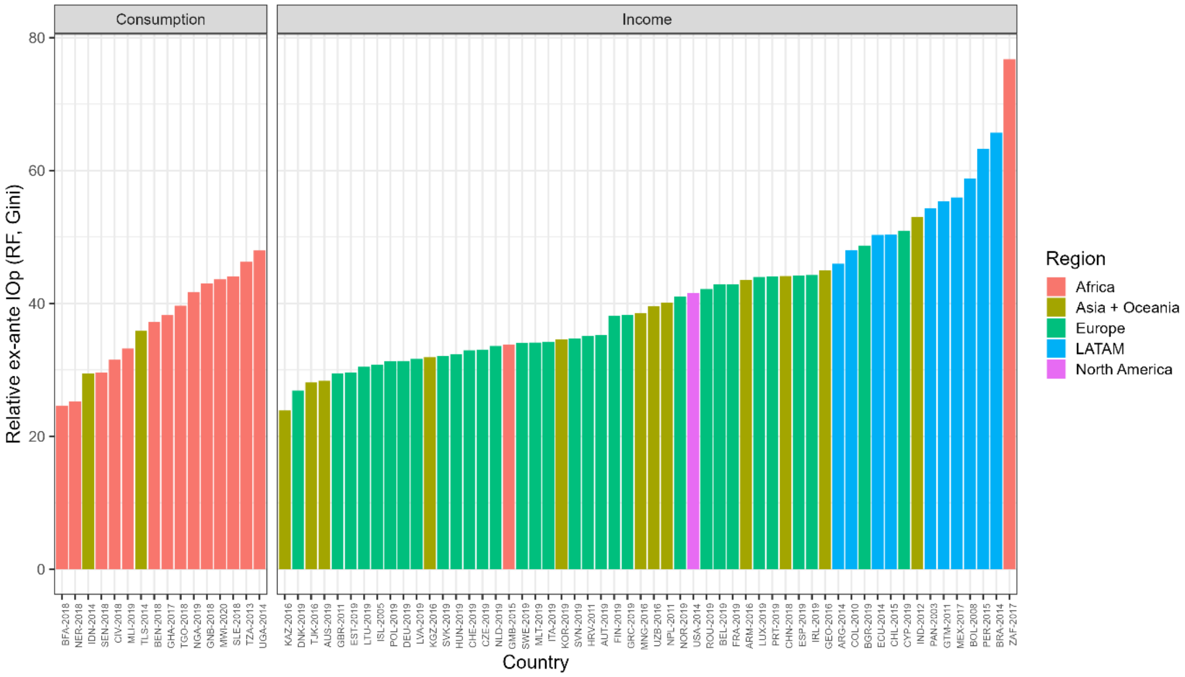
The relative measures in Table 2 (column 9) and Figure 4 range from 23.9% (Kazakhstan, 2016) to 76.8% (South Africa, 2017). There is considerable variation across regions, with Latin America substantially overrepresented at the top of the range, while Europe is overrepresented at the bottom. There is much more dispersion across Asia, with some countries like Kazakhstan and Indonesia at the bottom, and others like India or China reaching much higher values. It is difficult to compare Africa to the rest of the world since, as noted earlier, most African countries report consumption rather than income-based IOp.

We emphasize this comparability limitation more clearly in the bar chart in Figure 5, where consumption-based estimates are shown separately to the left of the figure. These are all still ex-ante Gini coefficients from random forests for the latest available surveys, with absolute levels in the top panel and relative indices in the bottom.[25] Both panels highlight the remarkable variation in IOp across the globe. This variability is, of course, higher when we look at absolute IOp, which is not normalized by the total inequality in the country. Absolute and relative IOp are, unsurprisingly, highly correlated ($\rho = 0.9$) but comparing them provides some interesting insights. For example, by looking at absolute IOp in Denmark one may conclude that, with a Gini coefficient of 0.07, IOp is hardly a problem for this country. While this might be true when we compare it with other absolute values around the world, the bottom panel in Figure 5 shows that IOp in Denmark still accounts for more than a quarter (26.9%) of its (rather low level of) total inequality. On the other hand, some countries, like Australia, move from the middle to the bottom part of the IOp distribution when passing from absolute to relative measures.

Figure 5 – Ex ante IOp Estimates from GEOM



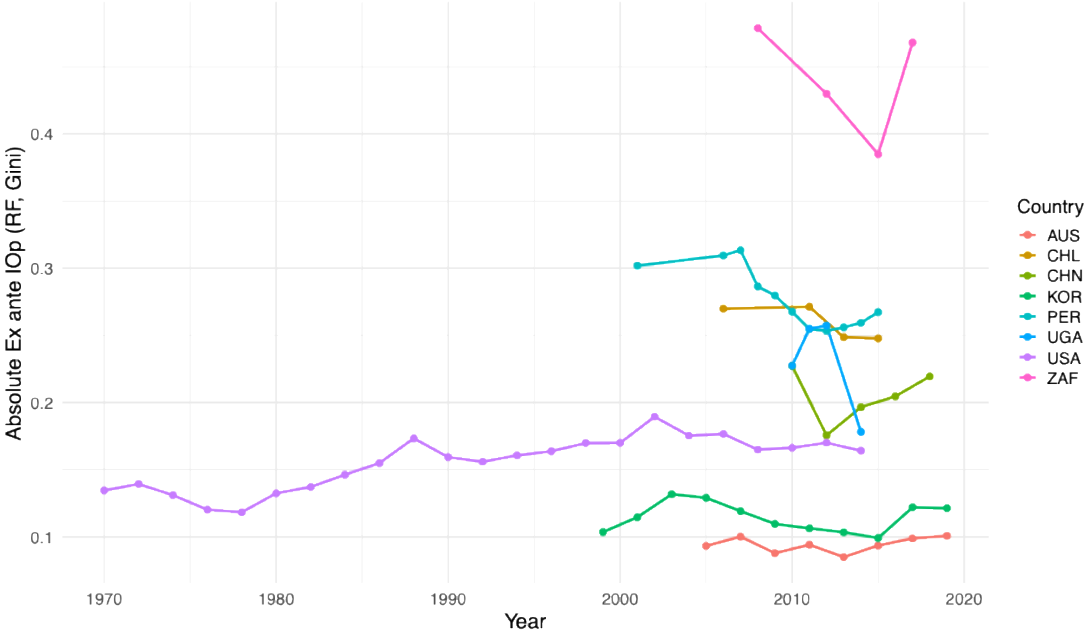
[25] Naturally, the bottom panel of Figure 5 is an alternative graphical representation of the information contained in Figure 4.

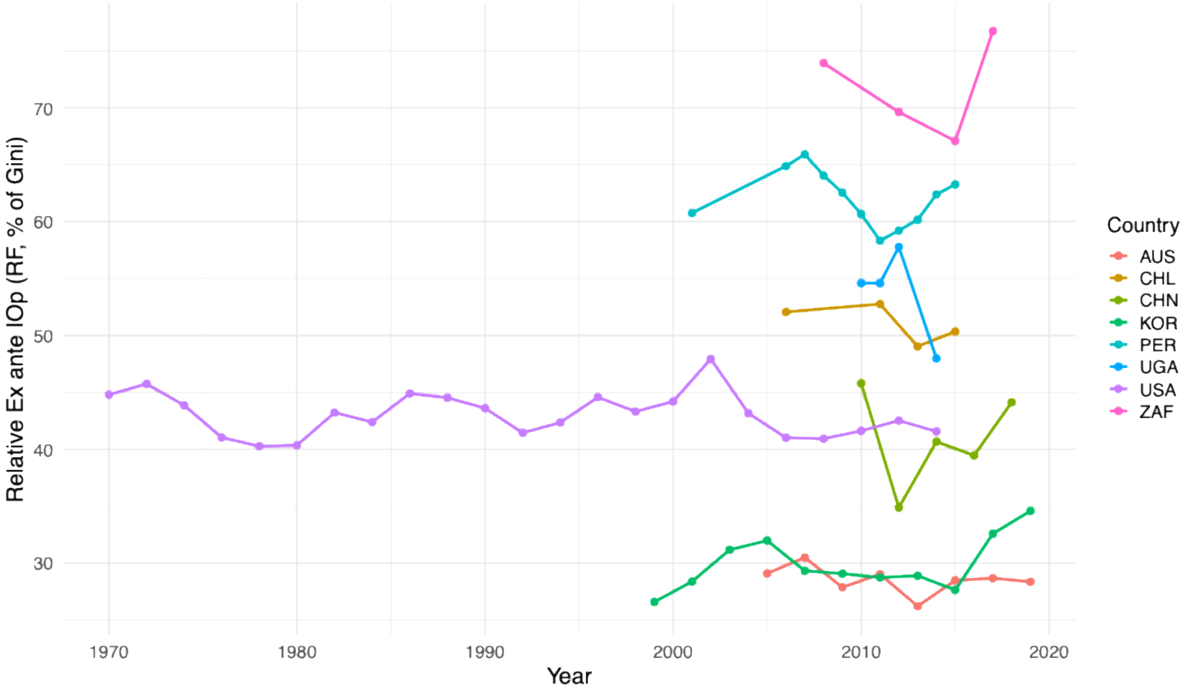


Note: elaboration based on GEOM data, latest available estimate for each country.

The GEOM database also allows users to follow the evolution of IOp over time for certain countries although, at present, time series are typically quite short and only available for a few countries. Figure 6 below reports IOp estimates for all countries that are observed for at least four years: Australia, Chile, China, South Korea, Peru, Uganda, the United States and South Africa. Once again, absolute (relative) estimates are on the top (bottom) panel.

Figure 6 - Trends in IOp



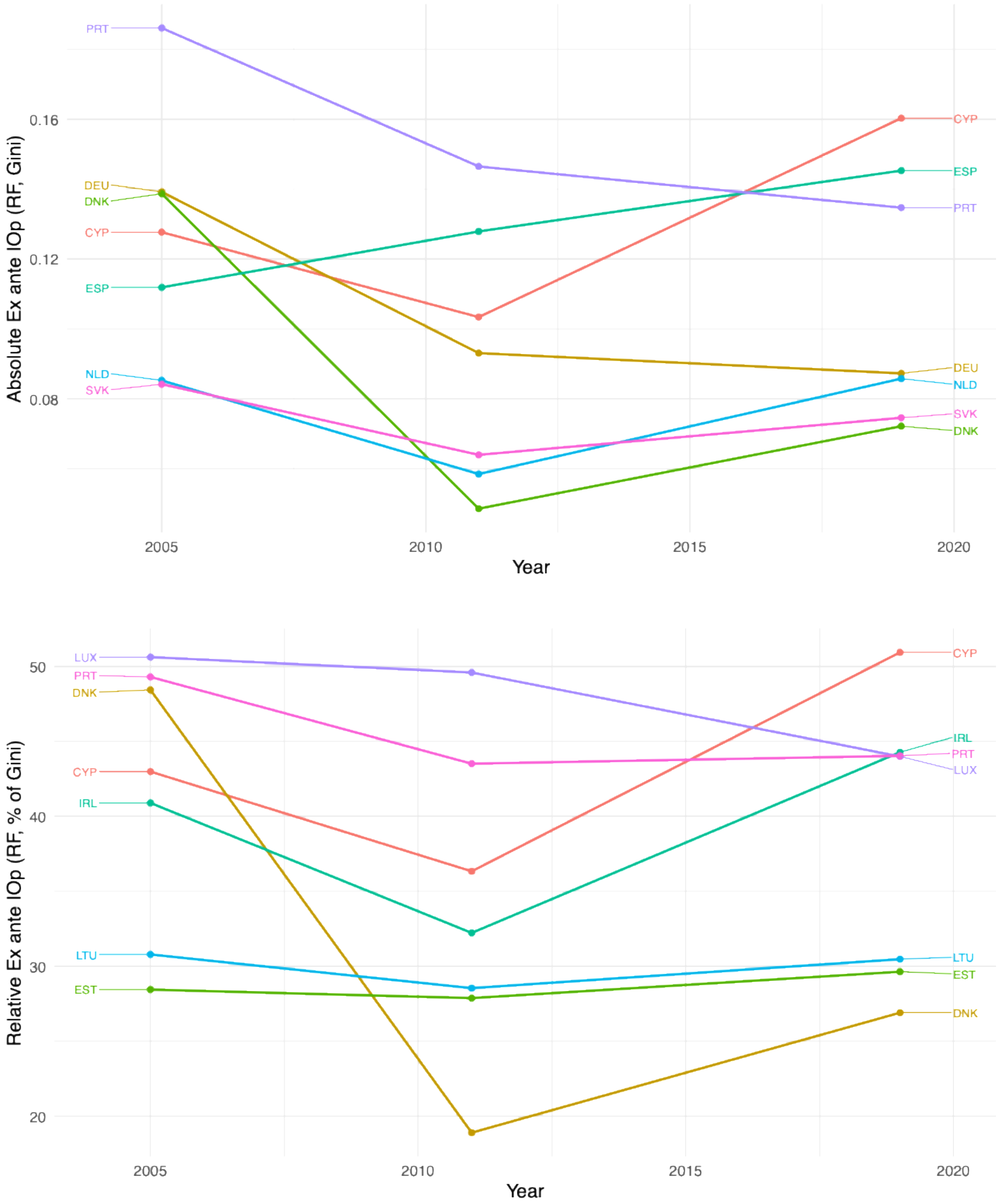


Note: Elaboration based on GEOM. The graph reports only countries that are observed for at least four years.

There is considerable variation in trends over time, with a steady increase in absolute IOp in the United States from 1978 to 2002, then followed by a slight decline. Throughout the period of observation, IOp in the United States accounts for between 40 and 50% of total inequality, placing it among the highest values for high-income economies. Focusing on the last 10 years of the available data, Figure 6 shows that the United States are no longer the "land of opportunity", in line with the conclusions in Chetty et al. (2014). Conversely, countries like Australia and South Korea display much lower levels of both absolute and relative IOp.[26] Both China (the world's largest country) and South Africa (the world's most unequal) display U-shaped patterns for both absolute and relative IOp. In China, absolute IOp fell sharply between 2010 and 2012 but then rose steadily back to its original level by 2018. In South Africa, the decline lasted between 2008 and 2015, offset by a sharp rise in 2017.

[26] Our results on South Korea are also in line with Moramarco et. al (2020).

Figure 7 - Trends in IOp: Europe



Note: Elaboration based on GEOM. The graph reports a subset of European countries.

Figure 7 zooms in on the evolution of IOp in Europe. As explained in Section 3, for most countries in this region GEOM draws information for only three years: 2005, 2011 and 2019. Figure 7 shows only those countries with the two lowest or highest estimates in 2005 and 2019, which is sufficient to illustrate the variability in both levels and trends of IOp within the European region. It is interesting that only three of the countries selected in the top panel appear also in the bottom one, highlighting again the importance of considering both absolute and relative measures of IOp when making these assessments.

Some changes were substantial. Denmark, for example, experienced a large drop in absolute IOp from 2005 to 2011 followed by a slight increase, transitioning from having one of the highest relative IOp levels in 2005 to the lowest in 2019. Other countries like the Netherlands or Slovakia (top panel) or Estonia and Lithuania (bottom panel) display more stable trends. Overall, of the seven countries in the top panel, four experienced declines in absolute ex-ante IOp between 2005 and 2019, two experienced increases (Cyprus and Spain), and one (the Netherlands) was basically unchanged. In the bottom panel, the same number of countries (three) saw rises as declines in relative ex-ante IOp, with Lithuania basically unchanged.

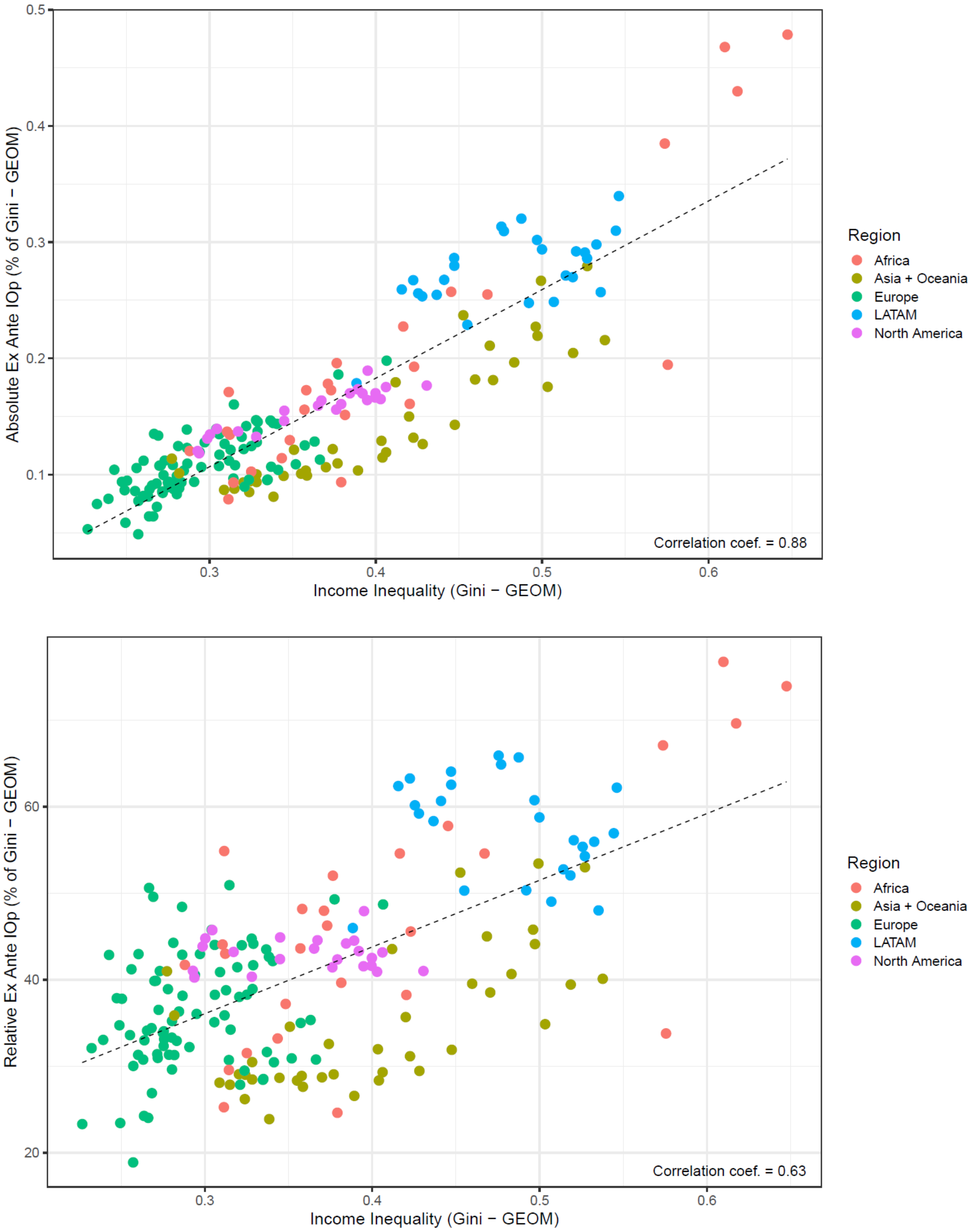
5.3. Empirical regularities: Opportunity "Great Gatsby" and Kuznets curves[27]

It might be of interest to ask how our measures of IOp co-vary with other important variables, such as overall income inequality and GDP per capita. Given the close conceptual and empirical association between IOp and (inverse) intergenerational mobility, one might expect to find a positive association between IOp and total inequality, analogous to the Great Gatsby curve first plotted by Corak (2013). Indeed, such a relationship for IOp was first reported by Brunori et al. (2013). After observing how intergenerational income persistence was closely linked to IOp in their sample of countries, those authors also reported a clear positive empirical association between IOp and income inequality across countries.[28] Our (more internally comparable) GEOM estimates confirm this positive association between IOp and inequality. Figure 8 displays pooled cross-sectional scatter plots, showing both absolute and relative (random forest) IOp estimates against total inequality.

[27] In this subsection, scatter plots draw on the full pooled cross-sectional data from GEOM, so that each point corresponds to a country-year observation.

[28] As mentioned in the Introduction, intergenerational income persistence can be interpreted as IOp under the simplifying assumption that parental income is the only circumstance beyond individual control.

Figure 8 - The Great Gatsby curve in GEOM



Note: Pooled cross section data. Elaboration based on GEOM data.

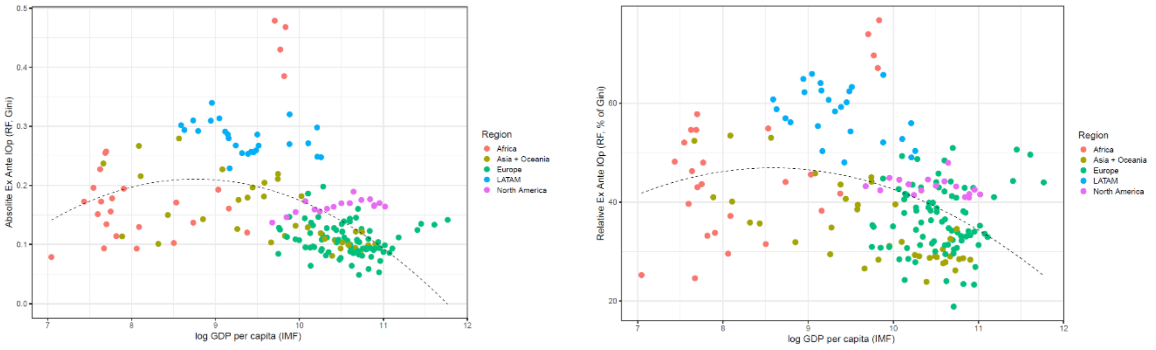
Of course, the positive association on the top panel of Figure 8 may be seen as at least partly mechanical, since absolute ex-ante IOp corresponds to the between-type component of total income inequality. However, there is no mechanical reason for this positive correlation to appear when looking at relative IOp (bottom panel). As Brunori et al. (2013) suggested, this relationship likely arises from the circular causation between the two concepts: today's income differences among families translate into differences in opportunities for their children, and those opportunity gaps in turn shape income differences in adulthood for that generation.

It is interesting to notice that in Figure 8 (bottom panel) Latin American countries (blue dots) tend to cluster above the fitted line, confirming the high weight of IOp over total inequality in those countries. Conversely, Asian countries (light green dots) fall consistently below the fitted line. The pink dots in the above figures all refer to the United States. So, inequality of opportunity is positively associated with current inequality, and the robust methodology used to construct the GEOM database makes it a unique source of data to verify this empirical regularity.

How might IOp vary with a country's development, or at least with development as proxied by per capita GDP? This question is of course reminiscent of the Kuznets hypotheses, which posits that inequality first tends to increase and then decrease as a country develops, leading to an inverted-U relationship between inequality and development. If the Kuznets hypothesis holds, then the positive association between IOp and overall income inequality documented above may be sufficient for an "Opportunity Kuznets Curve" to arise. Ferreira et al. (2025) argue that this dynamic is not only mechanical but can be explained by the interaction between economic development – which reduces the cost of accessing more productive technologies – and intergenerational transmission of incomes – which makes it less costly for the children of wealthy parents to invest in new productive sectors.

The GEOM database is, once again, a unique source of data for investigating the existence of an Opportunity Kuznets Curve. Figure 9 plots our preferred estimates of ex-ante IOp against the logarithm of GDP per capita, as a proxy for the level of development of each country. We draw information on GDP per capita from the IMF database (World Economic Outlook, October 2024) and express values in US dollars PPP2017 to align them with those in GEOM. The fitted quadratic regression seems to confirm the hypothesis in Ferreira et al. (2025), with an R^2 of 0.261 in the left panel, and 0.144 in the right one.[29]

Figure 9 - The Opportunity Kuznets curve



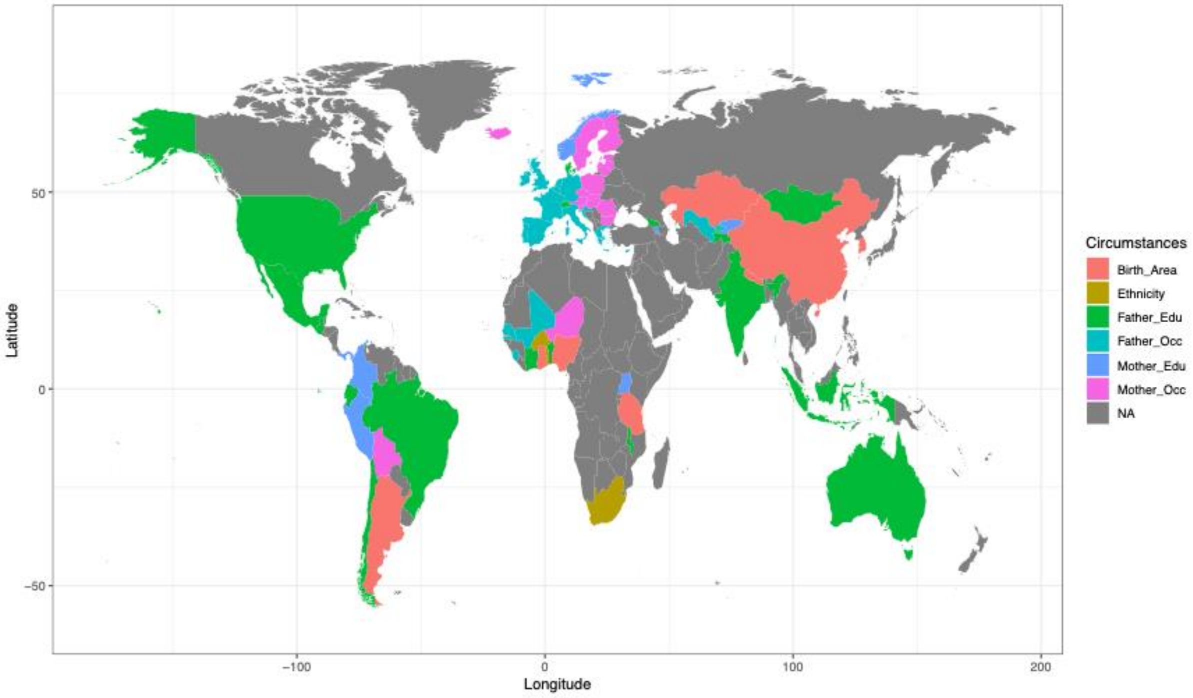
Note: Pooled cross section data. Elaboration based on GEOM and IMF data. GDP per capita expressed in US dollars PPP 2017.

[29] As a benchmark, the R^2 for the fitted regression in Figure 8 (right panel) is 0.386.

5.4. *Important individual circumstances*

As discussed in Section 4, the methods we employ to estimate IOp also allow us to conduct a Shapley-Shorrocks decomposition to estimate the contribution of specific circumstances to IOp. In Figure 10 we plot, for each country, the circumstance with the highest contribution to ex-ante IOp. The interested reader can access the complete Shapley value decomposition for each country on the GEOM website (https://geom.ecineq.org/country-profile/, looking at ex-ante and ex-post "Decomposition" format).

Figure 10 - Most important circumstance contributing to IOp



Note: Elaboration based on GEOM. Ex ante IOp, latest available estimates for each country.

Descriptively, parental occupation and educational levels appear to be the most important circumstances in most countries covered by GEOM. Father's education accounts for the largest share of IOp in the United States, India, Indonesia, Australia, Brazil and Chile, for example. Interestingly, Europe seems rather neatly divided between the West, where father's occupation dominates, and the East (including Sweden and Finland), whether the mother's occupation has the largest share. Unsurprisingly, ethnicity is the most important contributor to IOp in South Africa. In China, as well as in Argentina, Kazakhstan, Tanzania and some other African countries, place of birth is descriptively the most relevant circumstance.

6. Conclusions

Not all forms of inequality are the same, and there is growing evidence that inequality of opportunity – which is mostly inherited and reflects factors beyond individual control – is particularly harmful. This paper has introduced and described a novel public-access database containing internationally comparable estimates of IOp across seventy-two countries, which together account for just over two-thirds of the world's population. Estimates for more than one period are available for most countries and were constructed from the household-level microdata from 196 different household surveys, following definitions and protocols designed for maximum comparability.

The dataset contains a range of alternative measures, both absolute and relative, corresponding to the two main measurement approaches in the literature (ex-ante and ex-post) and using two different inequality indices (the Gini coefficient and the mean logarithmic deviation). Although these different indices are highly correlated among themselves, this summary paper focused on our preferred estimates, the ex-ante opportunity Gini coefficients computed from random forests.

There is considerable variation in this measure across countries, with absolute levels ranging from 0.05 (for Denmark in 2011), to 0.48 (for South Africa in 2008) and relative measures from 19% of total inequality also in Denmark, 2011, to 77% in South Africa, 2017. Across the entire sample, IOp is positively correlated with overall income or consumption inequality, confirming earlier findings of an Opportunity Great Gatsby curve. With respect to per capita GDP, IOp first rises and then falls, describing an inverted-U curve rather like that postulated by the Kuznets hypothesis for income inequality. There are sharp regional differences in IOp, with European (and some Asian) countries displaying lower levels, while Latin American (and some African) countries are at the upper range. Among high-income countries, the United States stands out for its high relative and absolute levels of inequality of opportunity.

Our use of data-driven machine learning tools to compute these indices was motivated by their superior properties in trading off upward (overfitting) and downward (omitted variable) estimation biases in the measurement of IOp. But the tree-based algorithms also generate interesting visualizations of the structure of inequality in different countries and allow users to compare income and population shares across types in informative ways. The approach also allows us to identify the most descriptively salient circumstances in each sample. Although family background variables such as parental education and occupation dominate in most cases, ethnic and geographical origins also play important roles in many countries.

Space constraints limit the results we can summarize in this paper, but readers are encouraged to visit the online database (www.geom.ecineq.org) to download data and conduct their own analysis. Future research to expand and update this database would greatly benefit from the inclusion of family background variables, such as place of birth, parental education and occupation, in more household surveys by statistical agencies around the world and, in particular, in those countries for which that information is currently not available for recent years.

References

Aaberge, R. , Mogstad, M., Peragine, V.: Measuring long-term inequality of opportunity. *Journal of Public Economics, 95(3-4), 193-204, (2011).*

Adermon, A., Brandén, G., Nybom, M.: The relationship between intergenerational mobility and equality of opportunity. *Institute for Evaluation of Labour Market and Education Policy (IFAU)*, Uppsala, Working Paper No. 2025:2, (2025).

Almås, I., Cappelen, A. W., Sørensen, E. Ø., Tungodden, B.: Fairness and the development of inequality acceptance. *Science*, *328*(5982), 1176-1178, (2010).

Björklund, A. Jäntti, M.: Intergenerational mobility, intergenerational effects, sibling correlations, and equality of opportunity: A comparison of four approaches*. Research in Social Stratification and Mobility,* 70, (2020).

Bourguignon, F, Ferreira, F. H. G., Menéndez, M.: Inequality of opportunity in Brazil. *Review of income and Wealth,* 53(4), 585-618, (2007)

Breiman, L.: Random Forest. *Machine Learning*, pp. 503–515, (2001).

Brunori, P., Neidhöfer, G.: The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach. *Review of Income and Wealth*, 67 (4), 900–927, (2021).

Brunori, P., Hufe, P., Mahler, D.: The roots of inequality: Estimating inequality of opportunity from regression trees and forests. *Scandinavian Journal of Economics*, 125 (4), 900–932, (2023).

Brunori, P., Ferreira, F. H. G., Peragine, V.: Inequality of opportunity, income inequality, and economic mobility: Some international comparisons. *Getting development right: Structural transformation, inclusion, and sustainability in the post-crisis era*. New York: Palgrave Macmillan US, 2013. 85-115.

Brunori, P., Peragine, V., Serlenga, L.: Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare* 52, 635-661, (2019).

Brunori, P, Ferreira, F. H. G.; Salas-Rojo, P.: Inherited Inequality: A General Framework and a 'Beyond-Averages' Application to South Africa. *IZA Discussion Papers, No. 17203, Institute of Labor Economics (IZA)*, Bonn, (2024).

Cappelen, A. W., Sørensen, E. Ø., Tungodden, B.: Responsibility for what? Fairness and individual responsibility. *European Economic Review, 54(3), 429-441, (2010).*

Cappelen, A. W., Konow, J., Sørensen, E. Ø., Tungodden, B*.:* Just luck: An experimental study of risk-taking and fairness. *American Economic Review,* 103(4), 1398-1413, (2013).

Chakravarty, S. R., Eichhorn, W.: Measurement of income inequality: Observed versus true data. In *Models and measurement of welfare and inequality* (28-32). Berlin, Heidelberg: Springer Berlin Heidelberg, (1994).

Checchi, D., Peragine, V.: Inequality of opportunity in Italy*. The Journal of Economic Inequality*, 8(4), 429-450, (2010).

Chetty, R., Hendren, N., Kline, P., Saez, E., Turner, N.: Is the United States still a land of opportunity? Recent trends in intergenerational mobility. *American Economic Review*, 104(5), 141-147, (2014).

Chow, G.: Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28 (3), 591–605, (1960).

Corak, M.: Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives,* 27 (3): 79-102, (2013).

Coulter, F., Cowell, F. A., Jenkins, S. P: Equivalence scale relativities and the extent of inequality and poverty. *Economic Journal,* 102, 1067-1082. (1992).

Ferreira, F. H. G., Brunori, P.: Inherited inequality, meritocracy, and the purpose of economic growth. International Inequalities Institute Working Paper 147, LSE. (2024).

Ferreira, F. H. G., Gignoux, J.: The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of income and wealth*, 57(4), 622-657, (2011).

Ferreira, F. H. G., Peragine, V.: Individual responsibility and equality of opportunity*. In: The Oxford Handbook of Well-Being and Public Policy, (2016).*

Ferreira, F. H. G., Moramarco, D., Peragine, V.*:* Economic development and inequality of opportunity: Kuznets meets the Great Gatsby?. No. wp-2025-25. World Institute for Development Economic Research (UNU-WIDER), 2025.

Fleurbaey, M., Moramarco, D., Peragine, V.: Measuring inequality and welfare when some inequalities matter more than others. *ECARES Working Papers*, (*2024)*.

Fleurbaey, M., Peragine, V.: Ex ante versus ex post equality of opportunity. *Economica*, 80(317), 118-130, (2013).

Fleurbaey, M.: On fair compensation. *Theory and decision, 36(3), 277-307, (1994).*

Friedman, J. H.: Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232, (2001).

Hothorn, T.: trtf: Transformation Trees and Forests. *CRAN,* (2023).

Hothorn, T., Zeileis, A.: partykit: A modular toolkit for recursive partitioning in R. *The Journal of Machine Learning Research*, 16 (1), 3905–3909, (2015).

Hothorn, T., Zeileis, A.: Transformation Forests, (2017). https://arxiv.org/abs/ 1701.02110.

Hothorn, T. , Zeileis, A.: Predictive Distribution Modeling Using Transformation Forests. *Journal of Computational and Graphical Statistics*, American Statistical Association, 30(4), 1181–1196, (2021).

Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15 (3), 651–674, (2006).

Hothorn, T., Seibold, H. and Zeileis, A.: Package ‘partykit’: A Toolkit for Recursive Partitioning*. CRAN,* (2023).

Hsieh, C. T., Hurst, E., Jones, C. I., Klenow, P. J. The allocation of talent and us economic growth. *Econometrica*, 87(5), 1439-1474, (2019).

Hufe, P., Peichl, A., Roemer, J., and Ungerer, M.: Inequality of income acquisition: The role of childhood circumstances. *Social Choice and Welfare*, 49(3-4), 499–544, (2017).

Konow, J.: Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review,* 90 (4), 1072-1092, (2000).

Marrero, G., Rodriguez, J.: Inequality of Opportunity and Growth. *Journal of Development Economics, 104, 107-122, (2013).*

Moramarco, D., Palmisano, F., Peragine, V.: Intertemporal inequality of opportunity. SERIES Working Papers 07-2020, Dipartimento di Economia e Finanza - Università degli Studi di Bari "Aldo Moro", (2020).

Muñoz, Ercio and Roy van der Weide. Intergenerational income mobility around the world: A new database. World Bank Policy Research Working Paper 11166.

Neidhöfer, G., Serrano, J., Gasparini, L: Educational inequality and intergenerational mobility in Latin America: a new database. *Journal of Development Economics,* 134, 329-349, (2018).

Peragine, V.: Opportunity egalitarianism and income inequality. *Mathematical Social Sciences*, *44*(1), 45-64, (2002).

Peragine, V., Serlenga, L.: Higher education and equality of opportunity in Italy. In *Inequality And Opportunity: Papers From The Second Ecineq Society Meeting* (pp. 67-97). Emerald Group Publishing Limited, (2008).

Pew Research Center: Trends in American values: 1987–2012. Partisan polarization surges in Bush, Obama years. 2012, June 4. https://www.pewresearch.org/wp-content/uploads/sites/4/legacy-pdf/06-04-12-Values-Release.pdf

Ramos, X., Van de Gaer, D.: Approaches to inequality of opportunity: Principles, measures and evidence. *Journal of Economic Surveys*, 30(5), 855-883, (2016).

Roemer, J. E. and Trannoy, A.: Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, 54 (4), 1288–1332, (2016).

Roemer, J. E.: A pragmatic theory of responsibility for the egalitarian planner. *Philosophy & Public Affairs*, 22 (2), 146-166, (1993).

Roemer, J. E.: Theories of distributive justice. *Harvard University Press, 1998.*

Shapley, L. S.: A Value For n-Person Games. *Proceedings of the National Academy of Sciences*, (1952).

Shorrocks, A. F.: Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *Journal of Economic Inequality*, 11 (1), 99–126, (2013).

Solon, G.: Intergenerational income mobility in the United States. *The American Economic Review*, 393-408, (1992).

van de Gaer, Dirk. *Equality of Opportunity and Investment in Human Capital*. PhD Dissertation. Katholieke Universiteit Leuven. (1993).

van der Weide, R., Lakner, C. , Mahler, D. G., Narayan, A., Gupta, R.: Intergenerational mobility around the world: A new database. *Journal of Development Economics* 166, 103-167, (2024).

Varian, H. R.: Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28(2), 3-28, (2014).

Appendix A

Table A1 provides information about the data sources used to construct GEOM. It includes the country name, the year, and the official name of each survey. The last column refers to the research team responsible for obtaining, cleaning, and harmonizing the data.

ADB & Monash University team includes: Gaurav Datt, Arturo Martinez Jr., John Nguyen, Albert Park

CEDLAS team includes: Matías Ciaschi

CEEY team includes: Pedro Torres López

LSE team includes: Luis Barajas, Paolo Brunori, Nancy Daza-Báez, Francisco H.G. Ferreira, Pedro Salas-Rojo, Louis Sirugue, Pedro Torres López

University of Bari team includes: Teresa Barbieri, Vito de Sandi, Fabio Farella, Domenico Moramarco, Vito Peragine, Enza Simeone, Giorgia Zotti

Table A1: Surveys used in GEOM

Country	Year/Years Covered	Survey and Acronym	Research team
Argentina	2014	Encuesta Nacional sobre la Estructura Social (ENES)	CEDLAS
Armenia	2016	Life in Transition Survey (LITS)	ADB & Monash University
Australia	2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019	Household, Income and Labour Dynamics in Australia (HILDA)	ADB & Monash University
Austria	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Belgium	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Benin	2018	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari
Bolivia	2008	Encuesta de Hogares (EH)	CEDLAS
Brazil	2014	Pesquisa Nacional por Amostra de Domicilios Contínua Anual (PNAD)	CEDLAS
Bulgaria	2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Burkina Faso	2018	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari

Chile	2006, 2011, 2013, 2015	Encuesta de Caracterización Socioeconómica Nacional (CASEN)	CEDLAS
China	2010, 2012, 2014, 2016, 2018	China Family Panel Studies (CFPS)	LSE
Colombia	2010	Encuesta Nacional de Calidad de Vida (ENCV)	CEDLAS
Croatia	2011	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Cyprus	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Czech Republic	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Denmark	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Ecuador	2006, 2014	Encuesta Condiciones de Vida (ECV)	CEDLAS
Estonia	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Finland	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
France	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Gambia	2015	Integrated Household Survey (HIS)	University of Bari
Georgia	2016	Life in Transition Survey (LITS)	ADB & Monash University
Germany	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Ghana	2013, 2017	Ghana Living Standard Survey (GLSS)	University of Bari
Greece	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Guatemala	2000, 2006, 2011	Encuesta Nacional sobre Condiciones de Vida (ENCOVI)	CEDLAS
Guinea Bissau	2018	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari

Hungary	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Iceland	2005	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
India	2005, 2012	India Human Development Survey (IHDS)	ADB & Monash University
Indonesia	2000, 2014	Indonesian Family Life Survey (IFLS)	ADB & Monash University
Ireland	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Italy	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Ivory Coast	2018	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari
Kazakhstan	2016	Life in Transition Survey (LITS)	ADB & Monash University
Kyrgyzstan	2016	Life in Transition Survey (LITS)	ADB & Monash University
Latvia	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Lithuania	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Luxembourg	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Malawi	2020	Malawi Fifth Integrated Household Survey (MFIHS)	University of Bari
Mali	2019	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari
Malta	2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Mexico	2017	Encuesta de Movilidad Social (EMOVI)	Provided by CEEY
Mongolia	2016	Life in Transition Survey (LITS)	ADB & Monash University
Nepal	2003, 2011	Nepal Living Standards Survey (NLSS)	ADB & Monash University
Netherlands	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE

Niger	2018	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari
Nigeria	2019	Nigeria Living Standards Survey (NLSS)	University of Bari
Norway	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Panama	2003	Encuesta de Niveles de Vida (ENV)	CEDLAS
Peru	2001, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015	Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza (ENAHO)	CEDLAS
Poland	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Portugal	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Romania	2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Senegal	2018	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari
Sierra Leone	2011, 2018	Sierra Leone Integrated Household Survey (SLIHS)	University of Bari
Slovakia	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Slovenia	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
South Africa	2008, 2012, 2015, 2017	National Income Dynamics Study (NIDS)	LSE
South Korea	1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019	Korean Labour and Income Panel Study (KLIPS)	ADB & Monash University
Spain	2005, 2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Sweden	2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Switzerland	2011, 2019	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
Tajikistan	2016	Life in Transition Survey (LITS)	ADB & Monash University

Tanzania	2009, 2011, 2013	National Panel Survey (NPS)	University of Bari
Timor-Leste	2007, 2014	Timor-Leste Survey of Living Standards (TSLS)	ADB & Monash University
Togo	2018	Enquete Harmonisee sur le Conditions de Vie des Menages (EHCVM)	University of Bari
Uganda	2010, 2011, 2012, 2014	National Panel Survey (NPS)	University of Bari
United Kingdom (UK)	2005, 2011	European Union Statistics on Income and Living Conditions (EU-SILC)	LSE
United States of America (USA)	1970, 1972, 1974, 1976, 1978, 1980, 1982, 1984, 1986, 1988, 1990, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014	Panel Study of Income Dynamics (PSID)	LSE
Uzbekistan	2016	Life in Transition Survey (LITS)	ADB & Monash University

Source: Own elaboration.

Table A2: GEOM main results from latest wave (MLD results)

Country	**Year**	**Variable**	**Sample MLD**	**Tree (ex-ante)**	**Random Forest (ex-ante**	**Tree (ex-post)**	**Tree (ex-ante) Relative (%)**	**Random Forest (ex-ante) Relative (%)**	**Tree (ex-post) Relative (%)**
Argentina	2014	Income	0.276	0.044	0.050	0.051	15.875	18.086	18.340
Armenia	2016	Income	0.309	0.032	0.056	0.091	10.395	18.102	29.501
Australia	2019	Income	0.222	0.014	0.016	0.011	6.385	6.969	4.856
Austria	2019	Income	0.156	0.017	0.016	0.016	10.740	10.289	10.096
Belgium	2019	Income	0.106	0.018	0.019	0.021	16.823	17.951	20.113
Benin	2018	Consumption	0.199	0.036	0.032	0.033	18.159	16.097	16.650
Bolivia	2008	Income	0.505	0.131	0.144	0.238	25.971	28.566	47.048
Brazil	2014	Income	0.427	0.165	0.163	0.191	38.725	38.115	44.843
Bulgaria	2019	Income	0.296	0.081	0.062	0.058	27.193	20.884	19.669
Burkina Faso	2018	Consumption	0.235	0.042	0.025	0.024	18.035	10.506	9.996
Chile	2015	Income	0.458	0.089	0.095	0.232	19.423	20.843	50.710
China	2018	Income	0.459	0.062	0.076	0.172	13.413	16.576	37.535
Colombia	2010	Income	0.547	0.101	0.106	0.252	18.424	19.430	46.061
Croatia	2011	Income	0.177	0.016	0.018	0.022	9.250	10.378	12.408
Cyprus	2019	Income	0.169	0.041	0.041	0.049	23.964	24.438	28.935
Czech Rep.	2019	Income	0.099	0.009	0.010	0.008	9.119	9.929	8.409
Denmark	2019	Income	0.143	0.006	0.008	0.005	3.860	5.614	3.649
Ecuador	2014	Income	0.377	0.081	0.082	0.142	21.497	21.815	37.712
Estonia	2019	Income	0.151	0.009	0.011	0.011	5.781	7.110	7.442
Finland	2019	Income	0.143	0.014	0.019	0.014	9.930	13.357	9.720
France	2019	Income	0.145	0.020	0.028	0.025	13.941	19.255	17.046
Gambia	2015	Income	0.670	0.066	0.058	0.112	9.885	8.705	16.694
Georgia	2016	Income	0.381	0.041	0.069	0.053	10.738	18.089	13.941
Germany	2019	Income	0.138	0.011	0.012	0.009	7.959	8.900	6.657

Country	Year	Variable	Sample MLD	Tree (ex-ante)	Random Forest (ex-ante	Tree (ex-post)	Tree (ex-ante) Relative (%)	Random Forest (ex-ante) Relative (%)	Tree (ex-post) Relative (%)
Ghana	2017	Consumption	0.314	0.039	0.042	0.052	12.460	13.225	16.412
Greece	2019	Income	0.166	0.020	0.023	0.032	12.205	13.595	19.215
Guatemala	2011	Income	0.519	0.142	0.136	0.251	27.395	26.258	48.429
Guinea Bissau	2018	Consumption	0.158	0.032	0.029	0.029	20.164	18.268	18.458
Hungary	2019	Income	0.146	0.008	0.012	0.010	5.693	8.505	6.996
Iceland	2005	Income	0.129	0.007	0.011	0.011	5.728	8.437	8.204
India	2012	Income	0.518	0.135	0.123	0.207	26.144	23.711	39.988
Indonesia	2014	Consumption	0.309	0.025	0.027	0.020	8.161	8.582	6.541
Ireland	2019	Income	0.136	0.018	0.024	0.022	12.932	17.634	15.871
Italy	2019	Income	0.207	0.022	0.019	0.026	10.516	9.214	12.590
Ivory Coast	2018	Consumption	0.173	0.028	0.022	0.020	16.338	12.572	11.298
Kazakhstan	2016	Income	0.203	0.011	0.011	0.008	5.170	5.170	4.037
Kyrgyzstan	2016	Income	0.362	0.021	0.032	0.168	5.668	8.930	46.337
Latvia	2019	Income	0.214	0.013	0.018	0.019	6.121	8.224	8.925
Lithuania	2019	Income	0.220	0.018	0.017	0.018	8.330	7.601	8.375
Luxembourg	2019	Income	0.194	0.031	0.032	0.031	15.866	16.486	16.021
Malawi	2020	Consumption	0.208	0.054	0.051	0.057	25.830	24.579	27.177
Mali	2019	Consumption	0.190	0.032	0.027	0.030	16.974	14.075	15.867
Malta	2019	Income	0.126	0.012	0.013	0.014	9.192	10.301	11.252
Mexico	2017	Income	0.495	0.152	0.142	0.126	30.662	28.643	25.454
Mongolia	2016	Income	0.412	0.036	0.053	0.057	8.614	12.788	13.832
Nepal	2011	Income	0.513	0.083	0.076	0.122	16.118	14.714	23.836
Netherlands	2019	Income	0.115	0.007	0.012	0.031	6.092	10.183	27.241
Niger	2018	Consumption	0.160	0.029	0.027	0.027	17.955	16.584	16.584

Country	Year	Variable	Sample MLD	Tree (ex-ante)	Random Forest (ex-ante	Tree (ex-post)	Tree (ex-ante) Relative (%)	Random Forest (ex-ante) Relative (%)	Tree (ex-post) Relative (%)
Nigeria	2019	Consumption	0.138	0.025	0.023	0.027	18.083	16.412	19.390
Norway	2019	Income	0.149	0.018	0.020	0.020	11.836	13.584	13.181
Panama	2003	Income	0.636	0.156	0.132	0.209	24.595	20.711	32.851
Peru	2015	Income	0.327	0.109	0.113	0.136	33.394	34.557	41.621
Poland	2019	Income	0.147	0.011	0.012	0.013	7.645	8.328	9.147
Portugal	2019	Income	0.172	0.024	0.030	0.030	14.095	17.401	17.459
Romania	2019	Income	0.233	0.038	0.033	0.074	16.452	14.003	31.830
Senegal	2018	Consumption	0.160	0.028	0.022	0.025	17.470	13.713	15.341
Sierra Leone	2018	Consumption	0.158	0.035	0.036	0.036	21.984	22.615	22.552
Slovakia	2019	Income	0.107	0.009	0.009	0.017	8.675	8.582	15.765
Slovenia	2019	Income	0.110	0.011	0.012	0.013	10.154	10.789	12.149
South Africa	2017	Income	0.690	0.288	0.360	0.413	41.707	52.108	59.844
South Korea	2019	Income	0.224	0.022	0.025	0.062	9.772	11.111	27.622
Spain	2019	Income	0.211	0.038	0.036	0.045	17.811	16.817	21.222
Sweden	2019	Income	0.166	0.027	0.017	0.021	16.325	10.000	12.470
Switzerland	2019	Income	0.145	0.011	0.015	0.009	7.634	10.110	6.465
Tajikistan	2016	Income	0.160	0.008	0.012	0.008	5.240	7.361	5.240
Tanzania	2013	Consumption	0.235	0.045	0.047	0.046	19.250	19.889	19.676
Timor Leste	2014	Consumption	0.129	0.023	0.017	0.021	17.702	13.432	16.149
Togo	2018	Consumption	0.242	0.044	0.037	0.037	18.249	15.400	15.153
Uganda	2014	Consumption	0.229	0.051	0.053	0.056	22.451	23.239	24.464
United Kingdom	2011	Income	0.187	0.011	0.015	0.016	5.678	7.874	8.516
United States of America	2014	Income	0.290	0.039	0.044	0.054	13.542	15.196	18.470
Uzbekistan	2016	Income	0.363	0.018	0.052	0.014	5.015	14.191	3.720

Source: Own elaboration.

Appendix B: Estimation methods

B1 - Ex-post Inequality of Opportunity

The approach to the estimation of ex-post Inequality of Opportunity (IOp) follows the method proposed by Brunori, Ferreira and Salas-Rojo (2024) based on the Transformation Trees (Trafotrees) introduced by Hothorn and Zeileis (2021) to define the most appropriate partition in Roemerian types.

The Transformation Tree algorithm

Trafotrees are analogous to Ctrees, but the sequence of tests used to grow the tree concerns the conditional distribution of the outcome instead of the sole conditional expectation. Trafotrees estimates distribution functions for each possible partition. Contrary to Ctree, for which the first step consists in a correlation test between the outcome and each circumstance, Trafotree first steps consists in a test on all possible ways of using values of each circumstance to partition the sample, making the algorithm substantially more computationally expensive. Then, the algorithm performs a binary splitting in the sample, sequentially creating two groups whose distribution functions are least likely to be the same. The population is therefore partitioned into an exhaustive and mutually exclusive set of subgroups (types). The algorithm can be summarized as follows:

1. Set a confidence level (1 - α).
2. Set a Bernstein polynomial order *m*.
3. Approximate the shape of the unconditional distribution of the outcome using a Bernstein polynomial of order *m*. Store the *m+1* parameter values.
4. Run a test to assess the instability of the parameters conditional on the value of each circumstance.[30] If the Bonferroni-adjusted *p-value* of the association test is higher than the chosen critical value α, exit. Otherwise, continue to step 5.
5. Among all regressors in which the null hypothesis of independence is rejected, select the variable producing the smallest adjusted *p-value* as splitting variable *[c].*
6. Consider how circumstance *[c]* can be used to partition the sample into two subsamples *[C]*. Among all possible binary partitions, compute the *p*-value for the null hypothesis that the distribution in two sub-samples is the same *(*$p^{[C]}$).
7. Choose $[\mathrm{C}]^* = \{[\mathrm{C}]\colon \mathrm{argmin}\ \ \mathrm{p}^{[\mathrm{C}]}\}$ as the most appropriate partition.

[30] This test can be understood as a test to detect a structural break in a time series (Chow, 1960). The cumulative sum of residuals is used to detect whether the parameters are unstable over time, while in this case the instability is investigated conditioning on the values or categories of each circumstance (see Hothorn and Zeileis, 2021) for details).

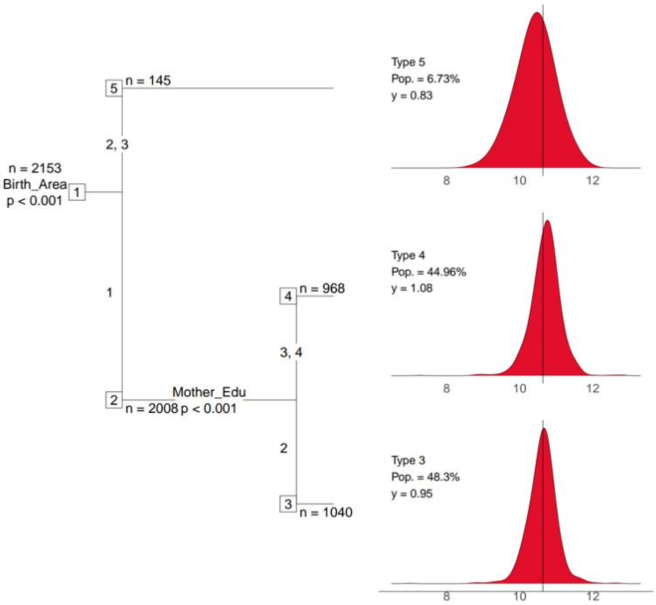
8. Repeat steps 3 – 7 for each resulting node (sub-sample) until the null hypothesis of step 3 cannot be rejected in any resulting sub-sample.

We use the “trafotree” R function developed by Hothorn and Zeileis (2021). The output of the estimation consists of a partition in types, that allows us to obtain a parametric interpolation of each type’s cumulative income distribution function.[31] These parametric conditional distributions can be inverted to yield the predicted 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 $\widehat{IOp} = I_q\left(\frac{\hat{\mu}}{\hat{\mu}_q}\hat{y}_{qc}\right)$.

Visualization of the results

Same as Ctrees, Trafotrees can also be graphically represented, providing information on the structure of inequality of opportunity within a particular observed population, now based on differences on the outcome distributions. Trafotrees in the database are displayed as in Figure B1. The type-specific parametric CDFs obtained in Trafotree can be represented as in Figure B2. They can also be aggregated into the overall density function as a mixture of type distributions.

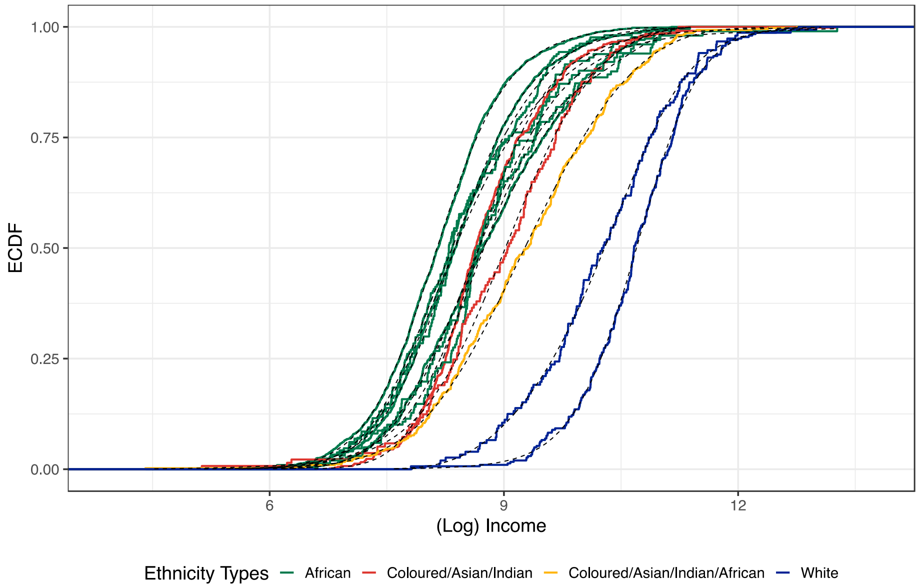
Figure B1: Transformation Tree example (Denmark, 2011).



Source: GEOM. Data from EUSILC 2011.

[31] The same interpolation was implemented in Brunori and Neidhöfer (2021) on a partition of types obtained with Ctree.

Figure B2: Type CDFs in South Africa (2017).



Source: Brunori, Ferreira, Salas-Rojo (2024).

As for the Ctrees described in the main text, we follow the convention and set $\alpha = 0.01$. We impose an additional requirement. Each terminal node must have a minimum of 1% of the observations in the sample (or 50 if the sample size is smaller than 5000). This country-specific minimum is set to minimize the effect of different sample sizes on the depth of the tree. For robustness, we run a second tree relaxing the previous requirement, such that minimum observations imposed in each terminal node is 0.1%. If the types obtained are different from those with the node-size restriction, we store the plot for further inspection. All remaining parameters are the default set in the "*ctree*" R function in the package "partykit" (Hothorn, Seibold and Zeileis, 2023).

We do not use weights to determine splits. Including sampling weights expands the sample size, such that individual observations are turn into hundreds or thousands of identical values. As a result, the tree becomes very deep, as null hypothesis are easily rejected. Weights are used to calculate the values of the counterfactual distribution and to estimate IOp.

Unlike Ctrees, transformation trees require the practitioner to choose the order of the Bernstein polynomial used to approximate the type-specific conditional distribution functions. We choose that order by setting a minimum improvement in the aggregate out-of-sample log-likelihood of 0.1%. All other parameters are the default parameters in the "*Trtf*" R function in the package "Trtf" (Hothorn, 2023).

B2. The role of individual circumstances: Shapley-Shorrocks decompositions and partial dependency plots

Since there is no guarantee that the contributions of all circumstance variables are additively separable, a plausible approach to identifying individual contributions is through a Shapley value decomposition (See Shapley, 1952; and Shorrocks, 2013)). Intuitively, a Shapley value decomposition calculates the overall contribution of each variable *c*, included among explanatory variables, to the variability of some outcome *y.* The Shapley value of variable *c* is calculated as the average decline in the explained variability of *y* resulting from all possible combinations of ways in which it can be explained without including *c* among explanatory variables.[32]

We follow Brunori, Ferreira and Salas-Rojo (2024) and obtain the Shapley value decompositions as follows:

1. Draw a sub-sample of the full sample. To favor computational speed, the sub-sample should consist of 5,000 observations, or 90% of the original sample size if it was smaller than 5,000.
2. Estimate IOp in the sub-sample by estimating a Ctree, allowing the tree to overfit and get deep ($\alpha = 0.9$), and minimum sample size in terminal nodes defined as in the random forest (0.1% of the sample size).
3. Re-estimate IOp in the sub-sample for all possible elimination sequences of each circumstance. Elimination of one or more circumstances is obtained by replacing their values with a vector of 1.
4. Estimate the difference between the overall IOp and the new IOp values obtained after different elimination sequences of each circumstance. Estimate the weighted average of these differences as the contribution of *c*.
5. Accounting for possible data or sample dependencies, repeat steps 1-4 one hundred times.
6. The final estimate of the contribution of c to IOp is the average contribution across 100 iterations.

Note that contributions of each circumstance are reported in relative terms. Absolute values are not directly comparable with IOp estimated with a single tree, because the sample sizes are smaller and confidence level is lower. We perform the decomposition estimating 100 trees on different sub-samples and we calculate average values across iterations. This procedure

[32] Note that according to Shapley (1952) each elimination sequence has a different probability that is used as weight to obtain such average decline.

makes our estimates robust to the high variance typical of a single tree. Shapley value decompositions implemented in this fashion were discussed in Section 5.4.

Partial Dependency Plots

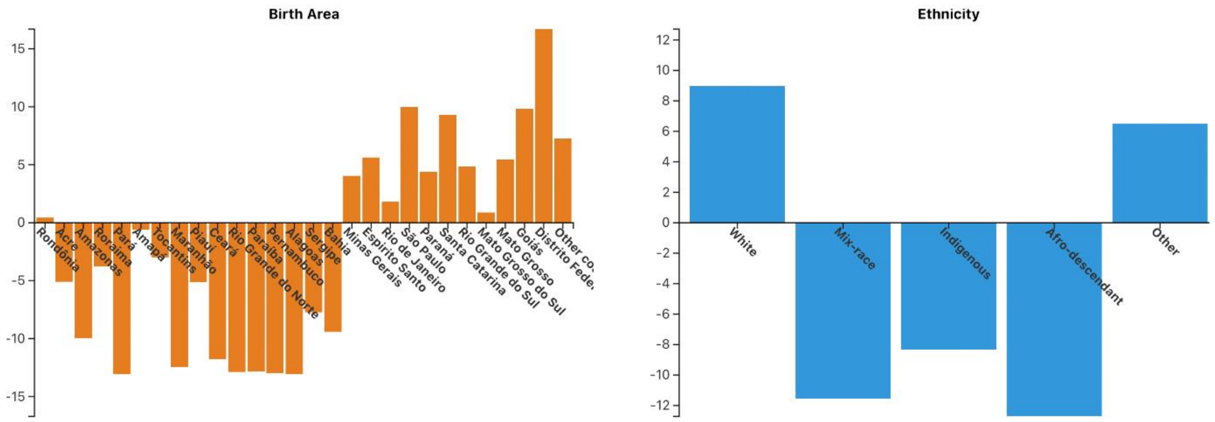
To overcome the limitations in the interpretation of trees we also derive partial dependence plots (PDPs) from random forests, who are conceptually similar to marginal effects from OLS regressions. The process for generating PDPs can be outlined as follows:

1. Run the random forest regression, where the outcome variable *y* is predicted as a function of regressors *X*. Formally, $\hat{y} = \hat{f}(X)$.
2. Duplicate the original dataset and select a predictor variable, X_k.
3. Take the initial value or category of the selected predictor variable, denoted as *j*.
4. Replace all values in X_k with *j* while leaving the remaining predictor variables unchanged. Utilize the random forest model estimated in the first step to predict $\hat{y}_{pdp}$, now using this modified dataset. Compute the average of $\hat{y}_{pdp}$, which is the average outcome associated to all individuals in the dataset in the counterfactual situation in which they all have value *j* in regressor *X*, all else equal.
5. Repeat steps 2-4 for all values *j* within each regressor *X*, and for all regressors.
6. Plot all mean predictions of $\hat{y}_{pdp}$.

For discrete or categorical predictor variables, such as the circumstances we use in our analysis, the PDP typically displays all categories on the x-axis and presents the associated conditioned expected values (mean of $\hat{y}_{pdp}$) on the *y-axis*. Figure B3 shows the results for Brazil 2014 as an example. The sample mean income is US$12,882 in 2017 prices. After running the random forest and obtaining the associated PDP, we find individuals born in Paraná to earn, on average, 5% more than the sample mean, while those born in Pará earn 13% less.

PDPs offer several advantages. They are derived from random forests, ensuring that each category receives the expected outcome by averaging across many trees. This property enhances their robustness, making their interpretation less dependent on specific data instances compared to single regression trees. Additionally, they are straightforward to interpret and complement Shapley value decompositions (see below), allowing for the interpretation of nonlinearities in the data-generating process beyond just average effects of predictor variables.

Figure B3: Partial Dependence Plot (PDP) for Brazil 2014.



Source: Own Elaboration. Data from PNAD (2014).

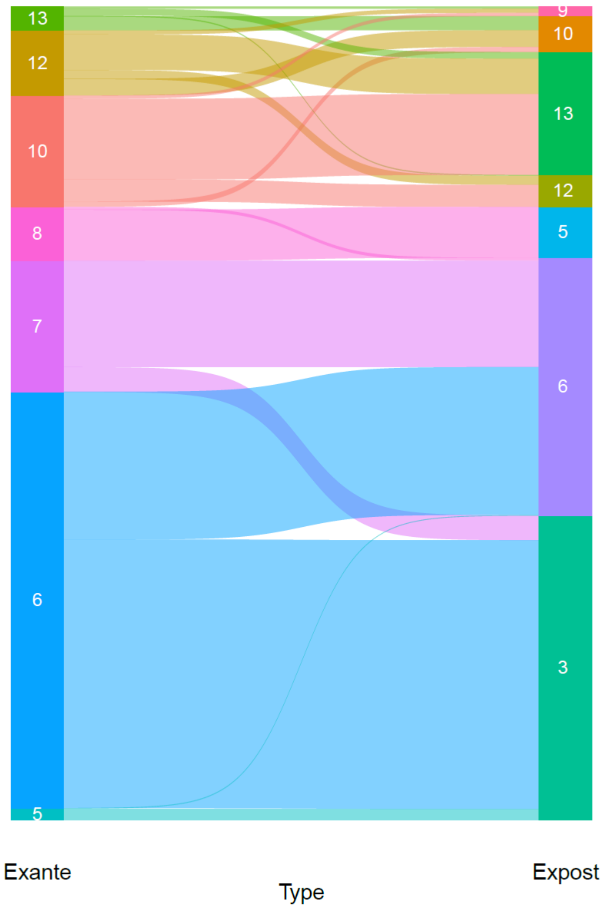
B3. Complementary analyses

The high variance and sample dependence issues exposed in the Ctree explanation also apply in Trafotrees. However, the aggregation of multiple overfitted Trafotrees into a Trafo forest turns out to be problematic. As ex-post IOp is not measured as between-type inequality but by assigning individuals a rank in the type-specific distribution and then evaluating inequality within-quantile across the distribution, the aggregation of multiple draws appears to induce a severe downward bias in the estimation of IOp (see Brunori, Ferreira, Salas-Rojo, 2024) for a discussion). For this reason, the GEOM database does not contain estimates of ex-post IOp based on Trafotree random forests.[33] However, to provide robust evidence also about ex-post IOp, the relative importance of circumstances is addressed with a Shapley value decomposition identical to the one described for the ex-ante IOp. Shapley values are again reported in relative term to avoid interpreting their absolute value obtained on overfitted trees estimated on subsamples of the entire sample.

Finally, since Ctrees and Trafotrees are different algorithms, that respectively consider the mean and the complete outcome distribution, GEOM also provides a visualization tool designed to compare the two kinds of partitions. These are Sankey (or alluvial) diagrams, like the one shown in Figure B4, which map the type to which each individual belongs across the two partitions: ex ante and ex post.

[33] Hothorn and Zeileis (2017) do propose a method to obtain prediction of a dependent variable from forest of transformation trees. However, this is conceptually different from estimating the counterfactual distribution needed to quantify ex-post IOp.

Figure B4: Sankey Diagram from Croatia (2019).



Source: Own elaboration from GEOM. Original data from EUSILC 2019.