Using Machine Learning to Unveil the Predictors of Intergenerational Mobility[*]

Luís Clemente-Casinhas
Instituto Universitário de Lisboa (ISCTE-IUL) and Business Research Unit (BRU-IUL)

Alexandra Ferreira-Lopes
Instituto Universitário de Lisboa (ISCTE-IUL) and Business Research Unit (BRU-IUL)

Luís Filipe Martins
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), and CIMS-University of Surrey

Abstract

We assess the predictors of intergenerational mobility in income and education for a sample of 137 countries, between 1960 and 2018, using the World Bank's Global Database on Intergenerational Mobility (GDIM). The Rigorous LASSO and the Random Forest and Gradient Boosting algorithms are considered, to avoid the consequences of an *ad-hoc* model selection in our high dimensionality context. We obtain variable importance plots and analyse the relationships between mobility and its predictors through Shapley values. Results show that intergenerational income mobility is expected to be positively predicted by the parental average education, the share of married individuals and negatively predicted by the share of children that have completed less than primary education, the growth rate of population density, and inequality. Mobility in education is expected to have a positive relationship with the adult literacy, government expenditures on primary education, and the stock of migrants. The unemployment and poverty rates matter for income mobility, although the direction of their relationship is not clear. The same occurs for education mobility and the growth rate of real GDP *per capita*, the degree of urbanization, the share of female population, and income mobility. Income mobility is found to be greater for the 1960s cohort. Countries belonging to the Latin America and Caribbean region present lower mobility in income and education. We find a positive relationship between predicted income mobility and observed mobility in education.

JEL Classification: E24; I24; J62; O15.

Keywords: Intergenerational Mobility in Income and Education; Predictors of Intergenerational Mobility; Rigorous Least Absolute Shrinkage and Selection Operator; Random Forest; Gradient Boosting; Shapley Values.

---

[*] We express gratitude to three referees and an editor for helpful contributions. We also thank Ambar Narayan, Daniel Mahler, and João Moura for their valuable help. We extend our appreciation to the participants of the 44th Meeting of the Association of Southern European Economic Theorists for their comments. We acknowledge financial support from FCT - Fundação para a Ciência e a Tecnologia (National Science and Technology Foundation) through grants 2020.04449.BD (DOI: 10.54499/2020.04449.BD) and UIDB/00315/2020 (DOI: 10.54499/UIDB/00315/2020).
The corresponding author is Luís Clemente-Casinhas. E-mail: luis_miguel_casinhas@iscte-iul.pt. Address: ISCTE Business School, Building II, Office D5.12, Avenida das Forças Armadas, 1649-026, Lisboa, Portugal.

1. Introduction

In the Organization for Economic Cooperation and Development (OECD) glossary of statistical terms, intergenerational mobility is defined as "the extent to which some key characteristics and outcomes of individuals differ from those of their parents."[1] These key characteristics, either positive or negative, have a direct impact not only on the individual who bears them, but on society as well. For example, intergenerational persistence in income and education have been identified as key predictors of inequality. Hence, the study of the predictors of intergenerational mobility (IM) of income and education is very important to properly define policies that can help to disentangle problems of IM and consequently problems of inequality.

In this work we use a database with data for 137 countries from 1960 to 2018 to assess the most important variables predicting income and education IM. We study IM in income as well as in education, since it is possible that these two dimensions may not evolve in the same direction, and thus avoid misinterpreting the results. Following most of the empirical literature, we consider IM measured in relative terms, which reflects the degree of dependence of children's future outcomes on their parents' outcomes; contrary to absolute upward mobility, which reflects the extent to which a specific generation's outcome is better than the previous generation outcome. We rely on a comprehensive literature review to assess the predictors of IM in income and education.

Most existing studies on intergenerational income and educational mobility rely on computing the value for IM and associating this value with differences in other variables through, for example, simple correlations (the most obvious is geography). This is mainly done for a single country or a limited small set of economies. Very few authors parametrically model these relationships in an attempt to find which factors may predict mobility. Deciding which variables to choose when computing mobility regressions is a non-trivial challenge and, due to authors' arbitrariness, may result in estimation biases. This is particularly important in light of the increasing availability of datasets.

The consequences of an *ad-hoc* model selection is shared by Brunori *et al*. (2023). The authors advocate that not selecting relevant variables limits the explanatory ability of a model, while introducing too many variables will result in overfitting. This may also occur if the model presents the incorrect functional form. They suggest the use of machine learning algorithms, which are not rigid regarding the relationships under study and, at the same time, use out-of-sample replicability criteria. Estimating equality of opportunity for 31 European countries, these authors show that conditional inference trees and forests minimize the discretion, which is inherent in the model selection employed by the researcher.

Our contribution to the literature is the following. Heterogeneity in mobility is conditional on different features, meaning that these features might themselves be potential predictors of intergenerational mobility/persistence in income and education. We are the first to present a Worldwide outlook to find the predictors of IM, filling this gap in existing research. For this we use the Global Database on Intergenerational Mobility (GDIM) from the World Bank, containing measures for income and educational mobility, which are comparable across countries. The intergenerational persistence of income is estimated using OLS or IV-type estimators by regressing the child's earnings on the parents' earnings around the reference age and the intergenerational persistence of education by regressing the children's years of schooling on parents' highest years of schooling. The information used on mobility is provided as 10-year averages, considering the cohorts of 1960s, '70s, and '80s. Grounded on the literature, we construct an inclusive database that includes all the predictors of IM identified in earlier works for which information

[1] https://stats.oecd.org/glossary/detail.asp?ID=7327.

is available for the widest number of countries possible. From an initial dataset for the period 1960-2018, we make these two datasets compatible by averaging over time the potential regressors considering different time periods, which are predicted by each generation in GDIM.

Lee and Lee's (2020) work is the closest to ours in the sense that they explore the predictors of educational mobility, although their study targets OECD countries and few predictors are considered. Only Kourtellos (2021) uses the same database as ours to explore the relationship between inequality and IM in education, but differs from our work as it considers a measure of absolute mobility and controls for few variables, thereby lacking a wider coverage of the literature regarding what predicts IM.

We are also the first to take advantage of Machine Learning algorithms in the context of IM research. We use the Random Forest and the Gradient Boosting methods to uncover the mobility in income and education predictors, through hyperparameters optimization to avoid overfitting and at the same time improve accuracy. These complement the evidence produced by the Rigorous Least Absolute Shrinkage and Selection Operator (RLASSO), which may penalize variables that are important for mobility but not selected. By using different methods we aim to validate our results, given our high dimensional database. Additionally, grounded on Brunori *et al*. (2023), trees and forests require minimal assumptions about which and how predictors influence mobility, accommodating in different ways the relationships that may occur, and incorporating less noise. We also consider Shapley values to understand the contribution of individual predictors to mobility predictions, because machine learning algorithms do not in principle provide information on the direction of the relationships we seek to reveal. Finally, after knowing which factors are the ones that mainly predict IM, we are able to predict intergenerational income persistence for the countries that present only observed values for intergenerational education mobility.

Results show a positive connection between intergenerational income mobility and both the parental average education and the share of married individuals, while and a negative relationship is expected with the share of children that have completed less than primary education, the growth rate of population density, and inequality. The 1960s cohort presents higher income mobility. Mobility in education is positively predicted by adult literacy, government expenditures on primary education, and the migrant stock. Although income mobility appears to be predicted by unemployment and poverty rates, the direction of their relationship is not clear. The same occurs regarding the relationship between education mobility and the real GDP *per capita* growth rate, the degree of urbanization, the share of female population and the intergenerational income mobility. Countries belonging to the Latin America and Caribbean region present lower mobility. Our evidence shows that developing economies face a disadvantage: they are the ones presenting the highest values of predicted income persistence. It also shows that persistence in education estimates are positively connected with income persistence predictions, although their relationship is modest. This implies that high-income countries appear to benefit in terms of IM in education when compared to the developing group.

This paper is organized as follows. Section 2 describes our empirical methodology, by defining the variables in our database as well as our statistical approach. Section 3 presents and discusses our benchmark results. Section 4 explores the contribution of each feature for individual predictions of mobility, using Shapley values. In Section 5 we use our model to predict income mobility for a specific set of countries and study the relationship between income and education mobility. Section 6 concludes.

2. Empirical Methodology

In this section we present the data and the statistical methods.

2.1. Data Description

Here we describe both the dependent and independent variables. Our dataset contains information for the period 1960-2018 with respect to 137 countries. The acronyms used in our database as well as in our results tables are presented in parentheses.

2.1.1. Intergenerational Mobility Measures

The variables we use are taken from the Global Database on Intergenerational Mobility (GDIM, 2018) constructed by the World Bank, containing mobility estimates by 10-year birth cohorts for the period 1940-1989.

IM can be interpreted in both absolute and relative terms[2]. According to GDIM (2018), absolute upward mobility reflects the extent to which a specific generation's outcome is better than the previous generation's outcome, while relative mobility measures the extent to which, for a given generation, individuals' outcomes are independent of their parents' outcomes. The authors of the database illustrate the differences between the two by considering that an individual's economic success relative to others may be reflected in the different rungs of the same economic ladder: having more absolute upward mobility means that the current generation was able to climb up the ladder relative to the previous generation (children are better off than their parents); while relative mobility occurring means that an individual will be on a different rung of the ladder among their peers (born in the same generation), when compared to the rung his or her parents occupied among their peers (e.g., children of parents relatively poor in the parents' generation attaining middle or upper class in their generation).

Our work considers only relative measures of mobility in income and education for three main reasons. First, we follow the leading literature on mobility. Second, the GDIM database presents only a relative measure for income and we also aim to study the relationship between mobility in income and education. Therefore, mobility in education should be measured in the same way as income. Third, it is clear from the different definitions presented that one can be mobile in absolute terms but that situation may not be translated into relative mobility. Deutscher and Mazumder (2023) state that changes in relative mobility could capture a diversity of income movements with divergent orientations, while absolute upward mobility could capture wide increasing levels of education and economic growth. Hence, upward mobility will result, with no doubtfulness, in a welfare improvement. Nevertheless, as countries change from developing to developed, as pointed out by the OECD (2018), absolute mobility will decelerate, and the focus should be on relative mobility at this point, since it allows for a better appraisal of inequality in a society. Policy makers should devote more of their attention to relative mobility because that is the measure that will allow them to assess if individuals are improving their current living standards (which must always be analysed in relative terms).

We work on the father-son pair, as Helsø (2020) states that comparisons between countries are often based on income mobility between sons and their fathers. We follow the same option since we are working with a worldwide view on mobility. The fact that we also study IM in education makes us study the father-son pair in this context as well. An important reason presented in the literature for this raises some concerns in the measurement of women's earnings in the event that they are married. Ermisch *et al.* (2006) consider that if male labour force participation of

[2] Although it is recognized that downward movements may occur, the focus of absolute mobility is usually the upward direction, since it is the one associated with higher income growth and shared prosperity (GDIM, 2018).

married men surpasses that for women, as is usually the case, this may reflect that there is no randomness in the choice women make regarding working. Cervini-Plá (2014) complements this view by arguing that this decision may reflect that they could be part of households with characteristics that justify their participation in the labour market, as belonging to a household in which a single person working is not enough to support the couple's expenses. Therefore, when married women are included in the analysis, their individual earnings may not accurately reflect their economic status. Also, women may be absent from the labour market due to maternity related issues, and therefore, their activity status may be intermittent, which supports the previous view. GDIM (2018) considers both genders for parents and children. For several countries, and for males and females, persistence in income is measured as persistence in individual earnings using an instrumental variable procedure with the Equalchances[3] methodology of 2018. This is not consistent, because if females are considered in GDIM (2018), individual earnings should not be used in their mobility estimates, since they are not a good measure of their "true" economic status, especially for married women. Adding to this, the values in the database divided by gender are typically equal. All of this leads us to consider that estimates for females may not be properly estimated, which reinforces the exclusive use of males in our work.

In the GDIM (2018) database, regarding the indicators constructed for relative measures, positive changes in the indicators are signs of more intergenerational persistence, negative changes in the indicator are signs of more intergenerational mobility.

Intergenerational Mobility in Income

- Intergenerational persistence of income or elasticity (IGPI), being the estimated coefficient from i) either regressing, through OLS, child's earnings on the parents' earnings around the reference age (both in logarithms), or ii) from the final of three sequential steps using four instrument-related estimation methods, namely the two-sample two-stage instrumental variables estimation (TSTSIV) or two-stage least squares (TSLS): two-sample instrumental variables estimation (TSIV) and instrumental variables estimation (IV) are substitute designations for these methods in the literature. First, one regresses income on a list of variables reflecting parents' characteristics such as parents' age and education on a sample, which represents the current parents' population when younger; second, the coefficients reflecting parental education and experience returns are used to predict parental earnings for a reference age; third, regress children's earnings on the predicted earnings for parents. This means that only the OLS estimates are computed using income information of both generations of parents and children. Persistence in income ranges between approximately 0.1 and 1.1, considering all countries and cohorts, meaning that increases in father's income will always be associated with increases in son's income, although not always in the same proportion: there may be the case in which the change for the child is greater than the one for the parent. The higher the value, the greater the dependence the second generation has regarding the first one.

Intergenerational Mobility in Education

- Intergenerational persistence of education (IGPE): coefficient obtained by regressing children's years of schooling on parents' highest years of schooling. It ranges between approximately -0.2 and 1.0, meaning that

[3] http://equalchances.org

there are countries in which the change in the child's education will never exceed that of the parent, in fact, it may actually decrease relative to that of the father. [4]

Summary statistics for both intergenerational persistence measures are in Table 1.

Table 1 – Summary Statistics for Intergenerational Persistence Measures

Variables	Cohorts	Obs.	Mean	Std. Dev.	Min	Max
IGPI	1960	34	0.38	0.15	0.11	0.68
	1970	36	0.67	0.25	0.24	1.10
IGPE	1960	101	0.40	0.19	0.05	0.98
	1970	101	0.40	0.14	0.08	0.71
	1980	136	0.38	0.14	-0.21	0.82

2.1.2. Intergenerational Mobility Predictors

We will use as explanatory variables for IM in income and/or education the predictors for which data are available at a Worldwide level, supported by an extensive literature review. Since we have 137 countries, we have sought to find proxies for the predictors that are available for the widest number of countries possible[5]. The definitions of the mobility predictors and a table (Table A1) summarizing those predictors and the impacts they are expected to have on income and education mobility, based on the literature review, can be found in Appendix A.[6] In this section we present some explanations about the expected relationship between the predictors and intergenerational mobility, based on the results of previous literature.

Human capital related variables (e.g., adult literacy, children's educational attainment, human capital index, parental average education) can have ambiguous relationships with intergenerational mobility, both for income and for education. For example, on one hand, richer parents will invest more in their children's human capital when compared to poorer parents. This evidence is reflected in the persistence of economic status differences between generations. On the other hand, being born in a society with a high level of human capital can increase mobility, due to positive externalities from this higher level of human capital.

Public expenditures on education are positively related with intergenerational mobility since increasing public expenditures on education helps to narrow the gap between rich and poor parents that present differences in their ability to invest in their children's education. Variables also related to education such as school quality, high school enrolment, and early childhood development can provide contrasting relationships with intergenerational mobility. While school quality and high school enrolment have shown to provide positive signs for the referred

[4] The negative value of –0.21 belongs to Lesotho, which had quite a few unstable years in terms of political and military environment. This value for the elasticity of intergenerational mobility of education means that one extra year of education of the father implies minus 0.21 years of education for the son.

[5] There are some factors that can explain mobility, but we can't include them in the analysis. For example, Hofstede's cultural dimensions (power, individuals, masculinity, risk aversion, time preference, indulgence) could be explored. However, we searched for data available in a country level but failed to find for almost all countries that we have in our dataset. And it is not just in a countries' level but also time: we would need to mimic our 60, 70 and 80 cohorts. In any case, these cultural dimensions are likely to also be caused by the economic and demographic variables that we include in our study.

[6] We use as sources the World Development Indicators from the World Bank Database (World Bank, 2018a), the Global Database on Intergenerational Mobility from the World Bank, the Penn World Tables compiled by Feenstra *et al*. (2015), the Our World in Data project from the Global Change Data Lab (2021), the Global Debt Database from the International Monetary Fund (2018), and the World Values Survey by Inglehart *et al*. (2018).

relationship, which indicates that the quality of schools (i.e., the quality of the curriculum and the teaching staff) and the number of people enrolled in secondary education are of significant relevance for intergenerational mobility, in what concerns early childhood development, results show that dividing children earlier in life in homogeneous groups, according to their capabilities, decreases intergenerational mobility, while the earlier that children attend kindergarten, the higher will be intergenerational mobility in education.

Labour market related variables present very clear-cut results. While the unemployment rate, and in particular for young workers and also for workers with advanced education, clearly jeopardizes intergenerational mobility perspectives, since people are out of a job, without training or formation, decreases their possibilities of raising income levels; the reverse, i.e., a higher participation in the labour market, for all workers, but also especially for women, clearly increases their hypotheses of ascending in social and economic terms. Hence, participation in the labour market is of extreme relevance for increasing intergenerational mobility.

Empirical evidence for demographic variables shows that instability in terms of household structure (for instance with an increasing number of divorces or even with the number of single-parent families or teen births) decreases intergenerational mobility, while an increasing share of marriages, increase the possibility of higher intergenerational mobility. A stable household and composed of more than one person that earns an income is a positive environment to foster investment in the education of children. Factors such as fertility rates and child and maternal mortality rates are also relevant for intergenerational mobility. In what concerns migration evidence is mixed. On one hand children of immigrant parents usually exhibit a higher mobility than children born in that country, since the school system may be better in the country in which they currently live and job market opportunities are also more. But the opposite can also happen, since people can also migrate to countries where educational and income levels may also be lower than in the countries where they lived previously.

In what concerns income, evidence points to the positive relevance of it since investment in human capital can foster both intergenerational mobility in education and income, and it is more likely to invest with available income. However, strong disparities in income in a country jeopardize income and education intergenerational mobility, since the more agents face income constraints the harder it is to invest in human capital.

Unstable macroeconomic conditions, such as inflation or a recession, are also harmful for intergenerational mobility since they bring uncertainty, and with uncertainty comes declining investment, both in physical and human capital. Additionally, the social and political context are also very important to promote an environment that increases intergenerational mobility.

2.2. Methodology

2.2.1. Sample Construction

The intergenerational persistence in income and education variables we use are defined as 10-year averages, corresponding to each cohort's mobility/persistence (i.e., cohorts of 1960s, '70s, and '80s). Regarding income mobility, countries differ between cohorts. For education mobility, they are the same for the 1960 and 1970 cohorts, while the sample differs for the 1980 cohort: countries in the last cohort are the same as in the two earlier ones (except for New Zealand) and 36 additional countries are considered. Each generation includes different individuals, answering each country's surveys, from which data are extracted to construct GDIM (2018). Our initial panel dataset for the potential mobility predictors contains yearly information for the period 1960-2018, covering all three cohorts of GDIM. To make these two datasets compatible we average over time the potential regressors considering different time periods, which are predicted by each generation in GDIM, and have for each cohort $c = \{1960;\ 1970; 1980\}$ a cross-section of N_c countries. Our sample size for income is equal to $N = 70$ and for education we have $N = 338$, as found in Table 1.

According to Narayan *et al*. (2018), the different estimation methods for the mobility in income measure use distinct reference ages concerning individuals' permanent earnings, i.e., the proxy of their lifetime earnings. For each country the potential predictors for mobility in income will be averaged over time considering as initial point the first year of a generation and as the end point the last year of a generation plus the reference age. In this way they will account for earnings from the moment agents are born until they obtain the income reflecting their lifetime earnings. For the OLS method, the reference age is 40 years old, while for the instrumental variables methods it is 37 years old. The last reference age is not only used for the majority of the cases, i.e., the ones covering two different surveys, but also for the cases where the same sample is used for children and pseudo-parents. Adopting these ages is in line with the suggestion made by Haider and Solon (2006). According to the authors, income must be measured during the period when there is a stable relationship between lifetime income and current income, i.e., around mid-ages. Otherwise, a lifecycle bias will exist when estimating the income elasticity. This is confirmed in the work of Nybom and Stuhler (2016). We consider the upper bound in time of 2018 because it is the year for which the most recent published information in GDIM (2018) is used.

Education mobility measures are always grounded on the individuals' educational attainment. Mobility data had to be harmonized by Narayan *et al*. (2018) due to the heterogeneity of information across countries. Two specific cases appear: the one for which only co-resident data on educational attainment is available and the one for which there are retrospective data on educational attainment[7]. Co-resident data is available for some countries for only the last generation – in this scenario respondents reside with their parents and only the age group 21-25 is considered. For retrospective data there are no age restrictions. Hence, for each country the potential predictors for mobility of education will be averaged over time considering as initial point the first year of a generation and as the end point one of two cases: the last year of a generation plus 25 years when considering co-resident data; or the last year for which there is a survey (i.e., 2016), for the case in which there are retrospective data, since with no age limit a respondent can at any age attain a specific education level and be part of the sample considered in the GDIM (2018).

The summary of the periods on which potential mobility predictors are averaged over time for each country is in Table 2.

Table 2 – Time Periods for Averaged Predictors of Mobility

Cohorts	Income Mobility		Educational Mobility	
	OLS	Instrumental Variables	Co-resident Data	Retrospective Data
1960	1960-2009	1960-2006	NA	1960-2016
1970	1970-2018	1970-2016	NA	1970-2016
1980	NA		1980-2014	1980-2016

Note: NA - not applicable.

We perform the Fisher-type test (Choi, 2001) for all cohorts and time periods for which averages were calculated (based on the information in Table 2), since the averages we construct rely on the evidence that for each of

[7] Co-resident data concerning parental education has to do with information that can only be gathered through respondents co-residing in the same household as their parents. Retrospective data differs from co-resident data in the sense that information about parental education can be obtained without needing to have respondents living with their parents.

the cohorts the variables are stationary over time[8]. The null hypothesis is of a unit root for all panels, which is tested against the alternative, in which stationarity is present in at least one panel. Overall, there is evidence of stationarity for almost all predictors for income as well as for education mobility. Considering the exceptions, we calculate their growth rates: for the income predictors, job density (jobden), population density (popden), and social capital (trust), and for the education predictors we have real GDP *per capita* (GDPpc) and human capital (HK). These growth rates are named jobdeng, popdeng, trustg, GDPpcg, and HKg, respectively. Some of the variables do not have enough observations to perform the test, so we assume for the baseline model that they are stationary. Later, we measure them by the last observation in the averaged time period, and if the conclusions drawn are not sensitive to the choice of their countries' values, there is no problem in using sample averages in the baseline estimation.

2.2.2. Variables Selection Models and Techniques

Our work deals with a large number of covariates. To analyse which ones matter most in explaining mobility we use three approaches. The first is the Rigorous Least Absolute Shrinkage and Selection Operator (RLASSO) with the purpose of variable selection in a high dimensional regression model[9] and the other two are the Random Forest and the Gradient Boosting Regressors that besides variable selection also allow us to compare their accuracy.

An initial challenge we face is the existence of missing data, given that our variable selection methods are contingent on a balanced set. The variables' prevalence of missing data (% of total observations for each predictor, presented in Table B3 and B4 in the Appendix B) varies from 2.86% to 70% in the case of the income persistence and 0.3% to 76.92% for education. The relatively low proportion of missing values, where almost all variables have a prevalence smaller than 50% (the exceptions are teenbirth with 51.43% and avinc with 70% for the IGPI, and IGPI with 78.99% and avinc with 76.92% for the IGPE) and half of them are below 10%, suggests that the dataset is quite informative. It also appears that missing data does not concentrate on specific variables, with a low risk of systemic biases occurring. Therefore, the RLASSO, Random Forest and Gradient Boosting results are obtained after imputing missing data with the Miss-Forest algorithm by Stekhoven and Bühlmann (2018).

The imputation method we employ works iteratively, training a Random Forest to the values that are observed and then predicting the missing ones. It handles mixed-type data, with high dimensions, nonlinear structures, and complex interactions due to, for example, scale. It does not need tunning parameters or any assumptions about data distribution. This does not always occur with other imputation techniques such as the KNNimpute (Troyanskaya, *et al*., 2001), the MICE (Van Buuren and Oudshoorn, 1999) or the MissPALasso (Städler and Bühlmann, 2014). Stekhoven and Bühlmann (2018) compare the accuracy (measured by the out-of-bag root mean squared error) of the Miss-Forest to the previous methods, using 10 different datasets with missing data artificially introduced at a 10% to

[8] Under stationarity, the variable's (sample) average is a good proxy for its (population) expected value. If variables are not stationary, this argument no longer applies. Also, when calculating the averages for which the period of data availability is shorter than the entire periods presented in Table 3 and treating them as the expected value considering the entire time span, we are assuming that the missing values share those same properties. This causes no problems if stationarity is verified.

[9] An alternative to penalty estimation is to use information criteria for selecting the most prominent predictors to explain mobility. However, this not practical for very large number of candidate covariates which is our case where there are 2^{K-1} different possible combinations. We took the simpler route to consider all possible combinations using the 10 covariates with the smallest individual AIC and BIC. Some similarities are found regarding the benchmark results obtained using the RLASSO (Tables 4 and 5): for income, both approaches select the cohort of individuals born in the 1960s (cohort60) and parental average education (MEANp); for education, this occurs for adult literacy (litadult), government expenditures on primary education (primexp), and the migrant stock (migstock).

a 30% rate. The authors find that Miss-Forest out-performs the other methods and reduces the imputation error by 50%.

After data is imputed, we are able to apply our selection models and techniques, described below[10].

Rigorous Least Absolute Shrinkage and Selection Operator

When parametrically examining mobility predictors, we define an econometric model, which takes the general form of

$$IM_i = \beta_0 + \sum_{k=1}^{K} \beta_k IMD_{i,k} + e_i, \tag{1}$$

where, for country i, IM_i corresponds to the mobility measure we seek to explain (mobility in income or education, interchangeably) by its predictors $IMD_{i,k}$ (K variables at most) considering $i = 1, \dots, N$ cross-sections; βs are the model's coefficients and the error term is given by e_i. One may expect correlates to differ by cohort and therefore $IMD_{i,k}$ also include cohort dummy variables. Our baseline estimation will use the mobility measures defined previously and obtained by using the outcomes of the father-son pair, as in most of the literature.

The LASSO shrinkage estimator is used to select and fit the covariates that constitute the model's regressors among all the K predictors we consider. This approach is robust to multicollinearity, leads to sparse solutions, and eases interpretation, being a proper shrinkage approach when the number of regressors is too large and the usual least squares method overfits the model. The method minimizes the residual sum of squares plus a penalty term, which controls for the coefficient estimates size in absolute terms. We use a version of the LASSO – the Rigorous LASSO (RLASSO) – in which there is an optimal penalty under non-Gaussian and heteroskedastic errors and the feasible algorithms developed by Belloni *et al.* (2012). The constraint introduced to the coefficients' sizes depends on the magnitude (units of scale) of the associated variables. For this reason, the covariates are normalized to unit variances, thereby preventing some variables from "dominating" others due to scale. The final estimation results are presented in the original units/scales. The penalty term induces a bias, and the way we use to smooth it (with no loss of performance) is to apply the OLS to the predictors that were previously selected, following Belloni and Chernozhukov (2013), and then interpret the results. In this selection method we use robust heteroskedasticity/clustered standard errors.

The use of a high-dimensional dataset has the potential of pinpointing the important predictors in explaining mobility. However, there is the risk throughout the process of eliminating predictors labelled as irrelevant but which are indeed relevant. We therefore complement the RLASSO estimation with two machine learning algorithms robust to multicollinearity: the Random Forest (originally created by Breiman, 2001) and the Gradient Boosting (by Friedman, 2001). Here, the search for the covariates that predict IM is done through the use of decision-trees, which learn how to split our data into subregions of the covariates' space and are used to make predictions on mobility.

[10] We could conduct a robustness exercise where we exclude observations with missing information for at least one variable. But this leads us to having an empty dataset. We therefore rely on the obvious strengths of the Miss-Forest that make us believe that it serves well our purpose without promoting a substantial bias. This way, we consider that the existence of missing data presents a minimal impact on the integrity and interpretability of our results.

Our dataset is randomly split into training and testing data, following the literature, which commonly uses an 80:20 ratio when splitting the data (Joseph, 2022). We chose to use two supervised learning algorithms because by combining weak learners they become more robust to the risk of overfitting, which occurs when the algorithm does not generalize to new data by memorizing too closely the training set, being sensitive to outliers or training data errors.

Random Forest Regressor

Random Forests start by repeatedly selecting B times a random sample from the training dataset, with $b = 1, \dots, B$. For each sample, a tree of unknown functional form $f_b(IMD_{i,k})$ is grown. Each tree has $j = 1, \dots, J$ nodes, which are the tree splitting points. At each node N_j, the sample is split in $r_j = 1, \dots, R_j$ regions/branches, according to a specific feature and a threshold s for the observations of that feature. Each time a split is considered, not all the features are candidates to that split. Instead, a random number of $u \leq K$ predictors is chosen among the full set of K mobility predictors, to decrease the correlation between the B regression trees. For each observation i and each tree, a prediction $\widehat{f_{i,b}}$ is obtained. The final prediction for a given observation is computed as $\widehat{f_b}(\cdot) = B^{-1} \sum_{b=1}^{B} \widehat{f_{i,b}}(\cdot)$.

Gradient Boosting Regressor

Gradient Boosting also considers an ensemble of trees but, unlike Random Forest trees, which are constructed independently at each b, here they are built sequentially to correct the previous fitted trees' errors, and are able to incorporate more complex data patterns. We define $L = MSE$ (mean squared error) as the loss function to be considered throughout the process. Using the training data, the algorithm initiates the model with a constant value $\widehat{f_0} = \operatorname{argmin}_{\chi} \sum_{i=1}^{n} L(IM_i, \chi)$. Then it grows $m = 1, \dots, M$ trees, as follows. First it calculates residuals as the negative gradient of the loss function, i.e., $r_{im} = -\left[\left[\partial L\left(IM_i, \hat{f}(\cdot)\right)\right]\left[\partial \hat{f}(\cdot)\right]^{-1}\right]_{\hat{f}(\cdot)=\widehat{f_{m-1}}(\cdot)}$. Then it fits a regression tree to the values r_{im}, creating $z = 1, \dots, Z$ terminal regions, $R_{z,m}$. For each terminal region a new output value is calculated as $\theta_{z,m} = \operatorname{argmin}_{\theta} \sum_{i=1}^{n} L\left(IM_i, \widehat{f_{m-1}}(\cdot) + \theta\right), \forall i \in R_{z,m}$. Finally it updates the model with $\widehat{f_{i,m}}(\cdot) = \widehat{f_{i,m-1}}(\cdot) + \rho \sum_{z=1}^{Z_m} \theta_{z,m} \mathbb{I}(i \in R_{z,m})$, where ρ corresponds to the learning rate, i.e., the contribution of each tree to the final prediction. As a final prediction, we will have $\widehat{f_i}(\cdot) = \widehat{f_{i,M}}(\cdot)$.

Considering both algorithms, for each tree, the feature and threshold used to split a node are the ones that maximize the quality of the split. The criterion used to measure the quality of a split is the mean squared error with Friedman's (2001) improvement score. Random Forest and Gradient Boosting algorithms do not usually follow the standard regression analysis. Instead, they present plots for each predictor's (features') importance. These correspond to the normalized total reduction of the criterion that a feature is responsible for (Gini importance). The greater is the feature importance, the more responsible for impurity decrease it also is and, therefore, the more accurate the model fit and predictions will be due to that feature.

Some concerns may be raised about high variance when using tree-based methods, because in general a single tree tends to overfit the data. However, as Louppe (2014) and Probst (2019) advocate, Random Forests and Gradient Boosting algorithms have the benefit of mitigating overfitting problems. In Random Forests, both observations and features are chosen in a random way for each tree. The aggregation principle behind the final prediction, according to which individual predictions are averaged, has been found to significantly reduce the variance of each tree. Boosted trees also present an advantage by being able to select different features that allow to correct the previous tree's fitted errors. Each time a tree is grown and evaluated, a larger weight on observations that were badly predicted is attributed.

Given these arguments, we feel confident that our methodology mitigates any problem that could exist regarding high variance.

- **Calibrating the Hyperparameters of Random Forest and Gradient Boosting through Cross-validation**

Hyperparameters are the settings of the algorithm that have to be tuned. Although some research has recommended the best values for the hyperparameters, these should be set with caution. With this in mind, we initially set different values that each hyperparameter may assume. Then we make a randomized hyperparameter search, which means that 100 combinations of the ones previously set are randomly chosen and tested, finding an optimal combination. After an optimal set of hyperparameters is found, we repeat the process and try a narrowed range of combinations around the one chosen through the randomized search. In this phase there is no random selection, and instead, all the combinations are tested. The final optimal combination chosen is the one we consider. Cross-validation is the search procedure that accounts for the overfitting that may arise from the optimization process. We use the 5-fold cross validation, in which the training dataset is split into 5 folds. Each time, data is trained in 4 subsets and the validation stage is performed on the 5th fold using the R^2 as metric. The tuned hyperparameters chosen are in Table B5.

- **Out-of-sample Model Evaluation**

The accuracy obtained in the testing set is measured through the mean absolute percentage error (MAPE) and the mean squared error (MSE). These respectively given by:

$$MAPE = 100N^{-1}\sum_{i=1}^{N}|{IM_i}^{-1}(\widehat{IM}_i - IM_i)| \text{ and } MSE = N^{-1}\sum_{i=1}^{N}(\widehat{IM}_i - IM_i)^2,$$

where, for country i, IM_i corresponds to mobility in income or education (interchangeably), $\widehat{IM}_i$ regards the predictions made mobility in income or education (interchangeably), for $i = 1, \dots, N$.

Accuracy is therefore given by $100 - MAPE$ or by $1 - MSE$. Regarding income mobility using the Random Forest and Gradient Boosting algorithms, the accuracy measured using MAPE is equal to 69.08% and 66.73%, while with the MSE is equal to 0.97 and 0.96, respectively. For education these equal 76.12% and 77.27% with MAPE, and 0.98 for both with MSE. These levels of accuracy are considered quite good to explain IM.

3. Empirical Results

Results for the estimated baseline mobility RLASSO regressions are presented for income in Table 3 and for education in Table 4 (only the statistically significant predictors are reported). Figures 1 and 2 contain features importance plots according to the Random Forest and Gradient Boosting algorithms, respectively, when considering income mobility correlates. The education predictors are in Figures 3 and 4, respectively. In this high-dimensional setting we analyse the features that, in descending order of importance, accumulate 75% of the total importance, represented by the shaded grey area in the figures. In particular, we choose the ones in the shaded areas for both Random Forest and Gradient Boosting algorithms.

Table 3 – Rigorous LASSO Results for Intergenerational Persistence in Income

Dependent Variable: Intergenerational Persistence in Income (IGPI)	
LatinAmericaCaribbean	0.20**

	(0.08)
cohort60	-0.21*** (0.04)
MEANp	-0.02*** (0.01)
Obs.	70
RESET test	**0.4011**

Notes: **, *** stand for statistical significance at 5%, and 1% levels, respectively. The estimated coefficients are presented with the robust standard errors in parentheses below. The p-value of the RESET test supports the null hypothesis of no omitted variables, i.e., a well specified model.

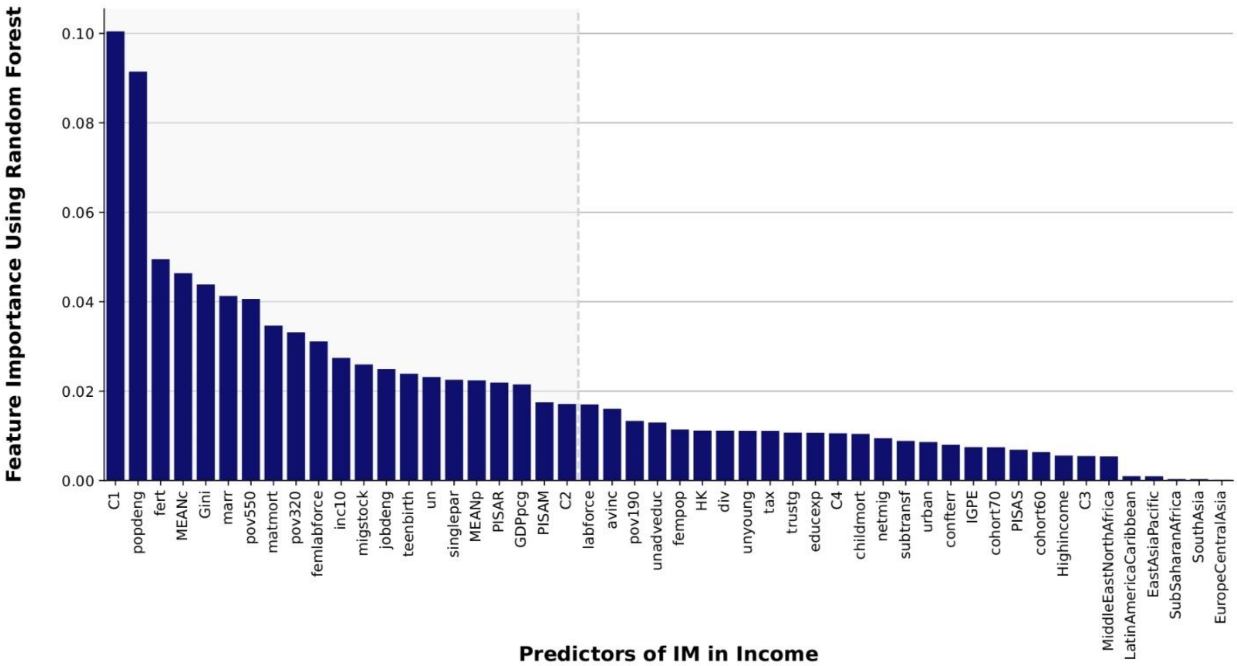
Using the Rigorous LASSO estimator for income mobility, we find that a country being part of the Latin America and Caribbean subsample (LatinAmericaCaribbean), the 1960s (cohort60), and intergenerational persistence in education (IGPE) appear to predict intergenerational persistence in income.

A country in the Latin America and Caribbean region shows more income persistence, in comparison with countries outside this region. This echoes the findings reported by Narayan *et al*. (2018) according to which relative mobility in income appears to be lower in the developing economies in comparison with high-income regions, namely in Latin America and Caribbean.

When considering individuals born in the 1960s, countries will present a higher IM compared to individuals born in the 1970s. The cohorts of 1960 and 1970 do not share any country for IM in income. In the cohort of 1960 countries are in general more developed than in the cohort of 1970, as seen in Figures C1 and C2 in Appendix C, and the persistence is greater for countries in the 1970 cohort, most of which are developing countries, enhancing results for the Latin American and Caribbean. As Narayan *et al.* (2018) stated in the World Bank document about fair progress, mobility differences depend on society preferences, which can change over time.

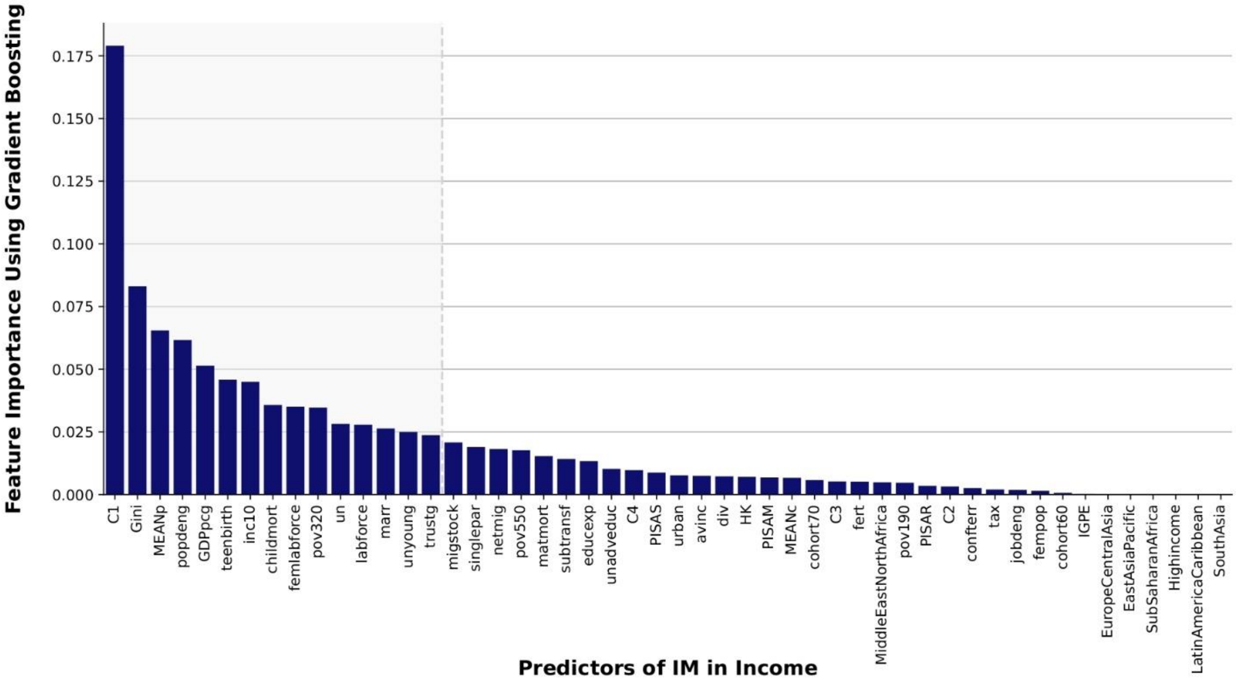
Expected evidence also emerges when analysing the relationship between intergenerational persistence in income and parental average years of education, which appear to have a negative and statistically significant relationship. This result was found in the work of Gallagher *et al.* (2019) for the USA.

Figure 1 – Feature Importance for IGPI Predictors Using the Random Forest Algorithm



Notes: C1 – children's educational attainment – less than primary level; popdeng – population density growth rate, fert – total fertility rate; MEANc – children's mean education years; Gini – Gini Index; marr – share of marriages; pov550 – shares of population living on less than $5.50 *per* day; matmort – share of women dying during pregnancy due to problems in gestation; pov320 - shares of population living on less than $3.20 *per* day; femlabforce – female labour force; inc10 – income share of the 10% richest individuals; migstock – migrants stock; jobdeng – job density growth rate; teenbirth – share of teen females who are pregnant or have had children; un – unemployment rate; singlepar – share of single parents; MEANp – parental average education; PISAR – test scores on the PISA reading scale; GDPpcg – GDP *per capita* growth rate; PISAM – test scores on the PISA mathematics scale; C2 - children's educational attainment – primary level; labforce – labour force participation rate; avinc – household disposable income; pov190 - shares of population living on less than $1.90 *per* day; unadveduc – unemployment rate for individuals with advanced education; fempop – share of female population; HK – human capital index; div – share of divorces; unyoung – youth unemployment; tax – taxes on income, profits, and capital gains; trustg – trust level growth rate; educexp – government expenditures on education as a share of GDP; C4 - children's educational attainment – upper secondary level; childmort – probability a child has of dying before the age of 5; netmig – migration movements; subtransf – subsidies and transfers; urban – degree of urbanization; confterr – deaths due to war, conflicts, and terrorism; IGPE – intergenerational persistence of education; cohort70 – cohort of the 1970; PISAS - test scores on the PISA science scale; cohort60 – cohort of the 1960; highincome – high income group of countries; C3 - children's educational attainment – lower secondary level.

Figure 2 – Feature Importance for IGPI Predictors Using the Gradient Boosting Algorithm



Notes: C1 – children's educational attainment – less than primary level; Gini – Gini Index; MEANp – parental average education; popdeng – population density growth rate, GDPpcg – GDP *per capita* growth rate; teenbirth – share of teen females who are pregnant or have had children; inc10 – income share of the 10% richest individuals; childmort – probability a child has of dying before the age of 5; femlabforce – female labour force; pov320 - shares of population living on less than $3.20 *per* day; un – unemployment rate; labforce – labour force participation rate; marr - share of marriages; unyoung – youth unemployment; trustg – trust level growth rate; migstock – migrants stock; singlepar – share of single parents; netmig – migration movements; pov550 – shares of population living on less than $5.50 *per* day; matmort – share of women dying during pregnancy due to problems in gestation; subtransf – subsidies and transfers; educexp – government expenditures on education as a share of GDP; unadveduc – unemployment rate for individuals with advanced education; C4 - children's educational attainment – upper secondary level; PISAS - test scores on the PISA science scale; urban – degree of urbanization; avinc – household disposable income; div – share of divorces; HK – human capital index; PISAM - test scores on the PISA mathematics scale; MEANc – children's mean education years; cohort70 – cohort of the 1970; C3 - children's educational attainment – lower secondary level; fert – total fertility rate; pov190 - shares of population living on less than $1.90 *per* day; PISAR – test scores on the PISA reading scale; C2 - children's educational attainment – primary level; confterr – deaths due to war, conflicts, and terrorism; tax – taxes on income, profits, and capital gains; jobdeng – job density growth rate; fempop – share of female population; cohort60 – cohort of the 1960; IGPE – intergenerational persistence of education; highincome – high income group of countries.

Considering the feature importance plots computed from the machine learning algorithms, there is clear evidence for what the significant predictors of intergenerational income persistence are, i.e., the opposite of mobility. Noticeably, none of them appear to be selected in the RLASSO estimation.

The most important variables that both algorithms identify are the share of individuals who have completed less than primary education (C1) and the average education of parents (MEANp), inequality (Gini), the growth rate of real GDP *per capita* (GDPpcg), the unemployment rate (un), the income share of the 10% richest individuals (inc10), female labour force (femlabforce), the poverty rate considering individuals living on less than $3.20 *per* day (pov320), the growth of population density (popdeng), teen birth (teenbirth), and the share of married individuals (marr). These results find theoretical grounds in Becker and Tomes (1979) and Becker *et al*. (2018) and empirical support in the works of Gallagher *et al.* (2019), Chetty *et al*. (2014a,b, 2017, 2020), Olivetti and Paserman (2015) and Chetty and Hendren (2018) regarding the USA, Corak (2019) and Lochner and Park (2022) for Canada, Deutscher

(2020) for Australia, Acciari *et al*. (2022) for Italy, Eriksen and Munk (2020) when comparing Denmark, USA, and Canada, and Deutscher and Mazumder (2020), who consider Australia and Denmark.

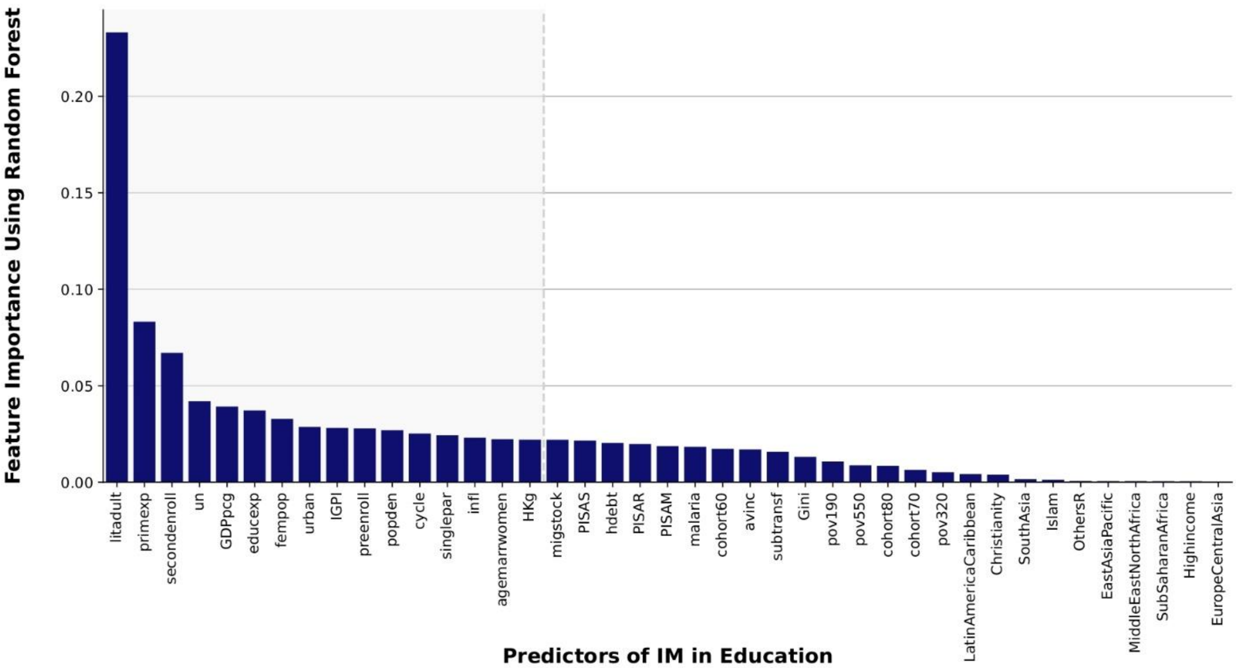
Table 4 – Rigorous LASSO Results for Intergenerational Persistence in Education

Dependent Variable: Intergenerational Persistence in Education (IGPE)	
LatinAmericaCaribbean	0.13*** (0.03)
litadult	-0.004*** (0.001)
migstock	-0.003*** (0.001)
PISAM	-0.0004** (0.0002)
primexp	-0.003* (0.001)
Intercept	0.90*** (0.10)
Obs.	338
RESET test	**0.4519**

Notes: *, **, *** stand for statistical significance at 10%, 5%, and 1% levels, respectively. The estimated coefficients are presented with the robust standard errors in parentheses below. The p-value of the RESET test supports the null hypothesis of no omitted variables, i.e., a well specified model. Results are computed using a robust estimator for the variance because there is heteroscedasticity.

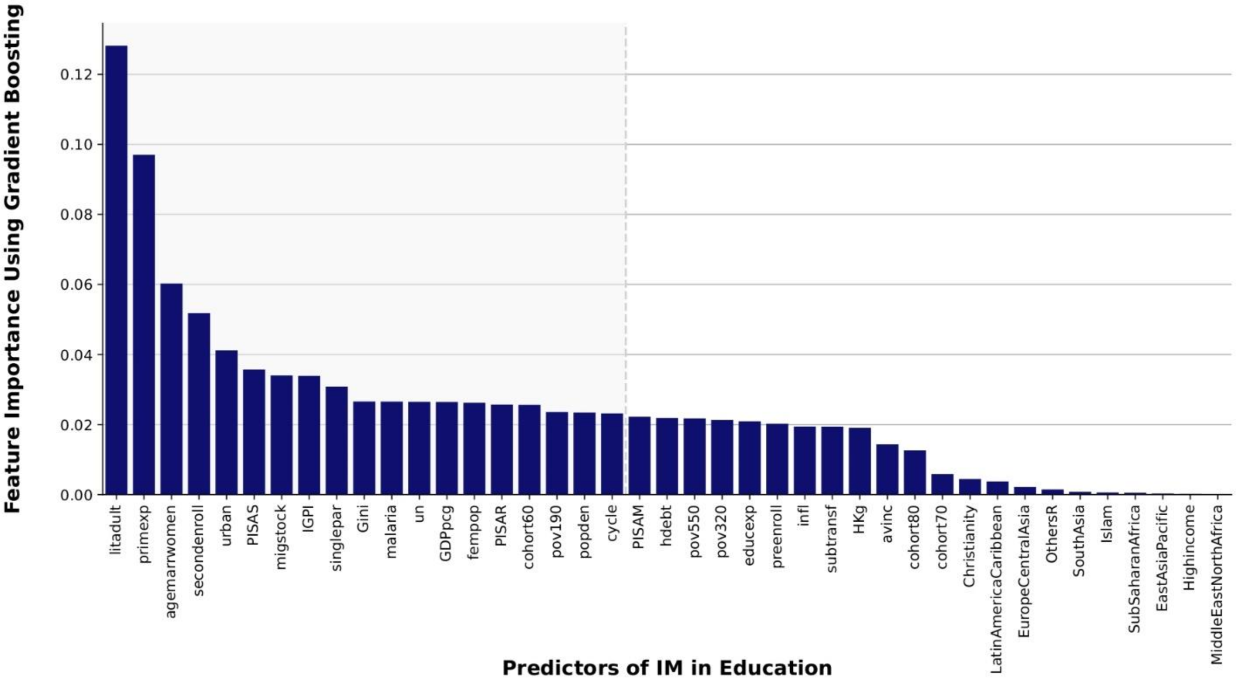
For intergenerational persistence in education, the RLASSO confirms that a country being part of the Latin America and Caribbean region has more education persistence, in comparison with countries outside this group. Narayan *et al*. (2018) also finds that the set of countries in this region present low education mobility. Adult literacy (litadult) appears to negatively predict intergenerational education persistence, i.e., it promotes mobility, as shown in the work of Alesina *et al*. (2021) for Africa. Our results for the migrant stock (migstock) are also in accordance with previous findings by Lam and Liu (2019), who suggest that mobility in Japan is higher for children of immigrants. However, the findings of Schneebaum *et al*. (2016) are mixed, presenting both negative and positive relationships between the migration stock and intergenerational mobility in education. The evidence we present for school quality (PISAM) also supports the findings reported in the literature. Hilger (2016) finds for the USA that the higher is the school quality, the higher is intergenerational mobility in education too. The expected positive prediction of government expenditures on primary education as a share of GDP (primexp) on education mobility is also verified, similar to the findings reported by Daude and Robano (2015) for Latin American, Urbina (2018) for Mexico, and Lee and Lee (2020) for OECD countries.

Figure 3 – Feature Importance for IGPE Predictors Using the Random Forest Algorithm



Notes: litadult – adult literacy; primexp – government expenditures on primary education as a share of GDP; secondenroll – gross secondary school enrolment; un – unemployment rate; GDPpcg – GDP *per capita* growth rate; educexp – government expenditures on education as a share of GDP; fempop – share of female population; urban – degree of urbanization; IGPI – intergenerational persistence of income; preenroll – gross pre-primary school enrolment; popden – population density; cycle – economic cycle; singlepar – share of single parents; infl – growth rate of the GDP deflator; agemarrwomen – average marriage age of the first marriage for women; HKg - human capital index growth rate; migstock – migrants stock; PISAS - test scores on the PISA science scale; hdebt – household debt; PISAR – test scores on the PISA reading scale; PISAM - test scores on the PISA mathematics scale; malaria – malaria incidence; cohort60 – cohort of the 1960; avinc – household disposable income; subtransf – subsidies and transfers; Gini – Gini Index; pov190 - shares of population living on less than \$1.90 *per* day; pov550 – shares of population living on less than \$5.50 *per* day; cohort80 – cohort of the 1980; cohort70 – cohort of the 1970; pov320 - shares of population living on less than \$3.20 *per* day; highincome – high income group of countries.

Figure 4 – Feature Importance for IGPE Predictors Using the Gradient Boosting Algorithm



Notes: litadult – adult literacy; primexp – government expenditures on primary education as a share of GDP; agemarrwomen – average marriage age of the first marriage for women; secondenroll – gross secondary school enrolment; urban – degree of urbanization; PISAS - test scores on the PISA science scale; migstock – migrants stock; IGPI – intergenerational persistence of income; singlepar – share of single parents; Gini – Gini Index; malaria – malaria incidence; un – unemployment rate; GDPpcg – GDP *per capita* growth rate; fempop – share of female population; PISAR – test scores on the PISA reading scale; cohort60 – cohort of the 1960; pov190 - shares of population living on less than $1.90 *per* day; popden – population density; cycle – economic cycle; PISAM - test scores on the PISA mathematics scale; hdebt – household debt; pov550 – shares of population living on less than $5.50 *per* day; pov320 - shares of population living on less than $3.20 *per* day; educexp – government expenditures on education as a share of GDP; preenroll – gross pre-primary school enrolment; infl – growth rate of the GDP deflator; subtransf – subsidies and transfers; HKg - human capital index growth rate; avinc – household disposable income; cohort80 – cohort of the 1980; cohort70 – cohort of the 1970; highincome – high income group of countries.

Both machine learning algorithms share the most important intergenerational persistence predictors in education. Interestingly, the two most important variables are also found to be statistically significant in the RLASSO estimation. These are the adult literacy (litadult) and the government expenditures on primary education (primexp).

The other most important features are the unemployment rate (un), the growth rate of real GDP *per capita* (GDPpcg), the share of population which is female (fempop), the degree of urbanization (urban), intergenerational persistence in income (IGPI), the economic cycle (cycle), population density (popden), the share of single parents (singlepar), and marriage age for women (agemarrwomen). This is in line with the findings of Alesina *et al*. (2021, 2023) regarding African countries, Hilger (2016) about the USA, Lee and Lee (2020) for the OECD, Schneebaum *et al*. (2016) for Austria, Daude and Robano (2015) for Latin America, Urbina (2018) regarding Mexico, and Neidhöfer and Stockhausen (2018) for the USA.

Finally, we replace the (time) sample averages by the last period observed as defined in Table 2, for those predictors that were assumed to be stationary given that we did not have enough observations to perform the unit root tests. For the RLASSO regression, a variable is a robust predictor of mobility if it continues to be statistically significant, and analogously, for the Random Forest and Gradient Boosting, if it continues to belong to the group of common variables found by both machine learning algorithms (considering the ones which, in descending order of importance, account for 75% of the total importance). The same reasoning is applied if, even after making the substitutions in the other variables, a variable that is not replaced continues to belong to the variables selected by RLASSO or to the group of common variables considered by the machine learning algorithms. This means that the conclusions drawn are robust and not sensitive to the choice of which of the two measures used: period averages or last year period. Table 5 summarizes the set of robust predictors.

Table 5 – Summary of Robust Predictors of Intergenerational Persistence

Methods	Intergenerational Persistence in Income	Intergenerational Persistence in Education
RLASSO	- Being a country in the Latin America and Caribbean region, *LatinAmericaCaribbean* (+) - Cohort of individuals born in the 1960s, *cohort60* (-) - Parental average education, *MEANp* (-)	- Being a country in the Latin America and Caribbean region, *LatinAmericaCaribbean* (+) - Adult literacy, *litadult* (-) - Migrant stock, *migstock* (-)
Machine Learning	- Share of children who have completed less than primary education, *C1* - Gini index, *Gini* - Unemployment rate, *un* - Share of people living on less than $3.20 *per* day, *pov320* - Growth rate of population density, *popdeng* - Share of married individuals, *marr*	- Adult literacy, *litadult* - Government expenditures on primary education, *primexp* - Growth rate of real GDP *per capita*, *GDPpcg* - Share of female population, *fempop* - Degree of urbanization, *urban* - Intergenerational Persistence in Income, *IGPI*

Notes: Impacts of robust predictors on IM are in parentheses for the RLASSO. Variables' acronyms are in italics.

Results show that according to the RLASSO estimator, being a country in the Latin America and Caribbean region (LatinAmericaCaribbean) appears to matter for intergenerational persistence in income and education. For income, the cohort of individuals born in the 1960s (cohort60) show more mobility in comparison with the cohort of individuals born in the 1970s. This also occurs regarding the parental average education (MEANp). For education, the migrant stock (migstock) appears to negatively predict intergenerational persistence. This is relevant, which seems to point to the fact that migrants have higher relative mobility than national citizens relative to their generation and in comparison to their fathers' generation. If migrants come from less developed countries than the one that they migrate to, this is only natural, since their sons now have access to a better educational system. Considering the Random Forest and Gradient Boosting algorithms, we found that the set of robust predictors of income persistence are the share of children who have completed less than primary education (C1), the Gini index (Gini), the unemployment rate (un), the share of people living on less than $3.20 *per* day (pov320), the growth rate of population density (popdeng), and the share of married individuals (marr). For education we have the government expenditures on primary education (primexp), the growth rate of real GDP *per capita* (GDPpcg), the share of female population (fempop), the degree of urbanization (urban), and the Intergenerational Persistence in Income (IGPI). Adult literacy (litadult) is the variable that is selected by all methodologies, when considering persistence in education.

- **Sensitivity Analysis to the Inclusion of Different Methods to Estimate IGPI**

In Section 2.1.1. Intergenerational Mobility Measures, we described how the relative mobility measures we use are constructed. Specifically, the IGPI can be estimated directly by OLS or by an instrumental variable estimator. Now, we conduct a sensitivity exercise, using the RLASSO and the Random Forest and Gradient Boosting algorithms with the hyperparameters reoptimized (see Table B6 in the Appendix) to check whether our results change when considering different estimation methods to compute income mobility. This is because, as pointed by Olivetti and

Paserman (2015) and Bloise *et al*. (2021), the two different methodologies may not be entirely comparable. With this purpose we consider only observations that use an IV-type estimator. Results are presented in Table 6.

Table 6 – Sensitivity Analysis for IGPI Estimation Methods

Methods	Benchmark	Without OLS
RLASSO	- Being a country in the Latin America and Caribbean region, *LatinAmericaCaribbean* (+) - Cohort of individuals born in the 1960s, *cohort60* (-) - Parental average education, *MEANp* (-)	- Cohort of individuals born in the 1960s, *cohort60* (-)
Machine Learning	- Share of children who have completed less than primary education, *C1* - Gini index, *Gini* - Unemployment rate, *un* - Share of people living on less than \$3.20 *per* day, *pov320* - Growth rate of population density, *popdeng* - Share of married individuals, *marr*	- Share of children who have completed less than primary education, *C1* - Share of children who have completed less than primary education, *C2* - Gini index, *Gini* - Income share of the 10% richest individuals, *inc10*

Notes: Impacts of robust predictors on IM are in parentheses for the RLASSO. Variables acronyms are in italics.

Results change when removing OLS estimates of mobility. For the RLASSO, only the cohort of individuals born in the 1960s (cohort60) is now significant. Regarding the Machine Learning algorithms, the ones that remain robust in comparison with the benchmark sample are the share of children completing less than primary education (C1) and the Gini index (Gini). Adding to these, the new robust variables are the share of children completing primary education (C2) and the income share of the 10% richest individuals.

Although results are different, we should note that this is done at the expense of the significant drop of around 13% observations. Therefore, we acknowledge that results may be predicted by the method that is used to compute mobility estimates and continue using the benchmark sample.

4. The Contribution of Each Predictor for Individual Predictions Using Shapley Values

The previous feature importance plots computed from the Random Forest and Gradient Boosting Algorithms reflects only the contribution of each feature for the model's fit, with no information about the direction of any possible relationship. Considering that feature importance may change in different ranges of the covariates' subspaces, we use a novel approach in machine learning – the computation of local features' importance. These give us the features' contributions for each model prediction and promotes the understanding of the possible relationships being modelled. We use the Shapley Additive Explanations Method (SHAP) by Lundberg and Lee (2017), an algorithm grounded on the work about cooperative game theory of Shapley (1953) and also developed by Shorrocks (2013), applied to decomposition analysis.

This approach was originally created to compute the expected marginal contribution of a player for the outcome of a game, given all the possible coalitions that player can join, i.e., the Shapley value. In a cooperative game, the Shapley value for player l is given by:

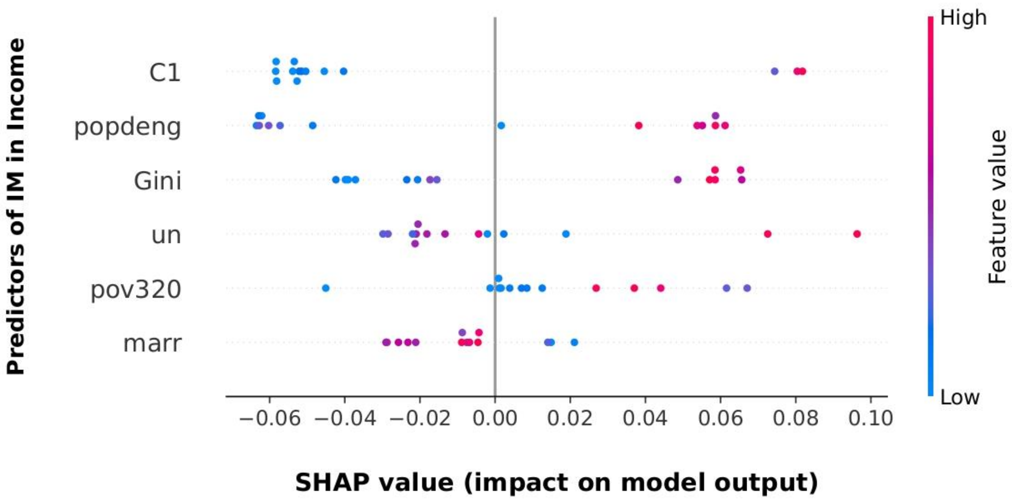
$$\phi_l(g) = \sum_{P \subseteq \{1,\dots,x\}\setminus\{l\}} \frac{|P|!\,(n - |P| - 1)!}{x!} [g(P \cup \{l\}) - g(l)] \tag{2}$$

where l is the total number of players, P considers the set of coalitions to which player l can make a marginal contribution, g is the function to obtain the outcome of the game. The same reasoning can be applied to our context where players become features.

The tree-based machine learning models we use are the Random Forest and Gradient Boosting algorithms applied to the variables presented in Table 5. Hyperparameters are again optimized (see Table B7 in Appendix B), training the models to be used in the computation of Shapley values for the testing dataset. When considering the mean absolute percentage error, this is done with improved accuracy regarding income persistence: accuracy in the Random Forest and Gradient Boosting algorithms is equal to 72.23% and 75.46%, respectively. For the education persistence model, these are similar: 74.36% and 76.16%, respectively. When accuracy is measured through the mean squared error, the corresponding values are roughly the same, being equal to 0.97, 0.96, 0.98 and 0.98, respectively. Figures 5 and 6 present the bee-swarm plots when considering income persistence and the Random Forest and the Gradient Boosting algorithms, respectively. The same is done for education persistence in Figures 7 and 8. Our interpretation again relies on the evidence considering both algorithms.

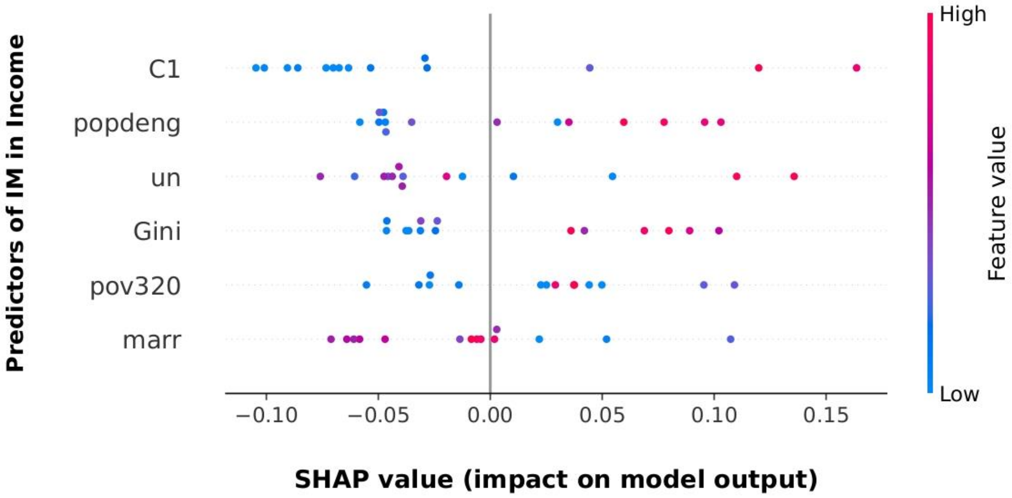
In each plot and for each variable, each dot represents an observation (country) in our testing set. Observations are placed horizontally according to the Shapley value associated with a given feature. If the frequency of Shapley values is high, the dots appear vertically. Moreover, each dot presents a different colour according to the bar scale at the right of the graph, depending on whether we have a low or high variable's raw value (not the Shapley value but the predictor's).

Figure 5 – Feature Contribution for Income Persistence Prediction Using the Random Forest Algorithm



Notes: C1 – children's educational attainment – less than primary level; popdeng – population density growth rate; Gini – Gini Index; un – unemployment rate; pov320 - shares of population living on less than \$3.20 *per* day; marr – share of marriages.

Figure 6 – Feature Contribution for Income Persistence Prediction Using the Gradient Boosting Algorithm



Notes: C1 – children's educational attainment – less than primary level; popdeng – population density growth rate; un – unemployment rate; Gini – Gini Index; pov320 - shares of population living on less than $3.20 *per* day; marr – share of marriages.

Overall, results show that lower values of the share of children who have completed less than primary education as the maximum education attainment (C1), the growth rate of population density (popdeng), and inequality (Gini) led to lower persistence in income predictions (higher mobility), while larger values promoted a higher predicted persistence (lower mobility). Although we should be careful about interpreting these findings, one may consider the possible inverse relationship between mobility and these variables, according to which we expect their increase to make IM in income lower.

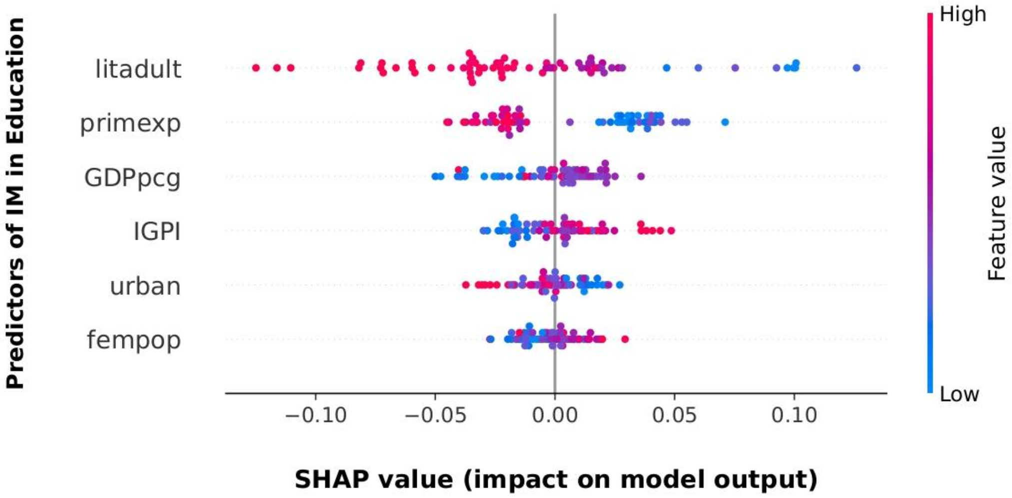
Results for educational attainment present ambiguous results in the work of Daruich and Kozlowski (2020), which contrasts with our findings, i.e., the share of children who have completed less than primary education as the maximum education attainment (C1) promotes higher predicted persistence (lower mobility).

The negative relationship between intergenerational income mobility and income inequality resembles the relationship known as the Great Gatsby curve and is confirmed by existing findings. Chetty *et al*. (2014a) present evidence that for individuals born in 1980-1982 across USA geographies, when inequality is high, mobility will be low. Chetty *et al.* (2014b) also study the USA considering the 1971-1993 birth cohorts and find IM to be stable throughout time, although by predicting the behaviour between persistence and middle-class inequality, results suggest that there is a positive relationship between both. Olivetti and Paserman (2015) find, for the USA between 1850 and 1940, that an increase in income inequality is one of the predictors of the decrease of IM between 1900 and 1920. Chetty *et al.* (2017) estimate mobility for children born in the period 1940-1984 in the USA. Results show a fall in upward mobility, with the highest decrease for middle-class families, due to greater inequality in the income distribution of the 1980s relative to the 1940s. Chetty and Hendren (2018) conclude that inequality correlates negatively with mobility in USA counties for the individuals born in 1980-1986. For Corak (2019), mobility is higher in Canada, where there is lower income inequality (the association is stronger for the lower half of the income distribution) in individuals born between 1963 and 1970. Lochner and Park (2022) suggest that there is also a positive relationship between intergenerational persistence with the variability of parental earnings across cities (in other words, a negative relationship between mobility and earnings inequality) in Canada, for the period 1978-2014. Acciari *et al*. (2022) show that for the 1942-2014 period the relationship between income inequality and individual's economic mobility (relative to their parents), has a negative slope despite mobility, in Italy's regions.

The opposite occurred for the share of married individuals (marr), which we expect to improve mobility in income predictions (persistence will be lower). In Eriksen and Munk (2020), for Denmark, in the period between 1980 and 2015, the result reported is that the share of married inhabitants relates in a positive way with mobility as well.

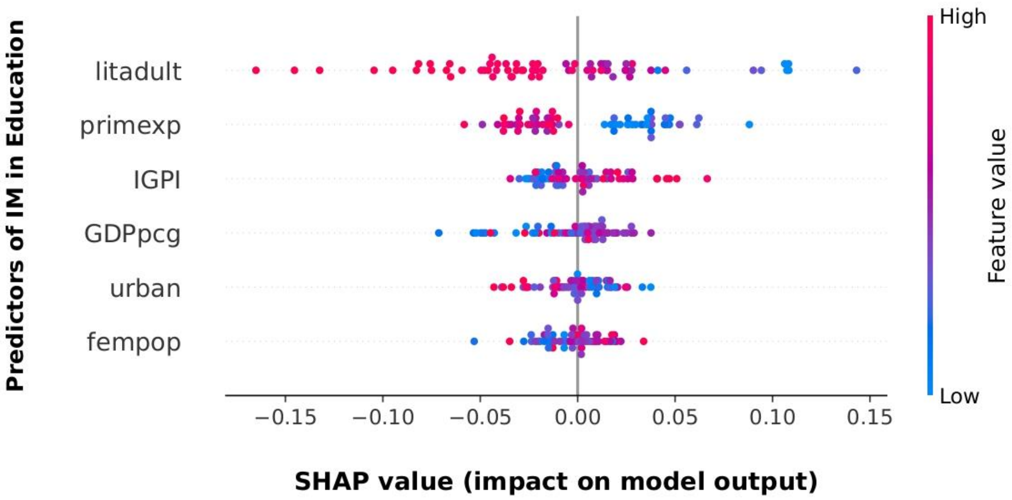
The previous conclusions may be drawn since there is a clear distinction between low and high values of the variables (blue and red, respectively). Our evidence for the poverty rate considering individuals living on less than \$3.20 *per* day (pov320) and the unemployment rate are not clear, because the dots for high and low values of these variables overlap in the negative or positive region of the Shapley Values.

Figure 7 – Feature Contribution for Education Persistence Prediction Using the Random Forest Algorithm



Notes: litadult – adult literacy; primexp – government expenditures on primary education as a share of GDP; GDPpcg – GDP *per capita* growth rate; IGPI – intergenerational persistence of income; urban – degree of urbanization; fempop – share of female population.

Figure 8 – Feature Contribution for Education Persistence Prediction Using the Gradient Boosting Algorithm



Notes: litadult – adult literacy; primexp – government expenditures on primary education as a share of GDP; IGPI – intergenerational persistence of income; GDPpcg – GDP *per capita* growth rate; urban – degree of urbanization; fempop – share of female population.

Regarding education, we have the higher values of adult literacy (litadult) and government expenditures on primary education as a share of GDP (primexp) contributing in a negative way for predictions of persistence (higher mobility). Lower values of these variables result in higher persistence predictions (lower mobility). A positive relationship between these variables and mobility is therefore expected.

Parental literacy appears to be positively correlated with mobility in Africa, from the 1960s on, in Alesina *et al*. (2021). For primary education expenditures we have Daude and Robano's (2015) finding that high-mobility countries present high progressive public investments in education, considering 18 Latin American countries for the year 2008. In Urbina (2018), evidence shows that mobility increased in this country for primary school completion (with significant increases for low and middle educational backgrounds) as well as for secondary school completion (due to improvement of the individuals' middle educational background), decreasing for some postsecondary education (with differences in the patterns by gender) for Mexican individuals born in 1947-1986. The "11-year plan" is a federal government policy having the goal of increasing primary and lower secondary enrolment. It is linked to the increase in mobility in those levels, and to the decrease for the higher ones by creating a bottleneck between lower and upper secondary education. Lee and Lee (2020) found that public expenditure on primary school, compared to expenditure on tertiary education, may improve mobility, considering OECD countries and 1947-1990 birth cohorts. Government expenditures on primary education as a share of GDP (primexp), which contributes in a positive way to intergenerational mobility in education, is related with the share of children who have completed less than primary education as the maximum education attainment (C1), a variable associated with a higher intergenerational mobility for income. Hence, it seems that the early stages of education and child development are essential for increasing both intergenerational mobility in income and education. Several organizations have stressed the relevance of the early years. For example, the UNICEF states in its website: "Science shows that life is a story for which the beginning sets the tone. That makes the early years of childhood a time of great opportunity, but also great risk." and "In the first few years of life, more than one million neural connections are formed each second – a pace never repeated again.". Additionally, the Center for the Developing Child at Harvard University asserts the following in its website: "Healthy development in the early years (particularly birth to three) provides the building blocks for educational achievement, economic productivity, responsible citizenship, lifelong health, strong communities, and successful parenting of the next generation." Hence, covariates related with early childhood development are of extreme relevance to intergenerational mobility.

Results are not clear for the intergenerational persistence in income (IGPI), the growth rate of real GDP *per capita* (GDPpcg), the degree of urbanization (urban), and the female population share (fempop). This is because low and high values of these variables overlap in both the negative or positive regions of the Shapley Values, unlike what occurs with the variables for which positive and negative connections with persistence are drawn.

5. Predicting Income Mobility

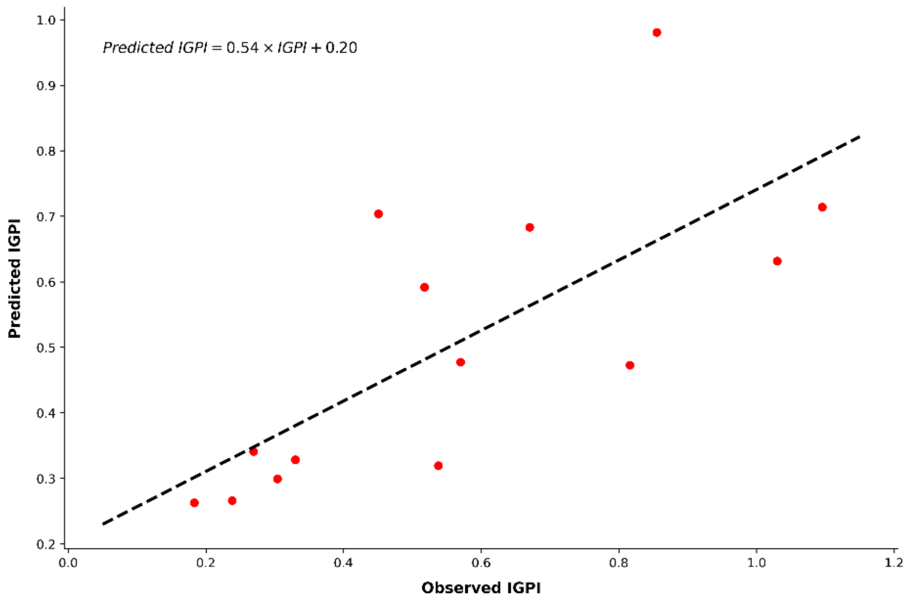
Narayan *et al*. (2018) point out that education mobility is important in its own right and should be an important element of income mobility: incomes persist due to the inherited endowments received from parents and to the investments parents make in children's education. A positive connection may be expected between these two dimensions. Having income mobility estimates for all the countries for which there are education mobility observed values is therefore important in the context of our study.

In the Global Database on Intergenerational Mobility (GDIM), only 70 countries present intergenerational income mobility values, below the 137 for which intergenerational educational mobility observations are available. Also, all countries' estimates of income mobility have a corresponding estimate of intergenerational education mobility. In Section 3 we found the set of robust predictors of intergenerational income mobility using the RLASSO as well as the Machine Learning algorithms. This means that we are able to predict the income mobility for those countries that for income do not belong to the GDIM and obtain a balanced dataset of income and education mobility measures. With this purpose, data are again pooled and we use the Gradient Boosting algorithm, which, as previously mentioned, is the one with the highest out-of-sample accuracy measured through the MAPE (75.46% vs 72.23% of the Random Forest). We also conclude that accuracy computed using the MSE is around the same values in both algorithms. Besides, we have a full-sample fit of 71.91% in the RLASSO.

Most countries in the entire dataset present intergenerational persistence in education estimates in subsequent cohorts. We will thus end up with different income mobility predictions for each country, depending on the cohort on which the IGPE is measured. This will allow us to compare predictions to the true values of income mobility, which are available by cohort. A joint analysis of income mobility predictions and education mobility observed values is also possible. To obtain the income mobility predictors averaged over time for each country, we consider the largest time period defined for each existing cohort, which regard the OLS estimator: 1960-2009 for the 1960 cohort and 1970-2018 for the 1970 cohort. Finally, we also compute income mobility predictions for the 1980 cohort: the period considered to compute the predictors' averages is 1980-2018.

Figure 9 plots the observed intergenerational persistence in income values against the predicted ones considering the testing set.

Figure 9 – Predicted Intergenerational Mobility of Income (IGPI) vs Observed IGPI in the Testing Set

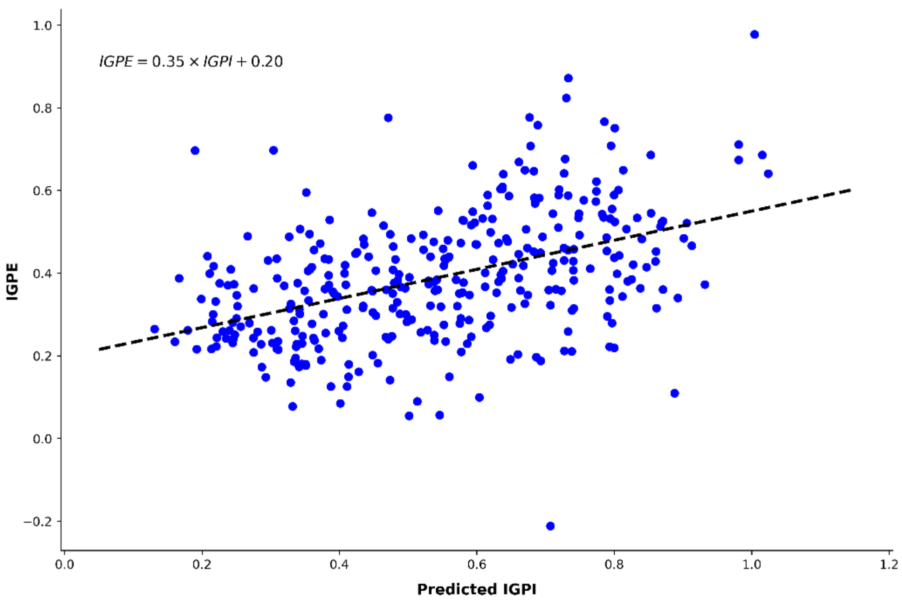


The relationship between the out-of-sample observed and predicted values of income persistence is positive with a correlation coefficient around 0.73. If we had considered the full training and testing sets, the relationship between observed and predicted income persistence is stronger, close to the 45º line, with a correlation coefficient around 0.90.

We now present the point predictions obtained for all the countries and cohorts for which IM in education was available from the GDIM. Figures C3, C4 and C5 in Appendix C show the absolute frequency of countries in a set of intervals of income persistence predictions. The corresponding countries are listed as well. The results show that high-income economies are the ones presenting the lowest values of predicted income persistence, i.e., highest values of predicted income mobility, a result which is consistent across cohorts. This was previously verified in the RLASSO estimation results for income mobility. Considering that the predicted income persistence mean value for the 1960 and 1970 cohorts is approximately 0.53, while for the 1980 cohort it is around 0.57, we have high-income economies comprehending always less than 10% of the countries above those values. Specifically, from the 35 high income economies, only 5 and 2 are above the mean for the 1960 and 1970 cohorts, respectively; while this is verified for 3 out of 34 high-income countries in the 1980 cohort. China, Guinea, Bulgaria, and Nepal are the only ones starting below the 1960 cohort average but ending up above the 1970 cohort mean. This occurs for Albania, Russia, Slovak Republic, Armenia, Greece, and Montenegro between the 1970 and 1980 generations.

Finally, we plot all the observed educational persistence values against the predicted income persistence values in our sample in Figure 10.

Figure 10 – The Relationship Between Intergenerational Mobility of Education (IGPE) and Predicted IGPI



The estimated slope is 0.35, a value very close to the one estimated by the LASSO approach in the baseline model (0.29). Although income-education persistence estimates are positively connected, their relationship is relatively modest, with the correlation coefficient being approximately equal to 0.45[11]. Since predicted income mobility is lower in developing economies and it appears to have a positive relationship with education mobility, this

[11] We performed the same exercise using the observed income mobility values from the GDIM and the predicted values for the countries with no information on income mobility. Conclusions are about the same.

means that the latter should also be compromised in the developing countries when compared to high-income economies.

Since we found in the baseline model that income inequality and the share of individuals with less than primary education are responsible for making incomes persist throughout generations, public policies aimed at reducing inequality and improving educational attainment of populations are of utmost importance. The same should be considered regarding improvements in government expenditures on primary education as a share of GDP and adult literacy, which are found to positively predict education mobility. Our evidence is corroborated by Narayan *et al*. (2018), according to which both income and education mobility are expected to be lower in the developing world. The authors point out that developing economies are the ones presenting the highest levels of inequality as well as the highest shares of individuals with a low education level. When correlating mobility in education and public spending on education for developing economies, a stronger association is found for the primary education level in comparison with the other levels. In addition, the World Bank (2018b) shows that the low-income countries' average student performs more poorly than 95% of high-income economies' average students, when considering literacy and numeracy assessments.

6. Conclusion

In this work we have assessed the predictors of income and education IM at a world-wide level. Literature about the predictors of intergenerational income and education mobility has been mainly country and period specific. Our analysis uses the recent database GDIM, which provides indicators and elasticities for both income and intergenerational education persistence for 137 developing and developed countries and considers the period from 1960 to 2018. We use the Rigorous Least Absolute Shrinkage and Selection Operator (RLASSO) as well as the Random Forest and Gradient Boosting algorithms to perform our analysis, avoiding the consequences of an *ad-hoc* model selection, particularly in a high dimensionality context such as the one we present. Since the two algorithms present only the variables' importance, we use Shapley values to obtain the expected relationship of mobility and its predictors. Finally, we predict mobility values for the countries for which only observed values for intergenerational education mobility are available, using the predictors of income mobility found. Grounded on our findings we propose policy measures aimed to improve IM, as follows.

Results show that income mobility is negatively predicted by the share of individuals that have completed less than primary education, while education mobility presents a positive relationship with adult literacy. Implementing strategies to promote human capital is considered to be essential as they should be translated in low-income individuals benefiting from cognitive and noncognitive skills that influence their returns in the labour market and improve income mobility. Inequality is also a driver of IM in income, predicting it in a negative way, resembling the popular Great Gatsby curve, according to which countries with greater inequality promote increases in income persistence. Narayan *et al*. (2018) consider the improvement and access to capital markets as a way to possibly mitigate inequality effects on mobility. Poor individuals will be able to invest with fewer constraints by borrowing to finance their children's education. Also, the likelihood that only individuals with inherited wealth have the opportunity of financing investments that are rewarded in the labour market should be lower. Finally, improving public spending on education will help to narrow the gap in private investments between offspring of rich and offspring of poor parents and thereby reduce the impact that parents have on children's outcomes. Specifically, we found government expenditure on primary education as a strong predictor of education mobility, confirming the argument that spending can produce stronger impacts when focused on early childhood (Herrington, 2015). Since parental average education is found to promote income mobility of the younger generation, we may think that policies targeting educational attainment today will have a feedback effect on future generations mobility outcomes as well.

These strategies are especially important for developing countries, reinforcing the conclusion that these countries are more penalized in terms of increasing income persistence through the significance of the 1960s cohort variable, but also in terms of education mobility, which appears to have a positive relationship with predicted income persistence. By improving income mobility and educational mobility, policy makers are promoting a feedback effect for future generations. Implementing all of these measures together is possible only with strong and sustained economic growth, meaning that their predictors should also be promoted.

We should note some of the limitations of our work, namely those concerning the dataset used. GDIM (2018) comprehends most countries in the World, but the existence of several estimation methods for mobility measures may bias the results. This means that differences in the evidence obtained may not be related with the predictors, countries, or cohorts used, but with the methodology adopted by the World Bank when constructing the database. The obstacle to include women in the analysis also relies on the fact that data and estimates about women are not reliable. If reliable estimates and data for women would surge this would be very useful, especially for the study of gender inequality in countries in which significant disparities exist. An alternative way to the GDIM database to study intergenerational mobility in income and in education is by means of a Meta analysis based on a collection of papers about the topic. Another limitation is that we can't examine the dynamic impact of a given predictor on mobility, since most the country specific predictors we use undergo gradual changes over time. Future research should consider undertaking the same analysis but with higher-frequency data. That is, instead of using 10-year averages for each cohort, use smaller intervals when data availability makes it possible. This would also allow a panel type analysis to complement our cross-sectional framework. There are factors as Hofstede's cultural dimensions that should predict mobility and could be explored at the time the related data becomes available. Finally, it would be interesting to consider multivariate splitting in Random Forests and Gradient Boosting where the interaction effects between variable pairs play a role at the same split.

Data availability statement

The data supporting the findings of this study are available upon request.

References

[1] Acciari, P., Polo, A., & Violante, G. (2022). "And Yet It Moves": Intergenerational Mobility in Italy. *American Economic Journal: Applied Economics*, 14(3), 118-163. https://doi.org/10.1257/app.20210151

[2] Alesina, A., Hohmann, S., Michalopoulos, S., & Papaioannou, E. (2021). Intergenerational Mobility in Africa. *Econometrica*, 89(1), 1-35. https://doi.org/10.3982/ECTA17018

[3] Alesina, A., Hohmann, S., Michalopoulos, S., & Papaioannou, E. (2023). Religion and Educational Mobility in Africa. *Nature*. https://doi.org/10.1038/s41586-023-06051-2

[4] Becker, G. & Tomes, N. (1979). An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility. *Journal of Political Economy*, 87(6), 1153-1189.

[5] Becker, G., Kominers, S., Murphy, K., & Spenkuch, J. (2018). A Theory of Intergenerational Mobility. *Journal of Political Economy*, 126(S1), S7-S25.

[6] Belloni, A. & Chernozhukov, V. (2013). Least Squares After Model Selection in High-Dimensional Sparse Models. *Bernoulli*, 19(2), 521-547. http://doi.org/10.3150/11-BEJ410

[7] Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica*, 80(6), 2369-2429. https://doi.org/10.3982/ECTA9626

[8] Bloise, F., Brunori, P., & Piraino, P. (2021). Estimating intergenerational income mobility on sub-optimal data: a machine learning approach. *The Journal of Economic Inequality*, 19(4), 643-665.

[9] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32. http://doi.org/10.1023/A:1010933404324

[10] Brunori, P., Hufe, P. & Mahler, D. (2023). The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests. *The Scandinavian Journal of Economics* (forthcoming). https://doi.org/10.1111/sjoe.12530

[11] Cervini-Plá, M. (2015). Intergenerational Earnings and Income Mobility in Spain. *Review of Income and Wealth*, 61(4), 812-828. https://doi.org/10.1111/roiw.12130

[12] Chetty, R. & Hendren, N. (2018). The Impacts of Neighborhoods on Intergenerational Mobility II: County-level Estimates. *The Quarterly Journal of Economics*, 133(3), 1163-1228. https://doi.org/10.1093/qje/qjy006

[13] Chetty, R., Friedman, J., Hendren, N., Jones, M., & Porter, S. (2020). The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility. *National Bureau of Economic Research Working Papers No. 25147*. http://doi.org/10.3386/w25147

[14] Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., & Narang, J. (2017). The Fading American Dream: Trends in Absolute Income Mobility Since 1940. *Science*, 356(6336), 398-406. https://doi.org/10.1126/science.aal4617

[15] Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014a). Where is The Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, 129(4), 1553-1623. https://doi.org/10.1093/qje/qju022

[16] Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014b). Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility. *American Economic Review: Papers & Proceedings*, 104(5), 141-147. http://doi.org/10.1257/aer.104.5.141

[17] Choi, I. (2001). Unit Root Tests for Panel Data. *Journal of International Money and Finance*, 20(2) 249-272. https://doi.org/10.1016/S0261-5606(00)00048-6

[18] Corak, M. (2019). The Canadian Geography of Intergenerational Income Mobility. *The Economic Journal*, 130(631), 2134-2174. https://doi.org/10.1093/ej/uez019

[19] Daruich, D. & Kozlowski, J. (2020). Explaining Intergenerational Mobility: The Role of Fertility and Family Transfers. *Review of Economic Dynamics*, 36, 220-245. https://doi.org/10.1016/j.red.2019.10.002

[20] Daude, C. & Robano, V. (2015). On Intergenerational (Im)mobility in Latin America. *Latin American Economic Review*, 24(9), 1-29. https://doi.org/10.1007/s40503-015-0030-x

[21] Deutscher, N. & Mazumder, B. (2020). Intergenerational Mobility across Australia and the Stability of Regional Estimates. *Labor Economics*, 66. https://doi.org/10.1016/j.labeco.2020.101861

[22] Deutscher, N. & Mazumder, B. (2023). Measuring Intergenerational Income Mobility: A Synthesis of Approaches. *Journal of Economic Literature*, 61(3), 988-1036. http://doi.org/10.1257/jel.20211413

[23] Deutscher, N. (2020). Place, Peers, and the Teenage Years: Long-Run Neighborhood Effects in Australia. *American Economic Journal: Applied Economics*, 12(2), 220-249. https://doi.org/10.1257/app.20180329

[24] Eriksen, J. & Munk, M. (2020). The Geography of Intergenerational Mobility – Danish Evidence. *Economic Letters*, 189. https://doi.org/10.1016/j.econlet.2020.109024

[25] Ermisch, J., Francesconi, M., & Siedler, T. (2006). Intergenerational Economic Mobility and Marital Sorting. *Economic Journal*, 116(513), 659-679. https://doi.org/10.1111/j.1468-0297.2006.01105.x

[26] Feenstra, R., Inklaar R., & Timmer M. (2015). The Next Generation of the Penn World Table. *American Economic Review*, 105(10), 3150-3182. https://doi.org/10.1257/aer.20130954

[27] Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[28] Gallagher, R., Kaestner, R., & Persky, J. (2019). The Geography of Family Differences and Intergenerational Mobility. *Journal of Economic Geography*, 19(3), 589-618. https://doi.org/10.1093/jeg/lby026

[29] GDIM (2018). *Global Database on Intergenerational Mobility*. Development Research Group, World Bank. Washington, D.C.: World Bank Group.

[30] Global Change Data Lab (2021). *Our World in Data Project*.

[31] Haider, S., & Solon, G. (2006). Life-cycle Variation in the Association Between Current and Lifetime Earnings. *American Economic Review*, 96(4), 1308-1320. https://doi.org/10.1257/aer.96.4.1308

[32] Helsø, A. (2020). Intergenerational Income Mobility in Denmark and the United States. *The Scandinavian Journal of Economics*, 1-24. https://doi.org/10.1111/sjoe.12420

[33] Herrington, C. (2015). Public Education Financing, Earnings Inequality, and Intergenerational Mobility. *Review of Economic Dynamics*, 18(4), 822-842. https://doi.org/10.1016/j.red.2015.07.006

[34] Hilger, N. (2016). The Great Escape: Intergenerational Mobility in the United States Since 1940. *National Bureau of Economic Research Working Papers No. 21217*. https://doi.org/10.3386/w21217

[35] Hodrick, R. & Prescott, E. (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking*, 29(1), 1-16.

[36] Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E. & Puranen, B. (2018). *World Values Survey: All Rounds – Country-Pooled Datafile*. Madrid, Spain & Vienna: J.D. Systems Institute & WVSA Secretariat.

[37] International Monetary Fund (2018). *Global Debt Database*. Washington, D.C.: International Monetary Fund.

[38] Joseph, V. (2022). Optimal Ratio for Data Splitting. *Statistical Analysis and Data Mining*, 531-538. https://doi.org/10.1002/sam.11583

[39] Kourtellos, A. (2021). The Great Gatsby Curve in Education with a Kink. *Economic Letters*, 208. https://doi.org/10.1016/j.econlet.2021.110054

[40] Lam, K. & Liu, P. (2019). Intergenerational Educational Mobility in Hong Kong: Are Immigrants More Mobile Than Natives? *Pacific Economic Review*, 24(1), 137-157. https://doi.org/10.1111/1468-0106.12215

[41] Lee, H. & Lee, J. (2020). Patterns and Determinants of Intergenerational Educational Mobility: Evidence Across Countries. *Pacific Economic Review*, 26(1), 70-90. https://doi.org/10.1111/1468-0106.12342

[42] Lochner, L. & Park, Y. (2022). Earnings Dynamics and Intergenerational Transmission of Skill. *Journal of Econometrics* (forthcoming). https://doi.org/10.1016/j.jeconom.2021.12.009

[43] Louppe, G. (2014). Understanding Random Forests: From Theory to Practice. PhD Thesis.

[44] Lundberg, S. & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*.

[45] Narayan, A., Weide, R., Cojocaru, A., Lakner, C., Redaelli, S., Mahler, D., Ramasubbaiah, R., & Thewissen, S. (2018). Fair Progress? Economic Mobility Across Generations Around the World. Equity and Development. Washington, D.C. World Bank.

[46] Neidhöfer, G. & Stockhausen, M. (2018). Dynastic Inequality Compared: Multigenerational Mobility in the United States, the United Kingdom, and Germany. *Review of Income and Wealth*, 65(2), 383-414. https://doi.org/10.1111/roiw.12364

[47] Nybom, M., & Stuhler, J. (2016). Heterogeneous income profiles and lifecycle bias in intergenerational mobility estimation. *Journal of Human Resources,* 51(1), 239-268.

[48] OECD (2018). *A Broken Social Elevator? How to Promote Social Mobility*. OECD Publishing, Paris.

[49] Olivetti, C. & Paserman, M. (2015). In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940. *American Economic Review*, 105(8), 2695-2724. https://doi.org/10.1257/aer.20130821

[50] Probst, P. (2019). Hyperparameters, Tuning and Meta-Learning for Random Forest and Other Machine Learning Algorithms. PhD Thesis.

[51] Schneebaum, A., Rumplmaier, B., & Altzinger, W. (2016). Gender and Migration Background in Intergenerational Educational Mobility. *Education Economics*, 24(3), 239-260. https://doi.org/10.1080/09645292.2015.1006181

[52] Shapley, L. (1953). A Value for N-person Games. In H. Kuhn & A. Tucker (Eds.), *Contributions to The Theory of Games II*, pp. 307-317, Princeton University Press, Princeton.

[53] Shorrocks, A. (2013). Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value. *The Journal of Economic Inequality*, 11 (1): 99-126. http://doi.org/10.1007/s10888-011-9214-z

[54] Solon, G. (2004). A Model of Intergenerational Mobility Variation over Time and Place. In M. Corak (Ed.), *Generational Income Mobility in North America and Europe*, pp. 38-47, Cambridge University Press, Cambridge.

[55] Städler, N. and Bühlmann, P. (2014). Pattern Alternating Maximization Algorithm for Missing Data in High-Dimensional Problems. *Journal of Machine Learning Research*, 15(55), 1903-1928.

[56] Stekhoven, D. & Bühlmann, P. (2018). MissForest – Non-Parametric Missing Value Imputation For Mixed-Type Data. *Bioinformatics*, 28(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

[57] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. (2001). Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, 17(6), 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

[58] Urbina, D. (2018). Intergenerational Educational Mobility During Expansion Reform: Evidence from Mexico. *Population Research and Policy Review*, 37(3), 367-417. https://doi.org/10.1007/s11113-018-9466-4

[59] Van Buuren, S. & Oudshoorn, K. (1999). Flexible multivariate imputation by MICE. Technical Report, TNO Prevention Center: Leiden, The Netherlands.
[60] World Bank (2018a). *World Development Indicators*. Washington, D.C. World Bank Group.
[61] World Bank (2018b). *World Development Report 2018: Learning to Realize Education's Promise.* Washington, D.C. World Bank Group.

Appendix A – Variables

Table A1 – Predictors of Intergenerational Income Mobility and Intergenerational Education Mobility

Predictors	Proxies	Expected relationship with Income Mobility	Expected relationship with Education Mobility
Human capital	Adult literacy, children's educational attainment, human capital index, parental average education	Ambiguous in Daruich and Kozlowski (2020). Negative in Becker and Tomes (1979), Solon (2004), and Becker *et al*. (2018). Positive in Gallagher *et al*. (2019), and Lochner and Park (2022).	Negative in Daude and Robano (2015), and Neidhöfer and Stockhausen (2018). Positive in Alesina *et al*. (2021).
Public expenditures on education	Government expenditure on education as a share of GDP, government expenditure on primary education as a share of GDP	Positive in Solon (2004).	Positive in Daude and Robano (2015), Urbina (2018), and Lee and Lee (2020).
School quality	Test scores	Positive in Chetty *et al*. (2014a, 2020), Chetty and Hendren (2018), and Acciari *et al*. (2022).	Positive in Hilger (2016).
Employment	Unemployment rate, unemployment for individuals with advanced education, youth unemployment	Negative in Chetty *et al*. (2020), Deutscher and Mazumder (2020), Eriksen and Munk (2020), and Acciari *et al*. (2022).	Negative in Alesina *et al*. (2021, 2023).
Labour market conditions	Female labour force participation, labour force participation	Positive in Eriksen and Munk (2020), and Acciari *et al*. (2022).	NA
Macroeconomic conditions	Economic cycle	Ambiguous in Becker and Tomes (1979). Positive in Chetty *et al.* (2017).	Positive in Hilger (2016), Urbina (2018), and Lee and Lee (2020).
Financial health	Household debt, household disposable income	Positive in Deutscher and Mazumder (2020).	Negative in Lee and Lee (2020). Positive in Daude and Robano (2015).
Segregation/Poverty rate	Poverty rates	Negative in Chetty *et al*. (2014a), Chetty and Hendren (2018), and Deutscher and Mazumder (2020).	Negative in Chetty *et al*. (2014a).
Location attributes	Degree of urbanization, population density, job density	Ambiguous in Chetty *et al*. (2020), and Eriksen and Munk (2020). Negative in Chetty and Hendren (2018), and Chetty *et al*. (2020). Positive in Deutscher (2020).	Ambiguous in Schneebaum *et al.* (2016). Positive in Alesina *et al*. (2021, 2023).
Migration	Migration movements, migrant stock	Negative in Eriksen and Munk (2020). Positive in Gallagher *et al*. (2019).	Ambiguous in Schneebaum *et al*. (2016). Positive in Lam and Liu (2019).
Early childhood development	Gross pre-primary school enrolment	NA	Ambiguous in Schneebaum *et al.* (2016). Positive in Daude and Robano (2015).
High school enrolment	Gross secondary school enrolment	NA	Positive in Hilger (2016).
Inflation	Growth rate of the GDP deflator	NA	Negative in Lee and Lee (2020).

Taxes	Taxes on income, profits, and capital gains	Ambiguous in Becker and Tomes (1979).	NA
Public policies	Subsidies and transfers	Positive in Chetty *et al*. (2020).	Positive in Daude and Robano (2015).
Income inequality	Gini index	Negative in Chetty *et al*. (2014a,b, 2017), Olivetti and Paserman (2015), Becker *et al*. (2018), Chetty and Hendren (2018), Corak (2019), Lochner and Park (2022), and Acciari *et al*. (2022).	Negative in Daude and Robano (2015), Hilger (2016), and Lee and Lee (2020).
Income shares	Income share of the 10% richest individuals	Positive in Acciari *et al*. (2022).	NA
Geography	Regions of the World	NA	NA
Household structure	Share of single parents	Negative in Chetty *et al*. (2014a, 2020), Chetty and Hendren (2018), Gallagher *et al.* (2019), and Eriksen and Munk (2020).	Negative in Alesina *et al*. (2023).
Family instability	Share of divorces	Negative in Acciari *et al*. (2022).	NA
Share of married individuals	Share of marriages	Positive in Eriksen and Munk (2020).	NA
Marriage age	Average marriage age of first marriage for women	NA	Positive in Alesina *et al*. (2023)
Total fertility rate	Total fertility rate	Positive in Daruich and Kozlowski (2020).	NA
Teen birth	Share teen females that are pregnant or have had children	Negative in Chetty *et al.* (2020), and Eriksen and Munk (2020).	NA
Child mortality	Probability a child has of dying before 5 years old	Negative in Olivetti and Paserman (2015).	NA
Maternal mortality	Share of women dying during pregnancy due to gestation problems	Negative in Olivetti and Paserman (2015).	NA
Gender	Share of female population	NA	NA
Social capital	Trust level	Positive in Chetty *et al.* (2014a, 2020), and Chetty and Hendren (2018).	NA
Wars	Deaths due to wars, conflicts, and terrorism	Negative in Olivetti and Paserman (2015).	NA
Religion	Religion followed by the greatest share of individuals	NA	NA
Malaria existence	Malaria incidence	NA	Negative in Alesina *et al*. (2021).

Note: NA - not applicable.

Variables Definitions

- **Human capital**

Adult literacy (litadult). Adult literacy consists of the percentage share of individuals aged 15 or older who are able to write, read, and comprehend short/simple statements regarding their ordinary life.

Children's educational attainment. Five variables contain information on children's educational attainment, namely, the share of children who have completed less than primary, primary, lower secondary, upper secondary, and tertiary education levels (C1, C2, C3, C4, and C5 respectively). We also consider another variable reflecting the children's mean education years (MEANc).

Human capital index (HK). The human capital index contains information about the mean school years and also returns to education.

Parental average education (MEANp). This variable accounts for the parents' average number of education years. It is calculated separately for each cohort.

- **Public expenditures on education**

Government expenditure on education as a share of GDP (educexp). This variable regards the expenditures of the general government devoted to education (% of GDP). It accounts for the expenses financed by the transfers the government receives from international sources.

Government expenditure on primary education as a share of GDP (primexp). This variable is given by the per-student government expenditure (transfers, current, and capital) as a share (%) of GDP *per capita*.

- **School quality.** Three test score-related measures are considered to measure school quality, namely the mean scores that 15-year-old individuals received on the PISA mathematics, reading, and science scales (PISAM, PISAR, and PISAS, respectively).

- **Employment**

Unemployment rate (un). The unemployment rate corresponds to the share of people (%) in the labour force looking and available for employment but have no job.

Unemployment with advanced education (unadveduc). This variable provides the percentage share of the labour force that has attained a higher education degree and is unemployed.

Youth unemployment (unyoung). The youth unemployment corresponds to the ratio (%) between the unemployed individuals, 15-24 years old, who are available/seeking to be employed and the labour force.

- **Labour market conditions**

Female labour force (femlabforce). This variable corresponds to the female subsample of the labour force participation rate, i.e., gives the ratio (%) between women, which supply labour for a specific period, and total labour force, with ages ranging from 15-64 years old.

Labour force participation rate (labforce). The labour force participation rate gives the ratio (%) between all those who supply labour for a specific period, and the labour force, with ages ranging between 15 and 64 years old.

- **Macroeconomic conditions**

Economic cycle (cycle). We compute this variable using the Hodrick and Prescott (1997) filter, which decomposes for each country the real GDP series on a trend and a business cycle component: the latter reflecting, if negative, an economic recession.

GDP *per capita* growth (GDPpcg). Annual percentage growth rate for real GDP *per capita* (at constant 2010 US$). Real GDP *per capita* is obtained by dividing real GDP (at constant 2010 US$) by the *de facto* mid-year population estimates.

- **Financial health**

Household debt (hdebt). This variable corresponds to the ratio (%) between the entire stock of loans and debt securities owned by households and a country's GDP.

Household disposable income (avinc). The household disposable income is obtained by subtracting taxes and contributions for social security from the income that results from employment and self-employment, capital, transfers (social security payments related to work insurance, assistance and universal benefits, and private transfers). The measure accounts for inflation and household size and is presented in 2011 international dollars.

- **Segregation/Poverty rate.** Poverty measures include the percentage shares of population living on less than $1.90, $3.20, and $5.50 *per* day (pov190, pov320, and pov550, respectively), considering international 2011 prices.

- **Location attributes**

Degree of urbanization (urban). This variable corresponds to the share (in %) of total population living in urban areas.

Job density (jobden). We calculate job density by dividing the employment rate by the land area in square kilometres. This last variable includes all the country's area excluding major rivers and lakes, continental shelf claims, and exclusive economic zones.

Population density (popden). The population density is given by the ratio between mid-year *de facto* population (which includes all residents, citizens and noncitizens, despite their legal status) and land area (measured in square kilometres).

- **Migration**

Migration movements (netmig). Migration movements can be measured by five-year estimates computed by subtracting the annual number of emigrants from the number of immigrants regardless of their citizenship.

Migrant stock (migstock). Migrant stock corresponds to the total number of individuals who are born in a country other than the one where they live, as a percentage of total population.

- **Early childhood development (preenrol).** Early childhood development can be measured by the gross pre-primary school enrolment (preenrol), which is given by the ratio (in %) between the number of individuals enrolled at the pre-primary level of education, independently of their age, and the total population with an age that officially matches the one for that level.

- **High school enrolment (secondenrol).** The gross secondary school enrolment is given by the ratio (in %) between the number of individuals enrolled at the secondary level of education, independently of their age, and the total population with an age that officially matches the one for that level.

- **Inflation (infl).** This variable corresponds to the annual percentage growth rate of the GDP deflator.

- **Taxes (tax).** Taxes on income, profits, and capital gains correspond to the sum of taxes applied on real or expected income from individuals, firms, profits, and capital gains (land, securities, among other assets).

- **Public policies (subtransf).** Subsidies and transfers include unilateral transfers, which are not repayable to either public or private companies; grants attributed to own government branches and to foreign governments, worldwide organizations; and social security, assistance benefits, and monetary and non-monetary benefits to employers. It is presented as a percentage of Government expenditures.

- **Income inequality (Gini).** The Gini index measures the area between a perfectly equal income distribution and the Lorenz curve (this plots cumulative income received against the cumulative population of receivers, both in percentages), and is expressed as a percentage of the maximum area below the first. It ranges between 0 and 100, meaning that there is no inequality in the first scenario and that there is no equality in the second.

- **Income shares (inc10).** We use the share of consumption or income of the 10% richest individuals of a population to measure the income share of the 10% richest.

- **Geography (region).** This variable corresponds to the geographic region of the world that a country belongs to, from among East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, South Asia, and Sub-Saharan Africa. Additionally, another group is presented and corresponds to a high-income category. We transform it into dummy variables, equal to the unit when we are in a specific geographic region and zero otherwise.

- **Household structure (singlepar).** We measure the household structure by the share of single parents, defined as the share of households composed by only a single parent and the corresponding children (from among adopted, biological, or stepchildren).

- **Family instability (div).** Family instability is measured by the number divorces (div) as a share of mean population, *per* 1000 inhabitants and *per* year.

- **Share of married individuals (marr).** Data on the number of marriages (marr) as a share of mean population, *per* 1000 inhabitants and *per* year. It is expected that the share of married inhabitants relates in a positive way with income mobility.

- **Marriage age (agemarrwomen).** For OECD countries we have direct information reflecting the average age of women when they first married. For countries outside the OECD, the marriage age considering women is an estimate of first marriage mean age.

- **Total fertility rate (fert).** Conditional on a woman living until the end of her childbearing years and bearing children according to fertility rates that are age-specific, the total fertility rate consists of the expected number of children she will give birth to.

- **Teen birth (teenbirth).** This variable corresponds to the percentage share of 15-19-year-old women who are pregnant or have had children.

- **Child mortality (childmort).** Child mortality reflects the probability that a child has of dying before the age of 5 years old, per 1000 live births, accounting for mortality rates associated with age.

- **Maternal mortality (matmort).** The maternal mortality reflects the number of women dying per 100,000 births, throughout pregnancy or in the last 42 days of pregnancy, due to gestation-related causes.

- **Gender (fempop).** Although we consider only men in our analysis, we introduce a variable that may contain information on gender differences of a country. We therefore consider a female population variable, which corresponds to the share of the *de facto* population, i.e., the population of all residents, not accounting for their citizenship or legal status, that is female.

- **Social capital (trust).** Trust is usually used as a proxy for social capital. The trust level in a society is evaluated through the following question: "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?". We consider only the answers "most people can be trusted" and "need to be very careful". For each country and each wave, we averaged all the respondent's valid answers.

- **Wars (confterr).** This variable considers the sum of deaths (*per* 100,000) related to war between states, conflicts between civilians, and terrorism.

- **Religion (religion).** We consider a categorical variable reflecting the religion, which is followed by the greatest share of individuals in a country as of 2010, from among Buddhism, Christianity, Islam, Folk Religions, Hinduism, Judaism, and Unaffiliated Religions. We created three main dummies grounded on the share of believers each religion has in the World according to the WGBH Educational Foundation: the first is for Christianity, which has the largest share of followers, the second for Islam, which has the second highest share of followers, and a third one containing all other religions (OthersR).

- **Malaria existence (malaria).** The malaria incidence is given by the number of malaria cases appearing *per* 1000 at-risk individuals in a given year.

Appendix B – Tables

Table B1 – Descriptive Statistics for Predictors of IM in Income

Variables	Obs.	Mean	Std. Dev.	Min	Max
avinc	70	22271.06	3448.42	11116.65	34088.88
C1	70	0.14	0.20	0.00	0.81
C2	70	0.12	0.13	0.00	0.65
C3	70	0.14	0.09	0.02	0.41
C4	70	0.36	0.20	0.02	0.81
childmort	70	63.47	62.15	9.02	226.32
cohort60	70	0.49	0.50	0.00	1.00

cohort70	70	0.51	0.50	0.00	1.00
confterr	70	6.85	45.65	0.00	381.60
div	70	1.31	0.88	0.13	4.10
EastAsiaPacific	70	0.07	0.26	0.00	1.00
educexp	70	4.31	1.30	1.62	8.85
EuropeCentralAsia	70	0.13	0.34	0.00	1.00
femlabforce	70	42.09	7.63	15.01	53.53
fempop	70	50.55	1.10	47.74	53.96
fert	70	3.46	1.72	1.63	6.93
GDPpcg	70	2.53	1.74	-1.95	11.16
Gini	70	37.44	8.70	24.60	60.89
Highincome	70	0.41	0.50	0.00	1.00
HK	70	2.31	0.68	1.13	3.44
IGEincome	70	0.53	0.25	0.11	1.10
IGP	70	0.39	0.16	0.08	0.78
inc10	70	29.63	7.01	20.98	48.65
jobdeng	70	0.00	0.00	0.00	0.01
labforce	70	69.37	8.99	43.16	88.35
LatinAmericaCaribbean	70	0.10	0.30	0.00	1.00
marr	70	6.12	1.87	2.10	11.80
matmort	70	189.41	278.10	3.82	1057.81
MEANc	70	10.48	3.15	2.16	14.75
MEANp	70	6.99	3.55	0.53	13.45
MiddleEastNorthAfrica	70	0.06	0.23	0.00	1.00
migstock	70	6.00	7.04	0.04	39.42
netmig	70	65301.27	544779.00	-837680.70	3939991.00
PISAM	70	454.30	59.60	292.00	600.08
PISAR	70	452.53	53.71	299.36	555.83
PISAS	70	468.29	48.46	325.79	574.62
popdeng	70	0.01	0.01	0.00	0.04
pov190	70	17.38	23.95	0.01	85.75
pov320	70	28.24	32.77	0.05	94.95
pov550	70	39.58	37.55	0.10	98.80
singlepar	70	8.89	2.66	5.08	19.40
SouthAsia	70	0.04	0.20	0.00	1.00
SubSaharanAfrica	70	0.19	0.39	0.00	1.00
subtransf	70	38.57	15.74	0.00	75.27
tax	70	35.91	16.95	9.32	88.71
teenbirth	70	12.40	8.16	2.90	40.83
trustg	70	0.00	0.00	-0.01	0.01
un	70	7.70	5.95	0.61	33.16
unadveduc	70	7.48	5.48	2.17	28.88
unyoung	70	15.62	11.85	0.99	58.53
urban	70	52.63	21.62	7.80	95.37

Table B2 – Descriptive Statistics for IM in Education

Variables	Obs.	Mean	Std. Dev.	Min	Max
agemarrwomen	338	24.01	3.08	17.48	30.94
avinc	338	17271.06	5287.45	10981.55	34859.54
Christianity	338	0.65	0.48	0.00	1.00
cohort60	338	0.30	0.46	0.00	1.00
cohort70	338	0.30	0.46	0.00	1.00
cohort80	338	0.40	0.49	0.00	1.00
cycle	338	-3.39e+08	1.63e+09	-1-94e+10	4.36e+09
EastAsiaPacific	338	0.09	0.29	0.00	1.00
educexp	338	4.30	1.49	1.07	11.35
EuropeCentralAsia	338	0.18	0.38	0.00	1.00
fempop	338	50.58	1.14	47.72	54.15
GDPpcg	338	2.25	1.87	-6.94	11.16
Gini	338	37.84	7.98	24.94	62.15
hdebt	338	24.88	23.80	1.25	111.82

Highincome	338	0.31	0.46	0.00	1.00
HKg	338	0.01	0.00	0.00	0.02
IGEincome	338	0.54	0.18	0.11	1.10
IGP	338	0.39	0.16	-0.21	0.98
infl	338	45.05	114.12	0.34	1082.08
Islam	338	0.22	0.42	0.00	1.00
LatinAmericaCaribbean	338	0.07	0.26	0.00	1.00
litadult	338	81.29	21.89	21.64	99.83
malaria	338	97.60	146.16	0.06	590.93
MiddleEastNorthAfrica	338	0.06	0.24	0.00	1.00
migstock	338	5.91	6.81	0.05	39.42
OthersR	338	0.12	0.33	0.00	1.00
PISAM	338	467.55	42.30	320.87	581.35
PISAR	338	464.92	40.06	299.36	539.79
PISAS	338	472.26	39.77	325.79	557.50
popden	338	97.04	115.90	1.29	915.09
pov190	338	17.25	22.53	0.00	85.75
pov320	338	29.88	31.00	0.05	94.95
pov550	338	44.15	35.85	0.12	98.80
preenrol	338	46.05	29.77	0.92	112.76
primexp	338	16.67	7.79	3.41	50.17
secondenrol	338	68.44	32.26	7.09	145.99
singlepar	338	8.64	2.57	3.86	19.40
SouthAsia	338	0.05	0.22	0.00	1.00
SubSaharanAfrica	338	0.24	0.43	0.00	1.00
subtransf	338	37.31	16.57	0.00	77.38
un	338	8.10	5.85	0.82	33.16
urban	338	50.70	21.77	8.72	96.87

Table B3 – IGPI Covariates with Missing Data

Features	Missing Rel. Freq (%)
Gini	2.86
HK	8.57
subtransf	14.29
tax	10.0
teenbirth	51.43
pov190	2.86
pov320	2.86
pov550	2.86
singlepar	15.71
avinc	70.0
PISAR	41.43
PISAM	41.43
PISAS	41.43
inc10	5.71
educexp	5.71
marr	20.0
div	27.14
unadveduc	8.57
trustg	24.29

Table B4 – IGPE Covariates with Missing Data

Features	Missing Rel. Freq (%)
IGPI	78.99
un	1.18
cycle	0.3
hdebt	46.15

Gini	1.18
litadult	18.05
subtransf	15.98
urban	0.89
secondenroll	2.66
migstock	0.89
pov190	1.18
pov320	1.18
pov550	1.18
agemarrwomen	0.89
singlepar	15.38
avinc	76.92
fempop	0.89
PISAR	47.34
PISAM	47.34
PISAS	47.34
malaria	49.11
preenroll	2.66
primexp	15.68
educexp	4.14
HKg	14.5
Christianity	0.89
Islam	0.89

Table B5 – Hyperparameters Chosen for Random Forest and Gradient Boosting

Hyperparameters	Random Forest		Gradient Boosting	
	Income	Education	Income	Education
Number of estimators Number of trees in the forest.	$B = 400$	$B = 1700$	$M = 1200$	$M = 1200$
Number of features for a split Number of features to consider when deciding which one will lead to the best split.	$\sqrt{K} = \sqrt{50}$	$K = 41$	$\sqrt{K} = \sqrt{50}$	$\sqrt{K} = \sqrt{41}$
Minimum sample size for a split The minimum number of observations required to split an internal node.	2	2	2	12
Maximum depth The maximum depth of the tree. If "None", the tree nodes expand until purity is reached in all leaves or these contain less than the minimum sample size for a split.	None	None	60	50
Minimum sample size in a leaf The minimum number of observations to be in a leaf.	1	2	2	3
Bootstrap Whether bootstrap samples are used when building trees. If "No", sampling is done without replacement.	No	No	NA	NA
Learning rate contribution of each tree The contribution of each tree to the final prediction.	NA	NA	$\rho = 0.1$	$\rho = 0.1$

Note: NA - not applicable.

Table B6 – Hyperparameters Chosen for Random Forest and Gradient Boosting removing studies using OLS (IGPI only)

Hyperparameters	Random Forest	Gradient Boosting
Number of estimators Number of trees in the forest.	$B = 1600$	$M = 1500$
Number of features for a split Number of features to consider when deciding which one will lead to the best split.	50	$\sqrt{K} = \sqrt{50}$
Minimum sample size for a split The minimum number of observations required to split an internal node.	11	3
Maximum depth	None	50

The maximum depth of the tree. If "None", the tree nodes expand until purity is reached in all leaves or these contain less than the minimum sample size for a split.		
Minimum sample size in a leaf The minimum number of observations to be in a leaf.	1	1
Bootstrap Whether bootstrap samples are used when building trees. If "No", sampling is done without replacement.	Yes	NA
Learning rate contribution of each tree The contribution of each tree to the final prediction.	NA	$\rho = 0.05$

Note: NA - not applicable.

Table B7 – Hyperparameters Chosen for Random Forest and Gradient Boosting with Robust Predictors of Mobility

Hyperparameters	Random Forest		Gradient Boosting	
	Income	Education	Income	Education
Number of estimators Number of trees in the forest.	$B = 2100$	$B = 1100$	$M = 1900$	$M = 900$
Number of features for a split Number of features to consider when deciding which one will lead to the best split.	$\sqrt{K} = \sqrt{6}$	$\sqrt{K} = \sqrt{6}$	$\sqrt{K} = \sqrt{6}$	$\sqrt{K} = \sqrt{6}$
Minimum sample size for a split The minimum number of observations required to split an internal node.	4	2	2	3
Maximum depth The maximum depth of the tree. If "None", the tree nodes expand until purity is reached in all leaves or these contain less than the minimum sample size for a split.	None	100	90	90
Minimum sample size in a leaf The minimum number of observations to be in a leaf.	5	1	5	1
Bootstrap Whether bootstrap samples are used when building trees. If "No", sampling is done without replacement.	No	Yes	NA	NA
Learning rate contribution of each tree The contribution of each tree to the final prediction.	NA	NA	$\rho = 0.25$	$\rho = 0.01$

Note: NA - not applicable.

Appendix C – Figures

Figure C1 – Intergenerational Persistence of Income for the 1960 Cohort

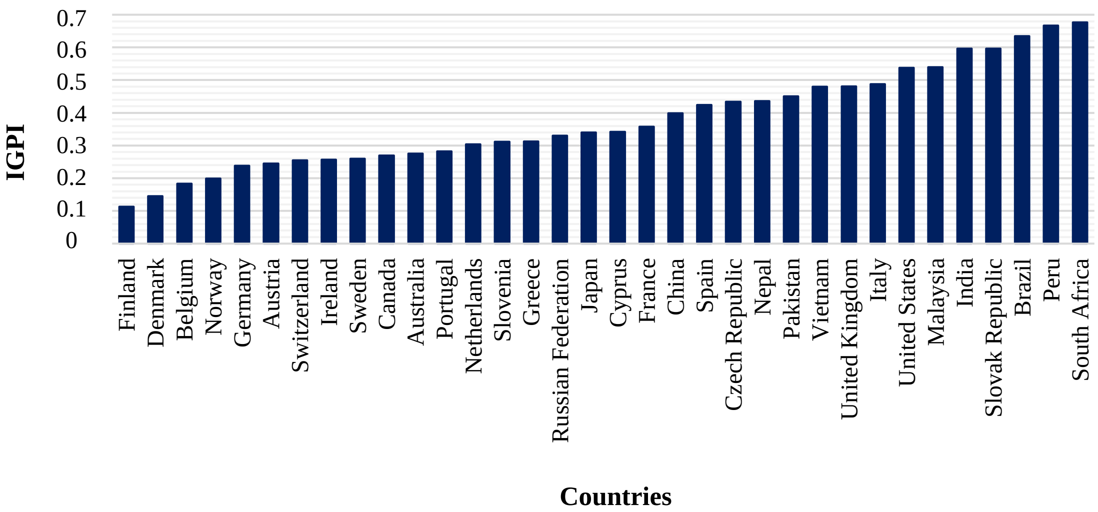


Figure C2 – Intergenerational Persistence of Income for the 1970 Cohort

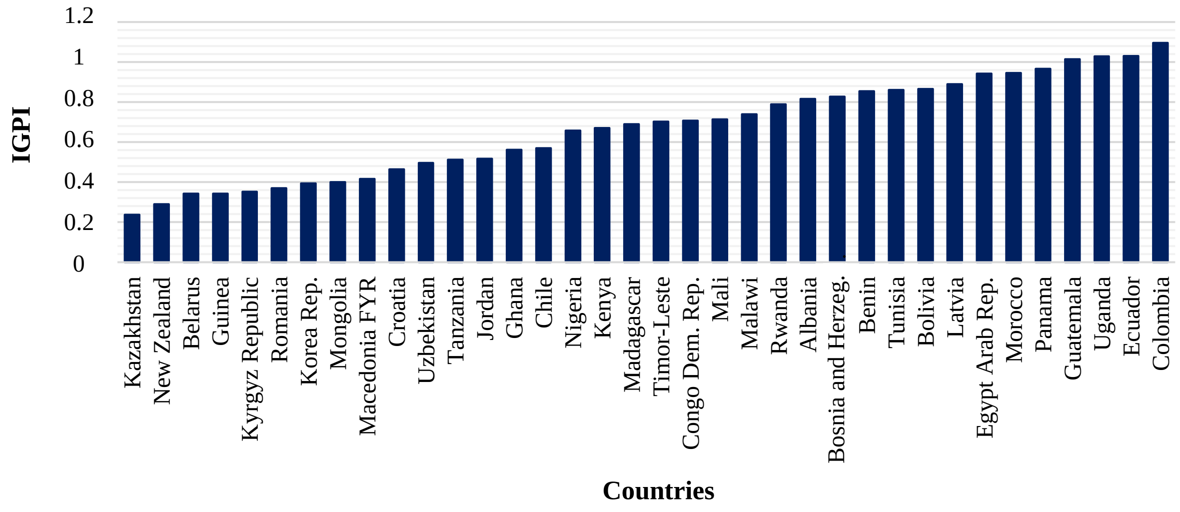
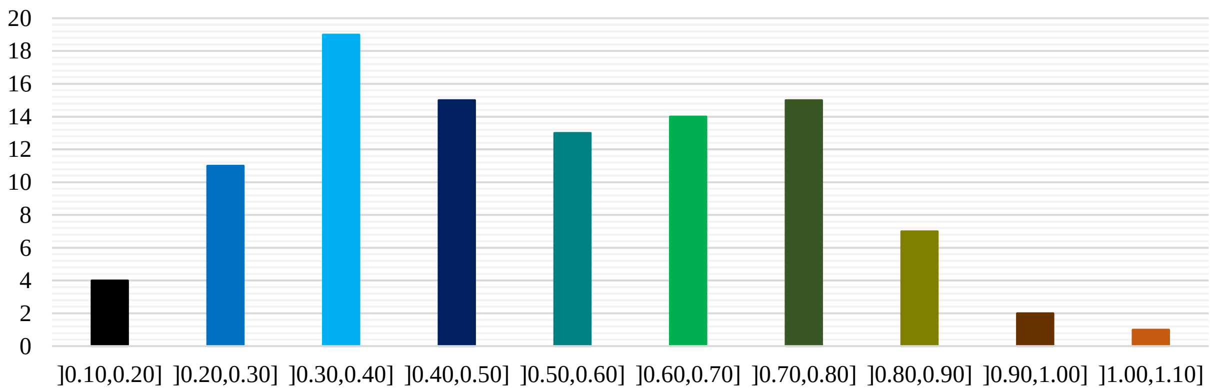


Figure C3 – Predictions of Intergenerational Income Persistence for the 1960 Cohort



Legend:

Belgium, Denmark, Finland, Norway

Australia, Austria, France, Ireland, Germany, Greece, Iceland, Kazakhstan, Slovenia, Sweden, Switzerland

Armenia, Azerbaijan, Belarus, Canada, China, Cyprus, Italy, Korea Rep., Kosovo, Moldova, Netherlands, New Zealand, Portugal, Romania, Russian Federation, Turkey, Ukraine, United Kingdom, United States

Albania, Bulgaria, Chile, Czech Republic, Estonia, Guinea, Hungary, Israel, Japan, Kyrgyz Republic, Mongolia, Montenegro, Nepal, Spain, Vietnam

Croatia, Ethiopia, Georgia, Liberia, Macedonia FYR, Malaysia, Philippines, Poland, Slovak Republic, Sri Lanka, Tajikistan, Tanzania, Uzbekistan

Benin, Congo Dem. Rep., Ghana, India, Indonesia, Jordan, Lao PDR, Latvia, Lithuania, Madagascar, Mali, Niger, Nigeria, South Africa

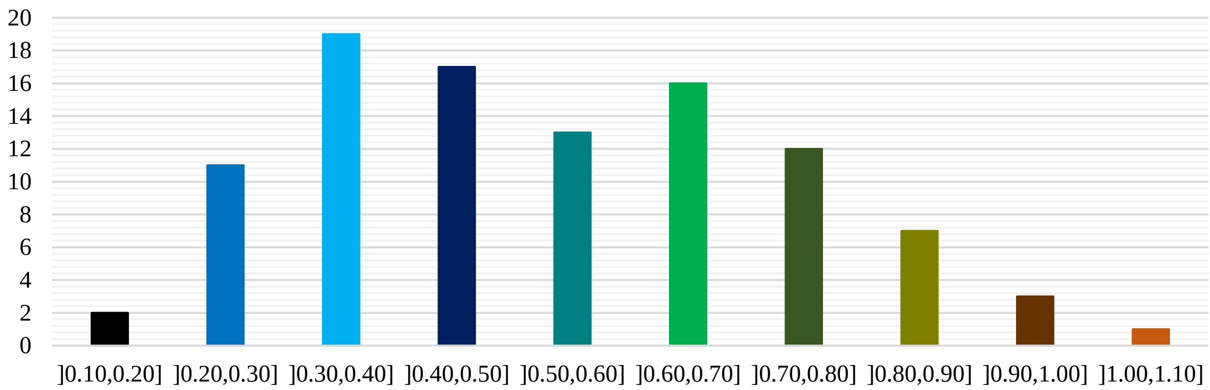
Bolivia, Bosnia and Herzegovina, Brazil, Colombia, Comoros, Ecuador, Egypt Arab Rep., Iraq, Malawi, Pakistan, Peru, Rwanda, Serbia, Timor-Leste, Togo

Bhutan, Kenya, Mauritania, Mexico, Panama, Senegal, Uganda

Morocco, Tunisia

Guatemala

Figure C4 – Predictions of Intergenerational Income Persistence for the 1970 Cohort



Legend:

Portugal, Slovenia

Australia, Azerbaijan, Belgium, Denmark, Finland, Iceland, Ireland, Kazakhstan, Netherlands, Norway, Switzerland

Armenia, Austria, Belarus, Canada, Cyprus, Czech Republic, France, Germany, Israel, Japan, Kosovo, New Zealand, Romania, Sri Lanka, Sweden, Turkey, United Kingdom, United States, Vietnam

Albania, Chile, Croatia, Estonia, Greece, Hungary, Korea Rep., Kyrgyz Republic, Macedonia FYR, Malaysia, Moldova, Mongolia, Montenegro, Poland, Slovak Republic, Spain, Ukraine

Bulgaria, China, Guinea, Indonesia, Italy, Jordan, Mexico, Niger, Philippines, Russian Federation, Tajikistan, Tanzania, Uzbekistan

Congo Dem. Rep., Ecuador, Ethiopia, Georgia, Ghana, India, Iraq, Lao PDR, Liberia, Lithuania, Madagascar, Mali, Nepal, Nigeria, South Africa, Timor-Leste

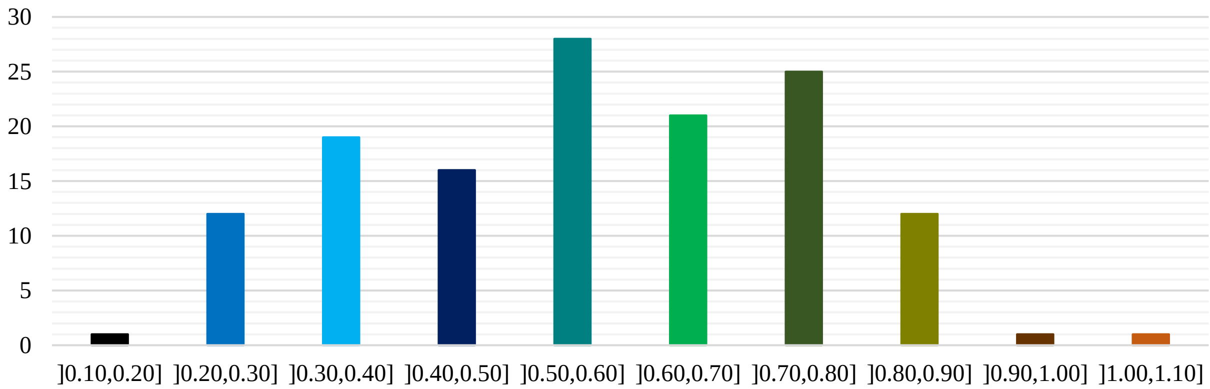
Bhutan, Brazil, Colombia, Comoros, Kenya, Latvia, Malawi, Pakistan, Peru, Rwanda, Senegal, Serbia

Bolivia, Bosnia and Herzegovina, Mauritania, Morocco, Togo, Tunisia, Uganda

Benin, Egypt Arab Rep., Panama

Guatemala

Figure C5 – Predictions of Intergenerational Income Persistence for the 1980 Cohort



Legend:

Belgium

Australia, Azerbaijan, Canada, Denmark, Fiji, Finland, Iceland, Ireland, Kazakhstan, Netherlands, Norway, Switzerland

Austria, Belarus, China, Cyprus, France, Germany, Indonesia, Israel, Kyrgyz Republic, Mauritius, Romania, Sri Lanka, Sweden, Thailand, Tonga, Turkey, United Kingdom, United States, Vietnam

Chile, Czech Republic, Estonia, Hungary, Iran Islamic Rep., Japan, Korea Rep., Malaysia, Maldives, Mongolia, Poland, Portugal, Slovenia, Spain, Ukraine, Uzbekistan

	Albania, Armenia, Bangladesh, Brazil, Central African Republic, Croatia, Guinea, India, Italy, Jordan, Kiribati, Kosovo, Lebanon, Liberia, Macedonia FYR, Mexico, Moldova, Namibia, Nepal, Niger, Nigeria, Russian Federation, Sierra Leone, Slovak Republic, South Africa, Tajikistan, Tanzania, Vanuatu
	Botswana, Bulgaria, Cambodia, Colombia, Congo Dem. Rep., Congo Rep., Ecuador, Ethiopia, Georgia, Ghana, Greece, Iraq, Lao PDR, Latvia, Lithuania, Mali, Montenegro, Peru, Philippines, Sao Tome and Principe, Timor-Leste
	Angola, Bolivia, Bosnia and Herzegovina, Burkina Faso, Cabo Verde, Cameroon, Chad, Comoros, Cote d'Ivoire, Egypt Arab Rep., Gabon, Kenya, Lesotho, Madagascar, Malawi, Morocco, Mozambique, Pakistan, Panama, Rwanda, Serbia, South Sudan, Swaziland, West Bank and Gaza, Yemen Rep.
	Afghanistan, Bhutan, Djibouti, Guinea-Bissau, Mauritania, Papua New Guinea, Senegal, Sudan, Togo, Tunisia, Uganda, Zambia,
	Benin
	Guatemala