Using Machine Learning to Unveil the Determinants 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

This version: November 2023

Abstract

We assess the determinants 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. The Rigorous LASSO and the Random Forest and Gradient Boosting algorithms are considered, avoiding the consequences of an *ad-hoc* model selection in our high dimensionality context. Through variable importance plots and Shapley values, we find that income mobility is expected to be positively influenced by the share of married individuals and negatively influenced by the share of children with less than primary education, the growth rate of population density, and inequality. Mobility in education appears to have a positive relationship with the adult literacy, government expenditures on primary education, and the stock of migrants. Income mobility is greater for the 1960s cohort. Latin America and Caribbean countries present lower mobility. Predicted income mobility and observed education mobility are positively related.

JEL Classification: E24; I24; J62; O15.

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

* We thank Ambar Narayan, Daniel Mahler, and João Moura for their helpful contributions. We also thank the participants of the 44th Meeting Association of Southern European Economic Theorists (ASSET) for the feedback and suggestions that improved this work. We acknowledge financial support from FCT - Fundação para a Ciência e a Tecnologia (National Science and Technology Foundation) through grants 2020.04449.BD and UIDB/00315/2020.

The corresponding author is Luís Clemente-Casinhas. E-mail: lmccs@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 determinants of inequality. Hence, the study of the determinants 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 from 137 countries from 1960 to 2018 to assess the most important variables determining 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 determinants 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 influence mobility. Deciding which variables to choose when computing mobility regressions is a non-trivial challenge and, as a result of 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 determinants of intergenerational mobility/persistence in income and education. We are the first to present a Worldwide outlook to find the determinants 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 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 determinants of IM identified in earlier works for which information is available.

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

Lee and Lee's (2020) work is the closest to ours in the sense that they explore the determinants of educational mobility, although their study targets OECD countries and few determinants 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 determines 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 determinants, 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 determinants 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 determinants 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 determine 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 the share of married individuals 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 influenced by adult literacy, government expenditures on primary education, and the migrant stock. Although income mobility appears to be influenced 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 and Section 4 some robustness exercises. Section 5 explores the contribution of each feature for individual predictions of mobility. In Section 6 we use our model to predict income mobility for a specific set of countries and study the relationship between income and education mobility. Section 7 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: policy makers should devote more of their attention to relative mobility because that is the measure that will allow policy makers to assess if individuals are improving their current living standards (which must always be analysed in relative terms).

There are no defined criteria in the published research about which individual exactly in the parent-child pair should be used in an analysis, even when authors are not constrained in terms of data. Although this is the case, 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 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. Also, women may be absent from the labour market due to maternity related issues, and therefore, their activity status may be intermittent,

[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).

which supports the previous view. Therefore, when married women are included in the analysis, their individual earnings may not accurately reflect their economic status. 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. 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, positive changes in the indicators are signs of intergenerational persistence, negative changes in the indicator are signs of 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 variable estimation (TSTSIV), the two-sample instrumental variables estimation (TSIV), instrumental variables estimation (IV), or two-stage least squares (TSLS). 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. 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 there are countries in which the change in the child's education will never exceed that of the parent.

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

[3] http://equalchances.org

2.1.2. Intergenerational Mobility Determinants

We will use as explanatory variables for IM in income and/or education the determinants, 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 determinants that are available for the widest number of countries possible. The definitions of the mobility determinants and the related literature review can be found in the Appendix A. Table 2 summarizes those determinants and the effects they are expected to have on income and education mobility.[4]

Table 2 – Determinants of Intergenerational Mobility in Income and Education

Determinants	Expected Effect	
	Income	**Education**
Human capital		
1. Adult literacy (litadult).	NA	Positive
2. Children's educational attainment: share of children who have completed less than primary, primary, lower secondary, upper secondary, and tertiary education levels and children's mean education years (C1, C2, C3, C4, C5, and MEANc, respectively).	Ambiguous	NA
3. Human capital index (HK).	Ambiguous	Negative
4. Parental average education (MEANp).	Positive	NA
Public expenditures on education		
1. Government expenditure on education as a share of GDP (educexp).	Positive	Positive
2. Government expenditure on primary education as a share of GDP (educexp).	NA	Positive
School quality		
1. Test scores on the PISA mathematics, reading, and science scales (PISAM, PISAR, and PISAS, respectively).	Positive	Positive
Employment		
1. Unemployment rate (un).	Negative	Negative
2. Unemployment rate for individuals with advanced education (unadveduc).	Negative	NA
3. Youth unemployment (unyoung).	Negative	NA
Labour market conditions		
1. Female labour force (femlabforce).	Positive	NA
2. Labour force participation rate (labforce).	Positive	NA
Macroeconomic conditions		
1. Economic cycle (cycle).	NA	Positive
2. GDP *per capita* growth (GDPpcg).	Ambiguous	Positive
Financial health		
1. Household debt (hdebt).	NA	Negative
2. Household disposable income (avinc).	Positive	Positive
Segregation/Poverty rate		
1. Shares of population living on less than \$1.90, \$3.20, and \$5.50 *per* day (pov190, pov320, and pov550, respectively).	Negative	Negative
Location attributes		
1. Degree of urbanization (urban).	Ambiguous	Ambiguous
2. Job density (jobden).	Negative	NA

[4] 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, the Global Debt Database from the International Monetary Fund, and the World Values Survey by Inglehart *et al*. (2018).

3. Population density (popden).	Positive	Positive
Migration		
1. Migration movements (netmig).	Positive	NA
2. Migrant stock (migstock).	Ambiguous	Ambiguous
Early childhood development		
1. Gross pre-primary school enrolment (preenroll).	NA	Ambiguous
High school enrolment		
1. Gross secondary school enrolment (secondenroll).	NA	Positive
Inflation		
1. Growth rate of the GDP deflator (infl).	NA	Negative
Taxes		
1. Taxes on income, profits, and capital gains (tax).	Ambiguous	NA
Public policies		
1. Subsidies and transfers (subtransf).	Positive	Positive
Income inequality		
1. Gini index (Gini).	Negative	Negative
Income shares		
1. Income share of the 10% richest individuals (inc10).	Positive	NA
Geography		
1. Geographic region of the world that a country belongs to among East Asia and Pacific (EastAsiaPacific), Europe and Central Asia (EuropeCentalAsia), Latin America and Caribbean (LatinAmericaCaribbean), Middle East and North Africa (MiddleEastNorthAfrica), South Asia (SouthAsia), and Sub-Saharan Africa (SubSaharanAfrica).	NA	NA
Household structure		
1. Share of single parents (singlepar).	Negative	Negative
Family instability		
1. Share of divorces (div).	Negative	NA
Share of married individuals		
1. Share of marriages (marr).	Positive	NA
Marriage age		
1. Average marriage age of the first marriage for women (agemarrwomen).	NA	Positive
Total Fertility Rate		
1. Total fertility rate (fert).	Positive	NA
Teen birth		
1. Share of teen females who are pregnant or have had children (teenbirth).	Negative	NA
Child mortality		
1. Probability a child has of dying before the age of 5 (childmort).	Negative	NA
Maternal mortality		
1. Share of women dying during pregnancy due to problems in gestation (matmort).	Negative	NA
Gender		
1. Share of female population (fempop).	NA	NA
Social capital		
1. Trust level (trust).	Positive	NA
Wars		
1. Deaths due to wars, conflicts, and terrorism (confterr).	Negative	NA
Religion		
1. Religion followed by the greatest share of individuals in a country, among Christianity (Christianity), Islam (Islam), and other religions (OthersR), which include Buddhism, Folk Religions, Hinduism, Judaism, and Unaffiliated Religions.	NA	NA

Malaria existence		
1. Malaria incidence (malaria)	NA	Negative

Note: NA - not applicable.

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 determinants 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 influenced 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 determinants 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. 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[5]. 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 determinants 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).

[5] 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 summary of the periods for which potential mobility determinants are averaged over time for each country is in Table 3.

Table 3 – Time Periods for Averaged Determinants 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 1), since the averages we construct rely on the evidence that for each of the cohorts the variables are stationary over time[6]. 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 determinants for income as well as for education mobility. Considering the exceptions, we calculate their growth rates: for the income determinants, job density (jobden), population density (popden), and social capital (trust), and for the education determinants 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 (LASSO) and the other two are the Random Forest and the Gradient Boosting Regressors[7]. They are described below.

Rigorous Least Absolute Shrinkage and Selection Operator

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

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

where, for country i, IM_i corresponds to the mobility measure we seek to explain (mobility in income or education, interchangeably) by its determinants $IMD_{i,k}$ (K variables at most, including cohort dummies) 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

[6] 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 1 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.

[7] These machine learning algorithms as well as the RLASSO are run by imputing missing data with the Miss-Forest algorithm. Details on this method can be found in Stekhoven and Bühlmann (2012).

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 determinants 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 determine 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.

Our dataset is randomly split into training and testing data, following the literature, which commonly uses an 80:20 ratio when splitting the data, grounded on the Pareto Principle (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 bootstrap 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 determinants, 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 $\widehat{f_0} = \text{argmin}_\chi \sum_{i=1}^{n} L(IM_i, \chi)$. 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 $\theta_{z,m} = \text{argmin}_\theta \sum_{i=1}^{n} L\left(IM_i, \widehat{f_{m-1}}(\cdot) + \theta\right), \forall i \in R_{z,m}$. Finally it updates the model with $\widehat{f_{\iota,m}}(\cdot) = \widehat{f_{\iota,m-1}}(\cdot) + \rho \sum_{z=1}^{Z_m} \theta_{z,m} \mathbb{I}(i \in R_{z,m})$, 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 $\widehat{f_\iota}(\cdot) = \widehat{f_{\iota,M}}(\cdot)$.

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 determinant'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.

- **Tuning 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. The best combination of hyperparameters is selected and evaluated on the full training and testing dataset. The metric used to evaluate the cross-validated model in the testing data is the R^2. The tuned hyperparameters chosen are in Table 4.

Table 4 – 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.

The accuracy obtained in the testing set regarding income mobility using the Random Forest and Gradient Boosting algorithms is equal to 69.08% and 66.73%, respectively. For education these equal 76.12% and 77.27%, respectively. This level of accuracy is considered quite good to explain the IM.

3. Empirical Results

Results for the estimated baseline mobility LASSO regressions are presented for income in Table 5 and for education in Table 6 (only the statistically significant determinants 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 determinants 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 5 – Rigorous LASSO Results for Intergenerational Persistence in Income

Dependent Variable: Intergenerational Persistence in Income (IGPI)	
LatinAmericaCaribbean	0.14** (0.06)
cohort60	-0.22*** (0.04)
IGPE	0.29*** (0.11)
Obs.	70
RESET Test	**0.7854**

Note: **, *** 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 determine 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 B1 and B2 in Appendix B, 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.

Interesting evidence also emerges when analysing the relationship between intergenerational persistence in education and intergenerational persistence in income, which appear to have a positive and statistically significant relationship. Narayan *et al*. (2018) identify this result as expected, given that education tends to strongly predict individual lifetime earnings for the parents' and children's generations, so mobility in education should influence income mobility. This view is supported in the argument of Solon (2004), that mobility in either education or income may be positively correlated, since income persistence is a result of endowments that are inherited and preferences of parents when deciding about investing in their children's education. In other words, highly educated parents, with higher income, are able to invest more in children's human capital in comparison with low educated parents, promoting education persistence and, as a consequence of education, income persistence. The empirical positive relationship between the two variables is found in the work of Fletcher and Han (2019) for the USA.

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

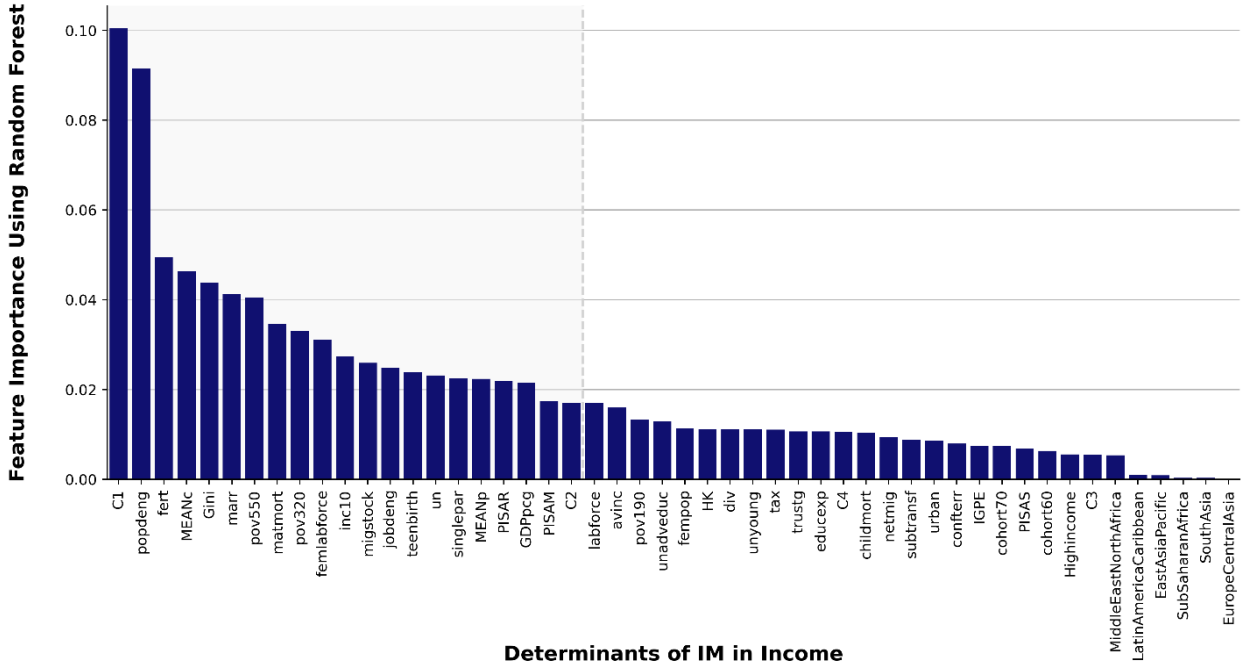
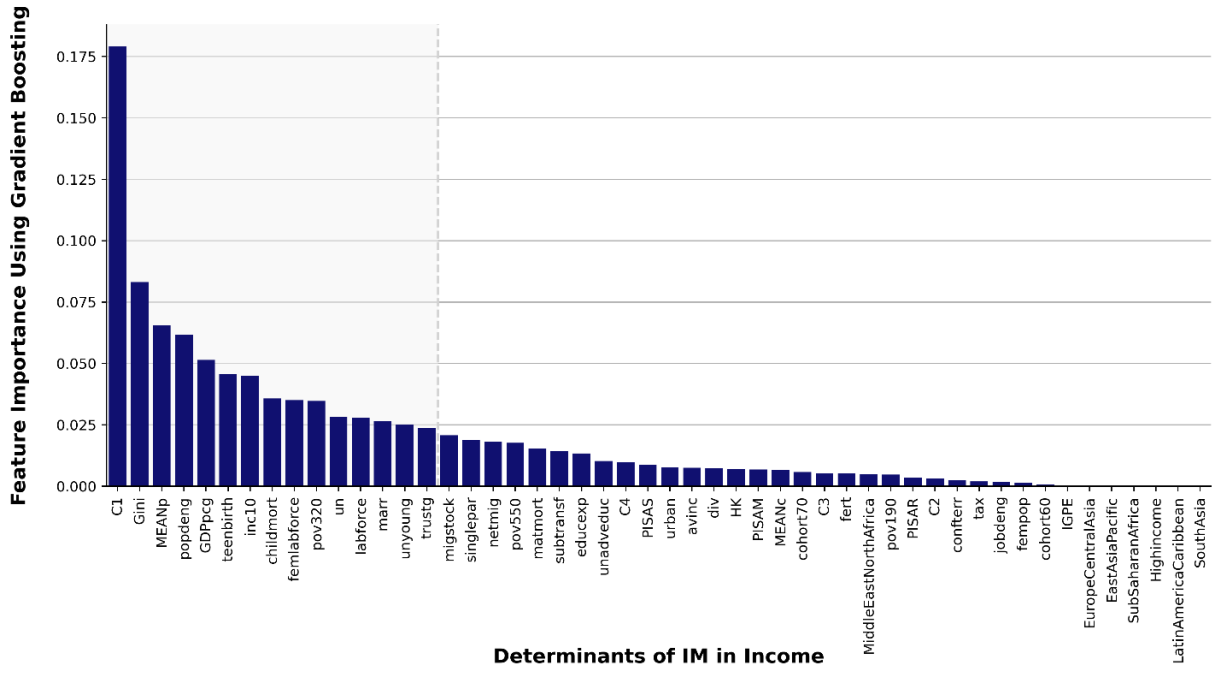


Figure 2 – Feature Importance for IGPI Determinants Using the Gradient Boosting Algorithm



Considering the feature importance plots computed from the machine learning algorithms, there is clear evidence for what the significant determinants 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, in order, 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 Causa and Johansson (2010) for the OECD, Gallagher *et al.* (2019), Chetty *et al*. (2014a,b, 2017, 2020a,b,c), Olivetti and Paserman (2015) and Chetty and Hendren (2018b) regarding the USA, Corak (2019) and Lochner and Park (2022) for Canada, Murray *et al*. (2018) and Deutscher (2020) for Australia, Kyzyma and Groh-Samberg (2020) regarding Germany, 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 6 – Rigorous LASSO Results for Intergenerational Persistence in Education

Dependent Variable: Intergenerational Persistence in Education	
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**

Note: *, **, *** 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.

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 influence 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. Lam and Liu (2019) suggest that mobility in Japan is higher for children of immigrants. Schneebaum *et al*. (2016) find that migrant men are more mobile than native men. 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 influence 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 Determinants Using the Random Forest Algorithm

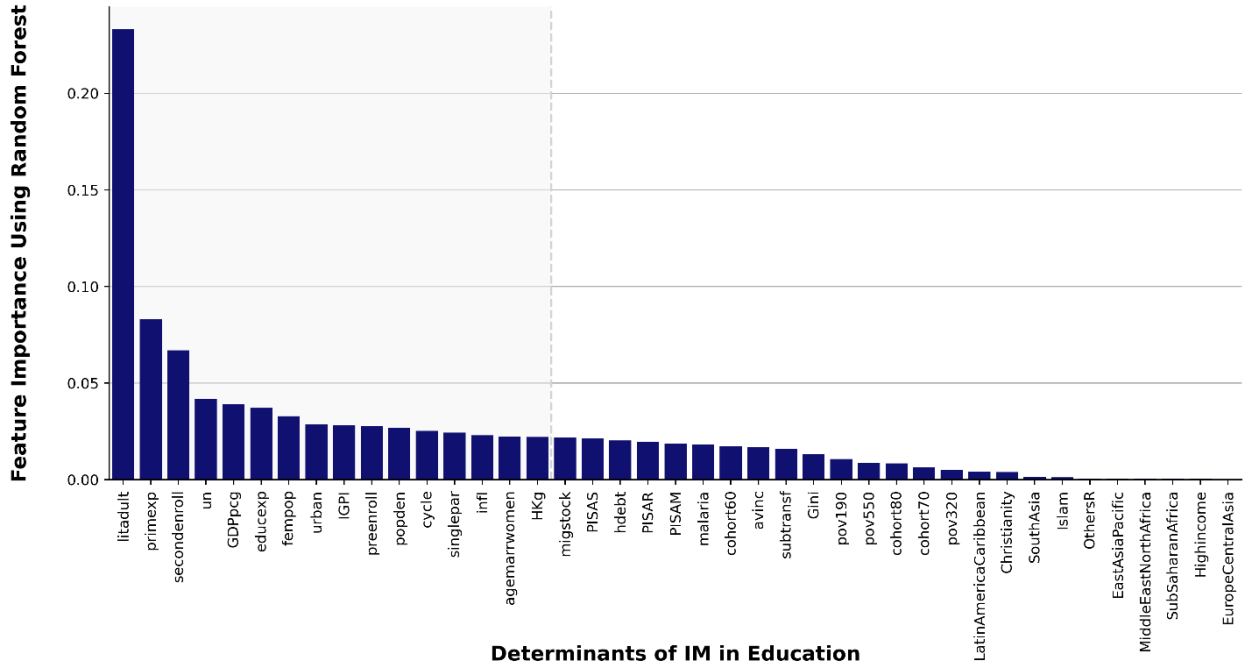
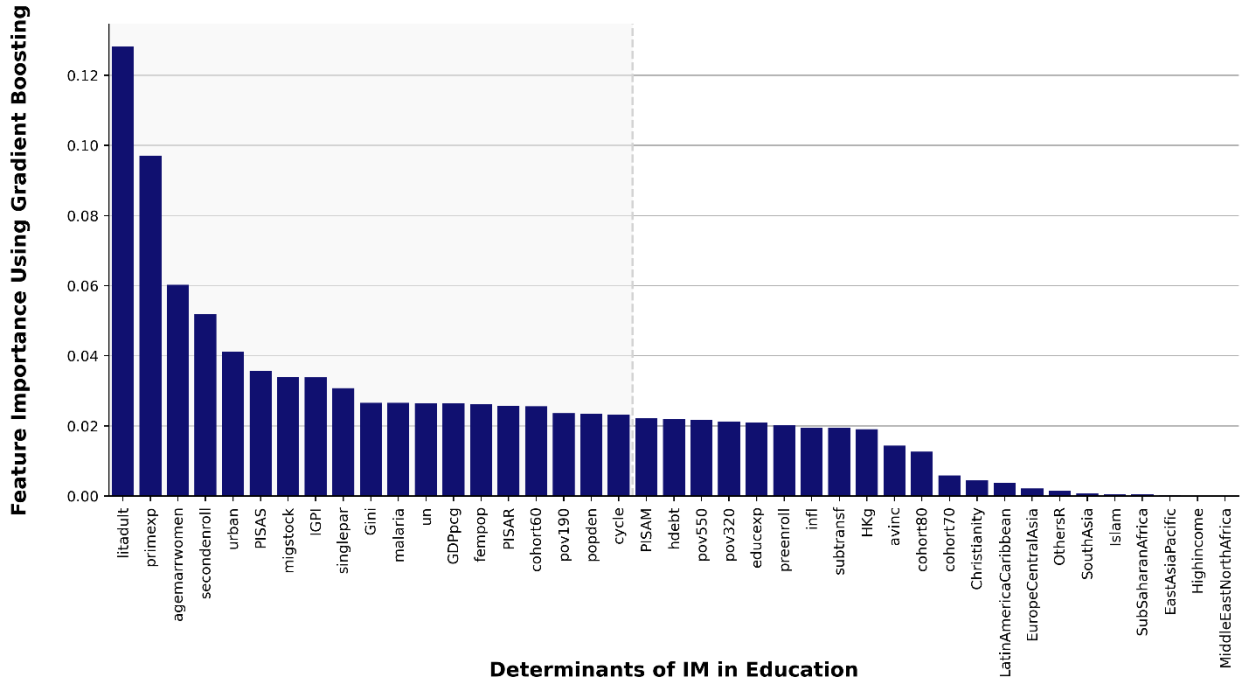


Figure 4 – Feature Importance for IGPE Determinants Using the Gradient Boosting Algorithm



Both machine learning algorithms share the most important intergenerational persistence determinants 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, Emran and Shilpi (2015) and Choudhary and Singh (2017) for India, Lee and Lee (2020) for the OECD, Nimubona and Vencatachellum (2007) for South Africa, Akarçay-Gürbuz and Polat (2016) concerning Turkey, Schneebaum *et al*. (2016) for Austria, Daude and Robano (2015) for Latin America, Urbina (2018) regarding Mexico, Latif (2017, 2018) for Canada, and Neidhöfer and Stockhausen (2018) and Fletcher and Han (2019) for the USA.

Finally, we replace the (time) sample averages by the last period observed as defined in Table 3, for those determinants 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 determinant 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 7 summarizes the set of robust determinants.

Table 7 – Summary of Robust Determinants 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* (-)	- 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*

Note: Effects of robust determinants 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. For education, the migrant stock (migstock) appears to negatively influence intergenerational persistence. Considering the Machine Learning algorithms, we found that the set of robust determinants 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.

4. The Contribution of Each Determinant 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 (which is grounded on an extensive literature review). 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).

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:

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

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 out 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 7. Hyperparameters are again optimized (see Table C1 in Appendix C), training the models to be used in the computation of Shapley values for the testing dataset. This is done with improved accuracy regarding income persistence, while accuracy for the education persistence model appears to be around the same values. The accuracy now obtained in the testing set for income mobility in the Random Forest and Gradient Boosting algorithms is equal to 72.23% and 75.46%, respectively. Regarding education, we have it equal to 74.36% and 76.16%, respectively. Figures 5 and 6 present the bee-swarm plots when considering income persistence and the Random Forest and the Gradient Boosting algorithms, respectively, in which each observation (country) is represented by each dot. The same is done for education persistence in Figures 7 and 8. Our interpretation again relies on the evidence considering both algorithms.

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

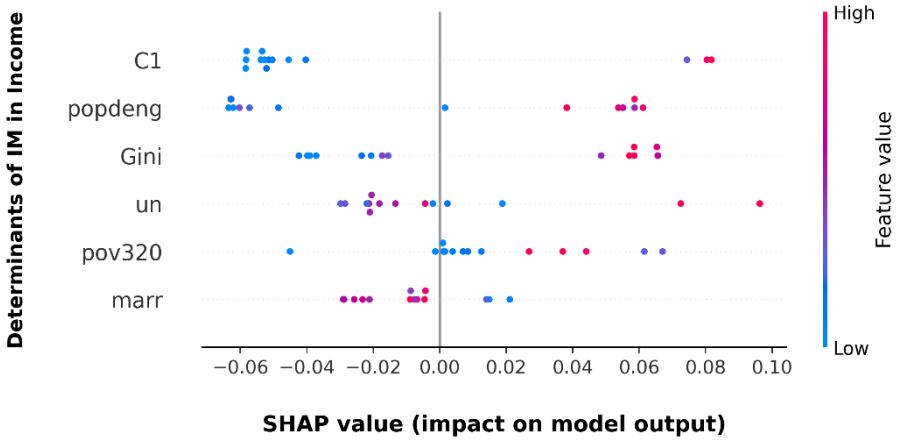
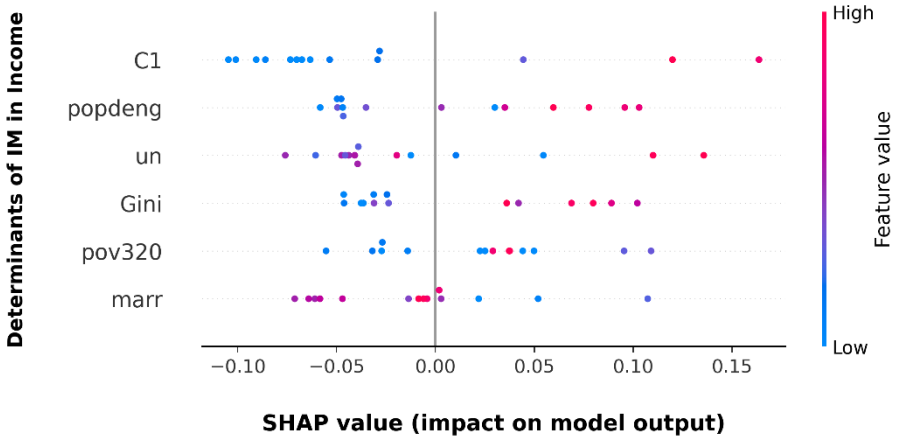


Figure 6 – Feature Contribution for Income Persistence Prediction Using the Gradient Boosting Algorithm



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.

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 determinants 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 (2018b) 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. Murray *et al*. (2018) consider that Australia is more mobile than the USA due to its lower inequality levels. In Kyzyma and Groh-Samberg (2020), German regions with lower inequality present higher mobility for individuals born between 1968 and 1977. 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. Our evidence for the poverty rate considering individuals living on less than $3.20 *per* day (pov320) and the unemployment rate are not clear.

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

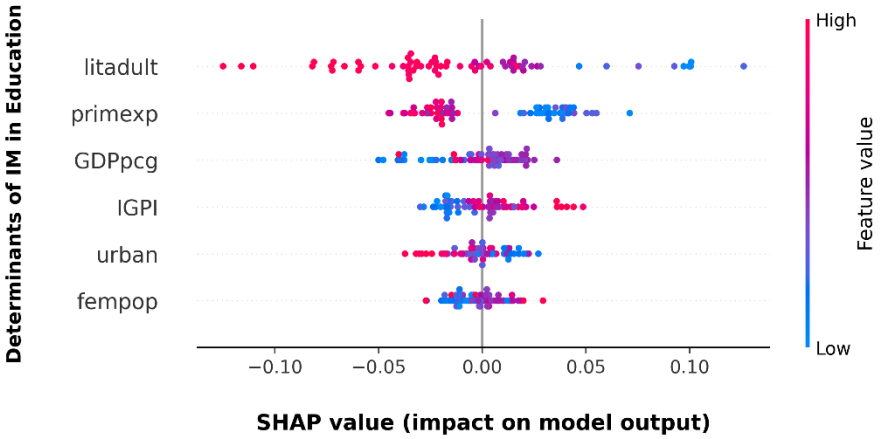
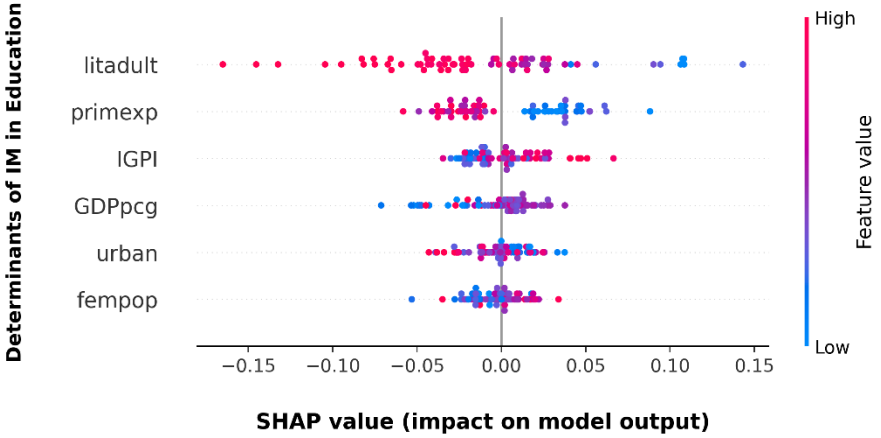


Figure 8 – Feature Contribution for Education Persistence Prediction Using the Gradient Boosting Algorithm



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.

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).

5. Predicting Income Mobility

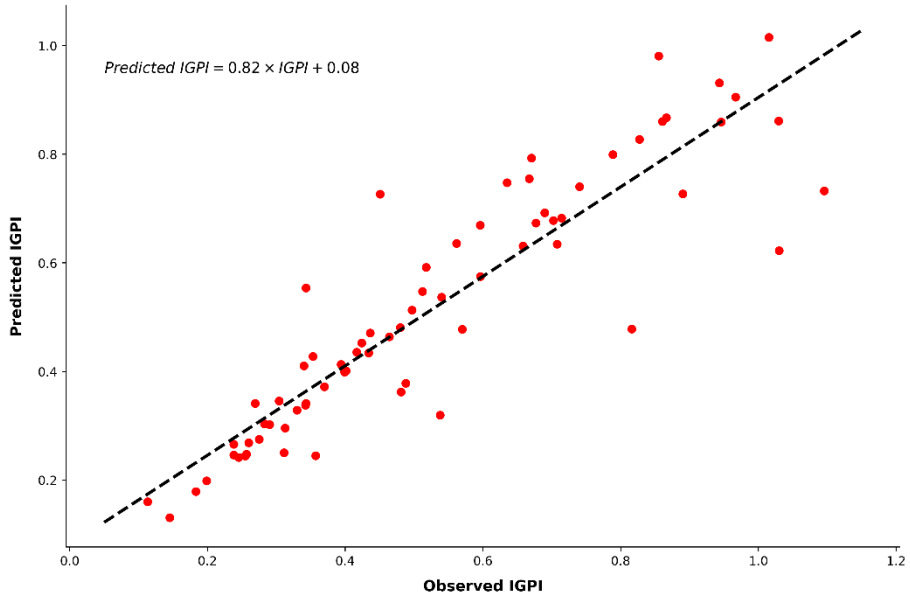
Narayan *et al*. (2018) point out that education mobility is important in its own right and should be an important element of income mobility: incomes persist due to the inherited endowments received from parents and to the investments parents make in children's education. A positive connection may be expected between these two dimensions. Although we were not able to confirm it when analysing income and education mobility determinants through the use of Shapley values, the RLASSO results in Table 5 seem to point to this relationship. 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 determinants 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 is the one with the highest accuracy (75.46%) when compared to either the Random Forest algorithm (72.23%) or even the RLASSO (64.10%).

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 determinants 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 determinants' averages is 1980-2018.

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

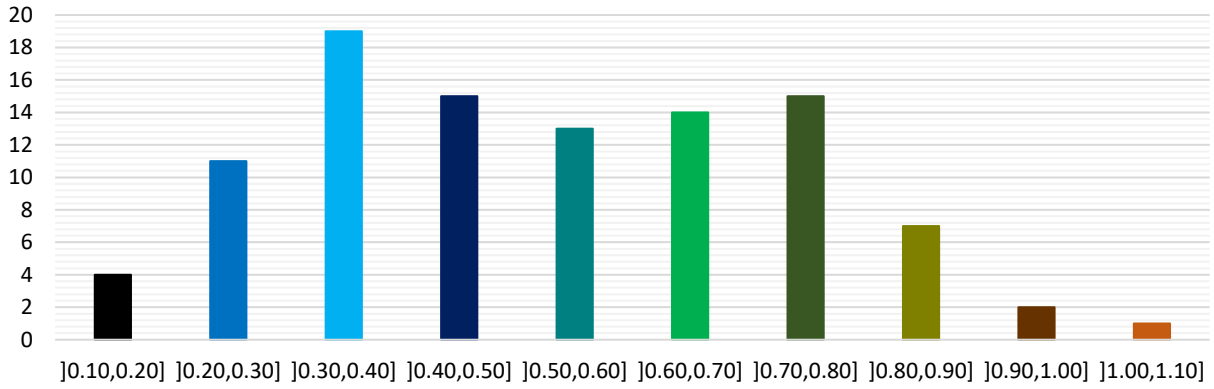
Figure 9 – Predicted IGPI vs Observed IGPI



The evidence presented in the graph shows that the accuracy of our predictions is high. The relationship between the observed and predicted values of income persistence is positive and strong, 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 10, 11 and 12 show the absolute frequency of countries in a set of intervals of income persistence predictions. The corresponding countries are listed as well.

Figure 10 – 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

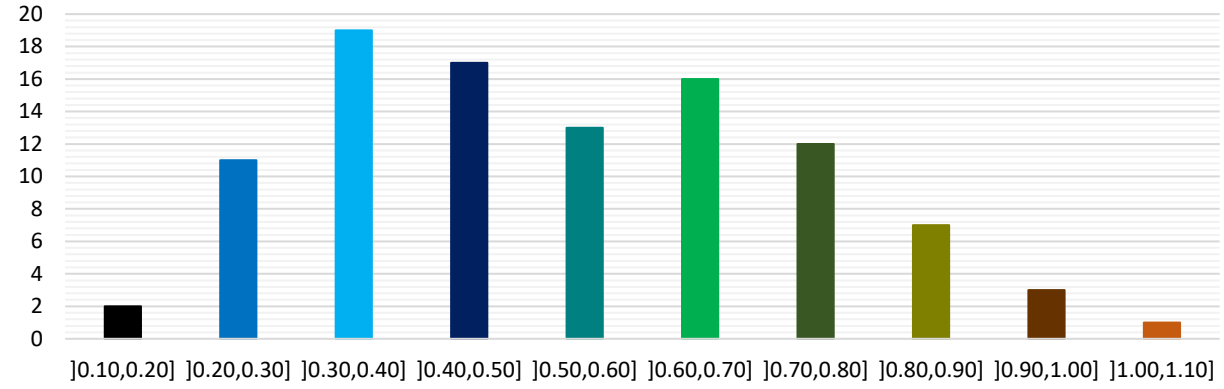
Armenia, Azerbaijan, Belarus, Canada, China, Cyprus, Italy, Korea Rep., Kosovo, Moldova, Netherlands, New Zealand, Portugal, Romania, Russian Federation, Turkey, Ukraine, United Kingdom, United States

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

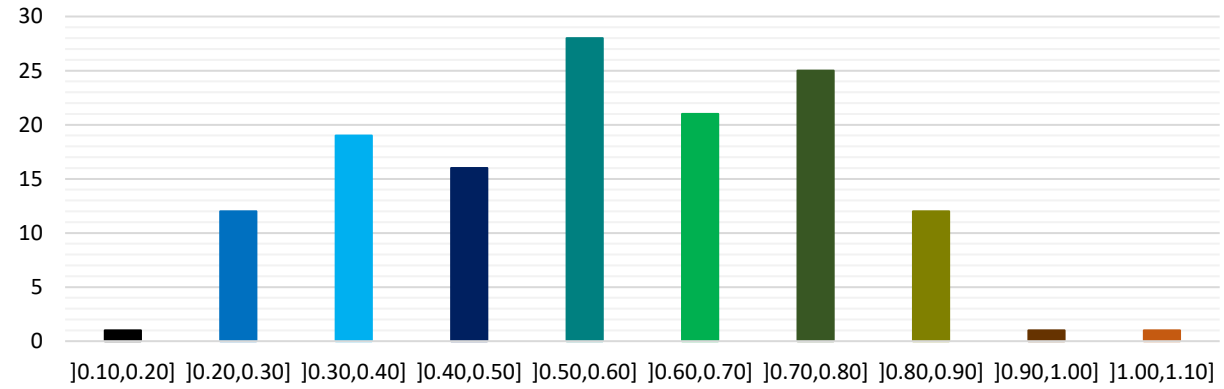
Figure 11 – Predictions of Intergenerational Income Persistence for the 1970 Cohort



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 12 – 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,

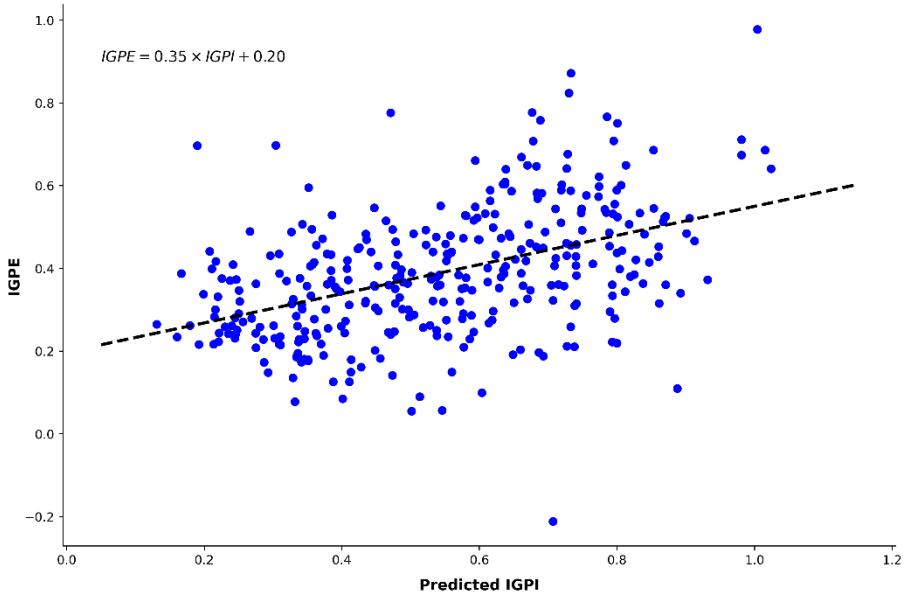
Benin

Guatemala

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 13.

Figure 13 – The Relationship Between 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[8]. Since predicted income mobility is lower in developing economies and it appears to have a positive relationship with education mobility, this 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 influence 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 determinants of income and education IM at a world-wide level. Literature about the determinants 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

[8] 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.

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 determinants. Finally, we predict mobility values for the countries for which only observed values for intergenerational education mobility are available, using the determinants 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 influenced 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, influencing 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 effect 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 effects when focused on early childhood (Herrington, 2015; Blankenou and Youderian, 2015).

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 determinants 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 determinants, countries, or cohorts used, but with the methodology adopted by the World Bank when constructing the database. 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.

References

[1] Abramitzky, R., Boustan, L., Jacome, E., & Pérez, S. (2021). Intergenerational Mobility of Immigrants in the United States Over Two Centuries. *American Economic Review*, 111(2), 580-608. https://doi.org/10.1257/aer.20191586

[2] 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

[3] Ahrens, A., Hansen, C., & Schaffer, M. (2020). Lassopack: Model Selection and Prediction with Regularized Regression In Stata. *The Stata Journal*, 20(1), 176-235. https://doi.org/10.1177/1536867X20909697

[4] Akarçay-Gürbüz, A. & Polat, S. (2017). Schooling Opportunities and Intergenerational Educational Mobility in Turkey: An IV Estimation Using Census Data. *The Journal of Development Studies*, 53(9), 1396-1413. https://doi.org/10.1080/00220388.2016.1234038

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

[6] 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

[7] Azam, M. & Bhatt, V. (2015). Like Father, Like Son? Intergenerational Educational Mobility in India. *Demography*, 52, 1929-1959. https://doi.org/10.1007/s13524-015-0428-8

[8] Bauer, P. & Riphahn, R. (2006). Timing of School Tracking as a Determinant of Intergenerational Transmission of Education. *Economics Letters*, 91(1), 90-97. https://doi.org/10.1016/j.econlet.2005.11.003

[9] Blankenau, W. & Youderian, X. (2015). Early Childhood Education Expenditures and the Intergenerational Persistence of Income. *Review of Economic Dynamics*, 18(2), 334-349. https://doi.org/10.1016/j.red.2014.06.001

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

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

[12] 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

[13] 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

[14] Bergman, P., Chetty, R., DeLuca, S., Hendren, N., Katz, L., & Palmer, C. (2023). Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice. *National Bureau of Economic Research Working Papers No. 26164*. http://10.3386/w26164

[15] Blanden, J., Gregg, P., & Machin, S. (2005). *Intergenerational Mobility in Europe and North America*. Centre for Economic Performance.

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

[17] 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

[18] Borisov, G. & Pissarides, C. (2019). Intergenerational Earnings Mobility in Post-Soviet Russia. *Economica*, 87(345), 1-27. https://doi.org/10.1111/ecca.12308

[19] Causa, O. & Johansson, Å. (2010). Intergenerational Social Mobility in OECD Countries. *OECD Journal: Economic Studies*, 1-44. https://doi.org/10.1787/eco_studies-2010-5km33scz5rjj

[20] 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

[21] Chetty, R. & Hendren, N. (2018a). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *The Quarterly Journal of Economics*, 133(3), 1107-1162. https://doi.org/10.1093/qje/qjy007

[22] Chetty, R. & Hendren, N. (2018b). 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

[23] Chetty, R., Friedman, J., Hendren, N., Jones, M., & Porter, S. (2020a). 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

[24] Chetty, R., Friedman, J., Saez, E., Turner, N., & Yagan, D. (2020b). The Determinants of Income Segregation and Intergenerational Mobility: Using Test Scores to Measure Undermatching. *National Bureau of Economic Research Working Papers No. 26748*. http://doi.org/10.3386/w26748

[25] 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

[26] Chetty, R., Hendren, N., Jones, M., & Porter, S. (2020c). Race and Economic Opportunity in the United States: An Intergenerational Perspective. *The Quarterly Journal of Economics*, 135(2), 711-783. https://doi.org/10.1093/qje/qjz042

[27] 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

[28] 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

[29] 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

[30] Choudhary, A. & Singh, A. (2017). Are Daughters Like Mothers: Evidence on Intergenerational Educational Mobility Among Young Females in India. *Social Indicators Research*, 133(2), 601-621. http://doi.org/10.1007/s11205-016-1380-8

[31] Chu, Y. & Lin, M. (2020). Intergenerational Earnings Mobility in Taiwan: 1990-2010. *Empirical Economics*, 59(1), 11-45. https://doi.org/10.1007/s00181-019-01637-0

[32] Connolly, M., Corak, M., & Haeck, C. (2019). Intergenerational Mobility between and within Canada and the United States. *Journal of Labor Economics*, 37(S2), S595-S641.

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

[34] 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

[35] 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

[36] 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

[37] 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

[38] Emran, M. & Shilpi, F. (2015). Gender, Geography and Generations: Intergenerational Educational Mobility in Post-reform India. *World Development*, 72(C), 362-380. https://doi.org/10.1016/j.worlddev.2015.03.009

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

[40] 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

[41] 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

[42] Fletcher, J. & Han, J. (2019). Intergenerational Mobility in Education: Variation in Geography and Time. *Journal of Human Capital*, 13(4), 585-634. https://doi.org/10.1086/705610

[43] 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

[44] 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

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

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

[47] 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

[48] 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

[49] 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

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

[51] 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.

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

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

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

[55] Kyzyma, I. & Groh-Samberg, O. (2020). Estimation of Intergenerational Mobility in Small Samples: Evidence from German Survey Data. *Social Indicators Research*, 151(4), 621-643. https://doi.org/10.1007/s11205-020-02378-9

[56] 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

[57] Latif, E. (2018). Trends in Intergenerational Educational Mobility in Canada. *The Australian Economic Review*, 52(1), 61-75. https://doi.org/10.1111/1467-8462.12297

[58] Latif, E. (2017). The Relationship Between Intergenerational Educational Mobility and Public Spending: Evidence from Canada. *Economic Papers*, 36(3), 335-350. https://doi.org/10.1111/1759-3441.12177

[59] 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

[60] Lefgren, L., Pope, J., & Sims, D. (2020). Contemporary State Policies and Intergenerational Income Mobility. *The Journal of Human Resources*, 0717-8921R1. https://doi.org/10.3368/jhr.57.4.0717-8921R1

[61] 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

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

[63] Murray, C., Clark, R., Mendolia, S., & Siminski, P. (2018). Direct Measures of Intergenerational Income Mobility for Australia. *Economic Record*, 94(307), 445-468. https://doi.org/10.1111/1475-4932.12445

[64] 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.

[65] 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

[66] Niimi, Y. (2018). Do Borrowing Constraints Matter for Intergenerational Educational Mobility? Evidence from Japan. *Journal of the Asia Pacific Economy*, 23(4), 628-656. http://doi.org/10.1080/13547860.2018.1515005

[67] Nimubona, A. & Vencatachellum, D. (2007). Intergenerational Education Mobility of Black and White South Africans. *Journal of Population Economics*, 20, 149-182. https://doi.org/10.1007/s00148-006-0120-9

[68] 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

[69] 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

[70] 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.

[71] 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.

[72] 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

[73] Štrumbelj, E. & Kononenko, I. (2014). Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems*. 41(3), 647-665. https://doi.org/10.1007/s10115-013-0679-x

[74] 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

[75] World Bank (2018a). *World Development Indicators*. Washington, D.C. World Bank Group.

[76] World Bank (2018b). *World Development Report 2018: Learning to Realize Education's Promise.* Washington, D.C. World Bank Group.

Appendix A

- **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. It is expected that this variable positively influences education mobility as the evidence presented by Alesina *et al*. (2021) for Africa.

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). One should expect that the higher the share of low educated individuals, the more likely incomes are of persisting. At the same time, Blanden *et al*. (2005) shows that for Britain a higher average education years of children is associated with lower income mobility. Therefore, ambiguous results appear in the literature and may be expected.

Human capital index (HK). The human capital index contains information about the mean school years and also returns to education. Different relationships between human capital and income mobility appear in the literature, so we can expect this relationship to occur in either way. Becker and Tomes (1979) develop a model in which parents' utility depends on parents' consumption, as well as on the number and quality (income when adults) of children they have. They show that the propensity to invest in children may be detrimental to intergenerational mobility. In Solon's (2004) model, the author shows that the higher are market returns to human capital and the marginal product associated with human capital investment, the lower mobility will be. In the model of Becker *et al.* (2018), richer parents will invest more in their children's human capital when compared to poorer parents: this is reflected in the persistence of economic status differences between generations. Disproportionate market returns to human capital may reduce mobility between generations. Evidence presented by Murray *et al*. (2018) for Australia and the USA, Connolly *et al*. (2019) for Canada and the USA, and Chu and Lin (2020) for Taiwan point to more human capital being associated with less mobility: the opposite occurs in the work of Lochner and Park (2022) for Canada. Daruich and Kozlowski (2020) find this variable to have ambiguous effects on mobility in the USA. When considering education mobility, we expect their relationship to be negative, such as the one found by Daude and Robano (2015) for Latin American countries and Neidhöfer and Stockhausen (2018) considering Germany, the USA, and the UK.

Parental average education (MEANp). This variable accounts for the parents' average number of education years. We expect this variable to positively influence income mobility, as found by Causa and Johansson (2010) regarding the OECD countries and Gallagher *et al.* (2019) for the USA.

- **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. Public expenditures on education should have a positive effect on income mobility as argued by Solon (2004) and found by Chu and Lin (2020) for Taiwan. The same is expected to occur regarding education mobility, as in Daude and Robano (2015) for Latin American countries, Latif (2017) for Canada, and Urbina (2018) considering Mexico.

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*. We also expect this variable to positively influence education mobility, grounded on the evidence presented for Latin American countries by Daude and Robano (2015), Mexico by Urbina (2018), and OECD countries by Lee and Lee (2020).

- **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). School quality is expected to have a positive relationship with mobility in income in the works of Chetty *et al*. (2014a, 2020a,b) and Chetty and Hendren (2018b) for the USA and Acciari *et al*. (2022) for Italy. The same should occur regarding mobility in education as the evidence presented by Nimubona and Vencatachellum (2007) for South Africa and Hilger (2016) for the USA.

- **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. The unemployment rate should negatively influence income mobility, considering the work developed for the USA by Chetty *et al*. (2020a), for Australia by Deutscher and Mazumder (2020), for Italy by Acciari *et al*. (2022), and for Denmark by Eriksen and Munk (2020). Grounded on evidence presented by Alesina *et al*. (2021, 2023) for African countries, we should obtain the same relationship with mobility in education.

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. The unemployment rate among college or higher educated individuals presents a negative relationship with mobility in income in the work of Acciari *et al*. (2022) for Italy.

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. Youth unemployment rates should also have a negative relationship with income mobility as shown by Acciari *et al*. (2022) for Italy.

- **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. As shown by Acciari *et al*. (2022) for Italy, female labour force is expected to have a positive relationship with mobility in income.

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. Grounded on the evidence presented by Acciari *et al*. (2022) for Italy and Eriksen and Munk (2020) for Denmark, a positive relationship is also expected between this variable and income mobility.

- **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. An economic boom is expected to have a positive effect on education mobility and an economic crisis should harm it, as considered by Urbina (2018) when examining Mexico.

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. In the model of Becker and Tomes (1979) economic growth has ambiguous effects on intergenerational income mobility since it depends on rates of return on investment or the degrees of inheritability. However, Chetty *et al.* (2017) found that lower GDP growth rates are associated with an income mobility decline in the USA. Better macroeconomic conditions are found to have a positive effect on education mobility in the works of Hilger (2016) for the USA, Choudhary and Singh (2017) for India, and Lee and Lee (2020) for the OECD.

- **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. We expect this variable to negatively influence education mobility, grounded on the works of Niimi (2018) for Japan and Lee and Lee (2020) for the 30 OECD.

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. This variable is expected to have a positive relationship with income mobility, as in Deutscher and Mazumder (2020) for Australia. We expect a positive connection between this variable and mobility in education, as suggested by Nimubona and Vencatachellum (2007) for South Africa and Daude and Robano (2015) for 18 Latin American countries.

- **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. It is expected that segregation/poverty rate has a negative connection with income mobility. This occurred in the work of Chetty *et al.* (2014a, 2020b,c) and Chetty and Hendren (2018b) for the USA, as well as in Deutscher and Mazumder (2020) for Australia. The same is suggested by Chetty *et al*. (2014a) regarding education mobility.

- **Location attributes**

Degree of urbanization (urban). This variable corresponds to the share (in %) of total population living in urban areas. Different effects are found in the literature regarding the relationship between income mobility and the degree of urbanization. While Chetty and Hendren (2018b) find mobility to be lower in urban areas in USA counties, Chetty *et al*. (2020a) and Eriksen and Munk (2020) find an ambiguous connection for the USA and Denmark, respectively. Corak (2019) finds a positive relationship when analysing Canada. Different results also occur for the relationship between education mobility and the degree of urbanization. Ambiguous effects occur in the work of

Schneebaum *et al.* (2016) about Austria. Positive relationships appear to exist regarding South Africa as determined by Nimubona and Vencatachellum (2007), African countries by Alesina *et al*. (2021, 2023), Turkey by Akarçay-Gürbüz and Polat (2017), and India by Emran and Shilpi (2015) and Choudhary and Singh (2017).

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. We expect a negative relationship between income mobility and job density, as found by Chetty *et al*. (2020a) regarding the USA.

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). This variable should have a positive connection with income mobility, as in the work of Deutscher (2020) about Australia. The same occurs for education mobility in Alesina *et al*. (2023) concerning African countries.

- **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. The work of Acciari *et al*. (2022) regarding Italy suggests that migration movements may be associated with more income mobility.

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. Different results appear in the literature regarding this variable's connection with income mobility. A positive relationship occurs in the work of Abramitzky *et al.* (2021) and Gallagher *et al*. (2019) for the USA. The opposite is found by Eriksen and Munk (2020) for Denmark. Regarding IM in education, mixed results also exist. While Schneebaum *et al*. (2016) find ambiguous results for Austria, a positive relationship between the share of migrants is suggested in the works of Abramitzky *et al.* (2021) regarding the USA and Lam and Liu (2019), who study Hong Kong.

- **Early childhood development (preenroll)**

Early childhood development can be measured by the gross pre-primary school enrolment (preenroll), 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. Results in the literature point in different directions regarding the relationship between education mobility and early childhood development. While Schneebaum *et al.* (2016) presents ambiguous evidence for Austria, a positive connection is found by Bauer and Riphahn (2006) regarding Switzerland and Daude and Robano (2015) considering Latin American countries.

- **High school enrolment (secondenroll)**

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. We expect this variable to positively influence mobility in education. This occurred in Hilger (2016), who studies the USA.

- **Inflation (infl)**

This variable corresponds to the annual percentage growth rate of the GDP deflator. Inflation and mobility in education are expected to present a negative relationship, as reported by Lee and Lee (2020) regarding the OECD.

- **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). Taxes may have an ambiguous effect on income mobility, as argued by Becker and Tomes (1979) concerning the application of a progressive tax reduction.

- **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. The works developed for Australia by Murray *et al*. (2018), Canada by Connolly *et al*. (2019), and the USA by Bergman *et al*. (2023) and Chetty *et al*. (2020a) suggest that public policies should have a positive relationship with income mobility. The same can be concluded regarding mobility in education in the work of Daude and Robano (2015) considering Latin American countries.

- **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. When inequality is high, income mobility should be low, as considered by Becker *et al*. (2018), and found by Chetty *et al*. (2014a,b, 2017), Olivetti and Paserman (2015), and Chetty and Hendren (2018b) for the USA, Corak (2019) and Lochner and Park (2022) for Canada, Murray *et al*. (2018) for Australia and the USA, Kyzyma and Groh-Samberg (2020) for Germany, and Acciari *et al*. (2022) for Italy. Mobility in education should also be negatively influenced by income inequality, as reported by Daude and Robano (2015) for Latin American countries, Hilger (2016) for the USA, and Lee and Lee (2020) for the OECD countries.

- **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. It is expected that this variable presents a positive relationship with income mobility, as found by Acciari *et al*. (2022) for Italy.

- **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. Differences in income mobility related to differences in within-country geography appear for the USA in the works of Chetty *et al.* (2014a), Olivetti and Paserman (2015), Chetty and Hendren (2018a), Chetty and Hendren (2018b), Lefgren *et al.* (2020), and Chetty *et al*. (2020a); for Russia in Borisov and Pissarides (2019); Canada in Corak (2019); Australia in Deutscher (2020); and Germany in Kyzyma and Groh-Samberg (2020). Cross-country differences are also found: Blanden *et al*. (2005) compare the UK to the USA; Deutscher and Mazumder (2020) compare Australia to the USA; Eriksen and Munk (2020) compare Denmark, USA, and Canada; Helsø (2020) compares the USA with Denmark; Kyzyma and Groh-Samberg (2020) compare Germany, Canada, USA, and Sweden. The same occurs for mobility in education. We have the case of Daude and Robano (2015) for Latin American countries; Causa and Johansson (2010) for the OECD; Neidhöfer and Stockhausen (2018) considering Germany; Fletcher and Han (2019) for the USA; Emran and Shilpi (2015) and Azam and Bhatt (2015) for India and Latin America; and Choudhary and Singh (2017) for India.

- **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). We expect the share of single parents to be negatively connected with intergenerational income mobility as in Chetty *et al*. (2014a, 2020a,c), Chetty and Hendren (2018b), and Gallagher *et al.* (2019) regarding the USA, and Eriksen and Munk (2020) for Denmark. The same is suggested in the work of Alesina *et al*. (2023) regarding African countries and education mobility.

- **Family instability (div)**

Family instability is measured by the number divorces (div) as a share of mean population, *per* 1000 inhabitants and *per* year. This variable should have a negative effect on income mobility, as in the work of Acciari *et al*. (2022) considering Italy.

- **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. This result was found by Eriksen and Munk (2020) for Denmark.

- **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.

Marriage age is expected to positively influence education mobility. The same occurs in Alesina *et al.* (2023) for African countries.

- **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. Daruich and Kozlowski (2020) build a heterogeneous agent life cycle model in which income mobility appears through the choice households make regarding the number of children they have (influencing child's education and future labour earnings, through the availability of resources). The authors' finding is that income mobility is improved by a constant and exogenous fertility, considering the USA.

- **Teen birth (teenbirth)**

This variable corresponds to the percentage share of 15-19-year-old women who are pregnant or have had children. As in the works of Chetty *et al.* (2020a) for the USA and Eriksen and Munk (2020) for Denmark, we expect a negative relationship between this variable and income mobility.

- **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. Child mortality is expected to have a negative relationship with income mobility. This was found by Olivetti and Paserman (2015) for the USA.

- **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. We consider that maternal mortality should negatively influence income mobility, as is the evidence presented by Olivetti and Paserman (2015) for the USA.

- **Gender (fempop)**

Gender differences in income mobility are found in the works of Causa and Johansson (2010) for the OECD, Borisov and Pissarides (2019) for Russia, Acciari *et al.* (2022) for Italy, Chetty *et al*. (2020c) for the USA, Helsø (2020) for Denmark, and Kyzyma and Groh-Samberg (2020) for Germany. These also occur regarding education mobility in Nimubona and Vencatachellum (2007) for South Africa, Alesina *et al.* (2023) for African countries, Emran and Shilpi (2015) for India, Akarçay-Gürbuz and Polat (2016) for Turkey, Schneebaum *et al.* (2016) for Austria, Daude and Robano (2015) for Latin America, Urbina (2018) for Mexico, Latif (2017, 2018) for Canada, and Neidhöfer and Stockhausen (2018) for the USA, Germany and the UK. 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. For example, according to Olivetti and Paserman (2015), when the share of men relative to women decreases, even the "lowest quality" males are desirable and can be matched

with a “high quality” partner, lowering the returns to human capital for men. Hence, persistence in income may increase when the share of female population increases as well.

- **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. Social capital is expected to have a positive effect on income mobility, as reported in Chetty *et al.* (2014a, 2020a) and Chetty and Hendren (2018b) for the USA.

- **Wars (confterr)**

This variable considers the sum of deaths (*per* 100,000) related to war between states, conflicts between civilians, and terrorism. Olivetti and Paserman (2015) show that a lower intergenerational income mobility can be due to factors such as wars, when analysing the USA, and we expect a negative connection between these variables.

- **Religion (religion)**

Alesina *et al*. (2021, 2023) found that there are differences in education mobility associated with heterogeneity in religion in African countries. We therefore 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. We should expect a negative effect of this variable on education mobility, as in Alesina *et al*. (2021) for African countries.

Appendix B

Table B1 – Descriptive Statistics for Determinants 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
preenroll	338	46.05	29.77	0.92	112.76
primexp	338	16.67	7.79	3.41	50.17
secondenroll	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

Figure B1 – Intergenerational Persistence of Income for the 1960 Cohort

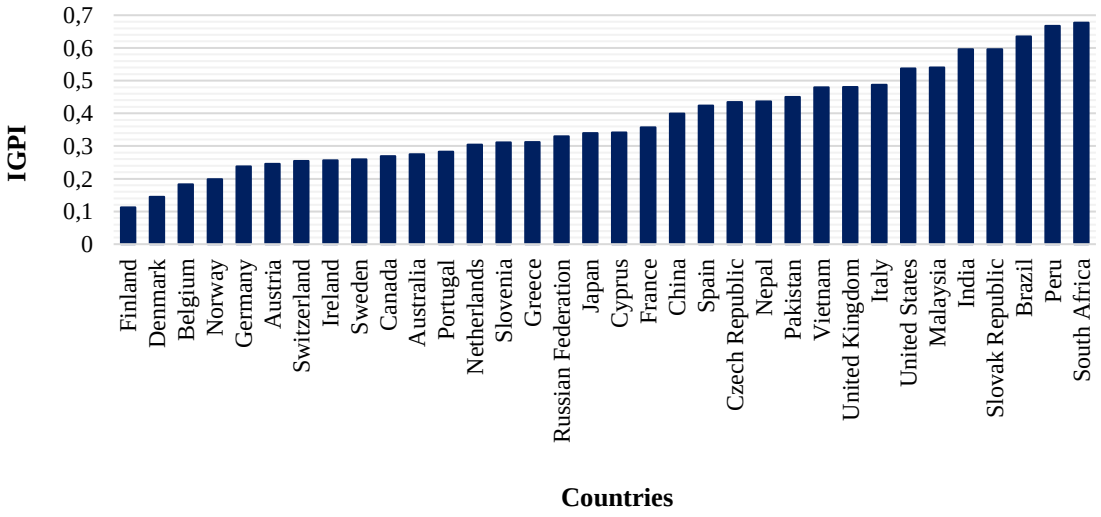
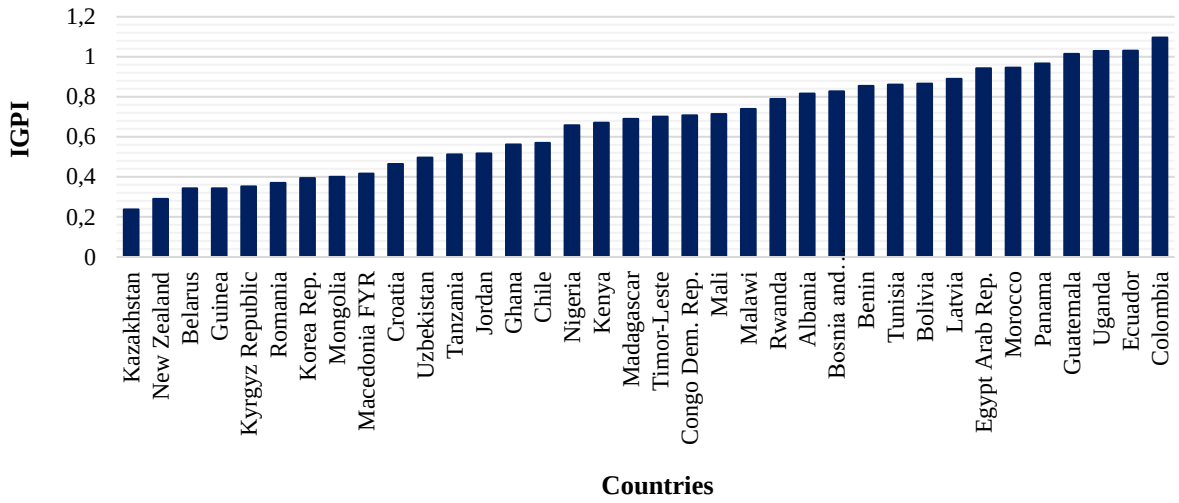


Figure B2 – Intergenerational Persistence of Income for the 1970 Cohort



Appendix C

Table C1 – Hyperparameters Chosen for Random Forest and Gradient Boosting with Robust Determinants 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.