INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Essays on Intergenerational Mobility in Income and Education

Luís Miguel Clemente Casinhas

PhD in Economics

Supervisors:

PhD Alexandra Maria do Nascimento Ferreira Lopes, Associate Professor,
ISCTE – Instituto Universitário de Lisboa

PhD Luís Filipe Farias de Sousa Martins, Associate Professor,
ISCTE – Instituto Universitário de Lisboa

June, 2023

BUSINESS SCHOOL

SOCIAL SCIENCES

Economics Department

Essays on Intergenerational Mobility in Income and Education

Luís Miguel Clemente Casinhas

PhD in Economics

Jury:

PhD Joaquim José dos Santos Ramalho, Full Professor, ISCTE – Instituto Universitário de Lisboa (President of the Jury)

PhD Ana Balcão Reis, Full Professor, Nova School of Business & Economics

PhD Fátima Suleman, Full Professor, ISCTE – Instituto Universitário de Lisboa

PhD Hugo Jorge Correia de Azambuja de Freitas Reis, Economist, Banco de Portugal

PhD Alexandra Maria do Nascimento Ferreira Lopes, Associate Professor, ISCTE – Instituto Universitário de Lisboa

June, 2023

To Margarida.

Acknowledgements

I would like to express my gratitude to Professor Alexandra Ferreira Lopes and Professor Luís Filipe Martins, which accompanied me during this PhD. Besides believing that I was able to achieve this stage, they always had invaluable patience for all the challenges I faced. Without their expertise, the quality of my research work would not be the same.

I want to acknowledge the financial support from FCT - Fundação para a Ciência e a Tecnologia (National Science and Technology Foundation) through grant 2020.04449.BD. This scholarship was the reason why I was able to dedicate most of my time doing research.

Two of the essays composing this thesis were presented and discussed in seminars and conferences. I would like to thank the participants and organizers of the BRU-IUL Research Seminar in Economics, the 29th Computing in Economics and Finance International Conference in Nice, the 16th Conference of the Portuguese Economic Journal in Braga, and the 44th Meeting of the Association of Southern European Economics Theorists. I thank BRU-IUL, which financed these participations. Some of the essays were also submitted to international peered reviewed journals. I want to show my appreciation to the anonymous referees and editors that enriched this research through positive comments on it. The same applies to Professor Ana Balcão Reis on her work on one of the essays.

I am also grateful to João Moura for helping me dealing with data treatment and programming with Python and MatLab. His precious contribution allowed me to develop this work more efficiently. Thanks must also go to my friends (with no specific order) António Urmal, Ricardo Gouveia Mendes, Ricardo Borges dos Santos, and Joana Barros Luís. The same applies to my colleagues, which showed their support during these years, namely Sofia Vale, Vivaldo Mendes, Joaquim Ramalho, Catarina Roseta Palma, and Clementina Barroso.

Besides being the initial step to a research career, this PhD is also the starting point to the teaching profession I want to have. This goal is the reflection of the example of my late aunt, which showed me early the precious role of a good teacher. I praise to my mother for always trying to give me the opportunities she didn't have. She was always there for me and always showed me that education is one of the most important keys to achieve anything. These are two of the fundamental sources of inspiration I have. The same applies to the rest of my family.

My deepest and forever gratitude goes to Margarida, the person to which I dedicate my thesis. There will never be enough words to thank her for the emotional support, motivation, patience, and for keeping my spirits high in the many lows a process like a PhD has attached. I also thank her for forgiving me for not always be there to share important and good moments. Her understanding about the choices I made, with no judgments, during half of our journey together, was crucial to successfully finish the PhD. For all the entertainment and company, for being my shadow in the long days of work, my final thanks go to Amelia, the most crazy and loving cat of this world.

Resumo

Nesta tese analisamos a mobilidade intergeracional do rendimento e da educação. No primeiro capítulo avaliamos os seus determinantes para 137 países, de 1960 a 2018. A mobilidade de rendimento relaciona-se positivamente com a proporção de casados e negativamente com a proporção de crianças com nível de educação inferior ao básico, a taxa de crescimento da densidade populacional e a desigualdade. A mobilidade educacional relaciona-se positivamente com a taxa de alfabetização adulta, a despesa pública com educação básica e a quantidade de migrantes.

No segundo capítulo construímos medidas de mobilidade intergeracional para Portugal, para gerações entre 1968-1988. As mulheres apresentam maiores mobilidade de rendimento e educacional absoluta. Filhos de pais com rendimento baixo apresentam maior persistência de rendimento. A percentagem de indivíduos cuja escolaridade é superior à dos pais ultrapassa os 80%. Filhos de pais com rendimento médio-alto apresentam mobilidade educacional superior. Educação elevada, ocupações que a exigem e rendimento médio-alto associam-se a maior mobilidade. As mobilidades relativas de rendimento e educação relacionam-se de forma positiva mas modesta.

No último capítulo analisamos o conteúdo informacional dos apelidos (ICS) ou nomes próprios (ICF) portugueses como medidas de mobilidade geracional, para indivíduos nascidos em 1956-1995. Sobrenomes e nomes próprios explicam, respetivamente, 14% e 2% das diferenças de escolaridade verificada. O ICS (ICF) é menor (maior) no litoral do país. O ICS correlaciona-se positivamente com a taxa de retenção e desistência no ensino básico, negativamente com o rácio P80/P20, e de forma convexa com o rácio P90/P10 e o Índice de Gini.

Classificação JEL: I24; J62.

Palavras-chave: Mobilidade Intergeracional de Rendimento; Mobilidade Intergeracional de Educação; Determinantes de Mobilidade Intergeracional; Indicadores de Mobilidade Intergeracional Relativa e Absoluta; Conteúdo Informativo dos Apelidos; Conteúdo Informativo dos Primeiros Nomes.

Abstract

In this thesis we analysis intergenerational mobility in income and education. In the first chapter we assess the determinants of intergenerational mobility in income and education for 137 countries, between 1960-2018. Income mobility has a positive relationship with the share of married individuals and a negative relationship with the share of children with less than primary education, the growth rate of population density, and inequality. Education mobility is positively linked with adult literacy, government expenditures on primary education, and the migrants' stock.

In the second chapter we construct mobility measures for Portugal, considering the 1968-1988 cohorts. Women present more income mobility and a greater absolute educational mobility. Children with low-income fathers present higher income persistence. The education of more than 80% of individuals surpasses their fathers' education. Children of medium-high-income fathers present higher education mobility. A high education level, occupations requiring it, and a medium-high income level are associated with more mobility. Income and education relative mobility are weakly positively connected.

Finally, we analyse the role of informational content of surnames (ICS) or first names (ICF) for generational mobility in Portugal, for 1956-1995 cohorts. Surnames and first names explain, respectively, 14% and 2% of the observed differences in educational attainment. The ICS is lower in the country's coast, as opposed to ICF. The ICS correlates positively with the retention and desistance rate in primary and lower secondary education, and negatively with the P80/P20 ratio, while a convex relationship occurs with the P90/P10 ratio and the Gini Index.

JEL Classification: I24; J62.

Keywords: Intergenerational Mobility in Income; Intergenerational Mobility in Education; Determinants of Intergenerational Mobility; Relative and Absolute Intergenerational Mobility Indicators; Informational Content of Surnames; Informational Content of First Names.

Contents

List of Tables

1. Using Machine Learning to Unveil the Determinants of Intergenerational Mobility.

2. From Rags to Riches? Using Survey Data to Estimate Intergenerational Mobility in Portugal.

3. Mother! Father! What Have You Done? The Contributions of First Names or Surnames for Generational Mobility.

Appendix A

Appendix B

Appendix C

List of Figures

1. Using Machine Learning to Unveil the Determinants of Intergenerational Mobility.

2. From Rags to Riches? Using Survey Data to Estimate Intergenerational Mobility in Portugal.

3. Mother! Father! What Have You Done? The Contributions of First Names or Surnames for Generational Mobility.

Appendix A

Appendix C

List of Abbreviations

Abbreviation	Meaning
CPI	Consumer Price Index
ECHP	European Community Household Panel
EU	European Union
EU-SILC	European Union Statistics on Income and Living Conditions
FYR	Former Yugoslav Republic
GDIM	Global Database on Intergenerational Mobility
GDP	Gross Domestic Product
ICF	Informational Content of First Names
ICS	Informational Content of Surnames
ICS30	Informational Content of Surnames held by less than 30 individuals
IGE	Intergenerational Income Elasticity
IGPE	Intergenerational Persistence of Education
IGPI	Intergenerational Persistence of Income
ILO	International Labour Organization
IM	Intergenerational Mobility
IMF	International Monetary Fund
IMF	International Monetary Fund
INE	Instituto Nacional de Estatística
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
IV	Instrumental Variables Estimation
JEL	Journal of Economic Literature
LASSO	Least Absolute Shrinkage and Selection Operator
Min.	Minimum
Max.	Maximum
MSE	Mean Squared Error
NA	Not applicable
NACE	Statistical Classification of Economic Activities
No.	Number
NUTS 3	Nomenclature of Territorial Units for Statistics 3
Obs.	Number of Observations
OECD	Organization for Economic Cooperation and Development
OLS	Ordinary Least Squares
p.p.	Percentage points
PDR	People's Democratic Republic
PISA	Programme for International Student Assessment

RESET	Ramsey Regression Equation Specification
RLASSO	Rigorous Least Absolute Shrinkage and Selection Operator
SHAP	Shapley Additive Explanations Method
Std. Dev.	Standard deviation
TH	Taxable Household
TP	Taxable Person
TSIV	Two-sample Instrumental Variables Estimation
TSLS	Two-stage Least Squares
TSTSIV	Two-sample Two-stage Instrumental Variables Estimation
UK	United Kingdom
UNESCO	United Nations Scientific and Cultural Organization
USA	United States of America
WGBH	Great Blue Hill

Introduction

In this thesis we study intergenerational mobility, which is the extent to which individuals' socioeconomic outcomes are related to the ones of their parents. Comprehending the determinants of intergenerational mobility, may help in mitigating inequality and to promote economic growth, development, living standards, and social cohesion.

Intergenerational mobility is usually measured through income and education. There are two types of mobility. The first one is relative mobility, which regards the relationship between the relative position of children and the relative position of parents in their respective generations, ignoring if the latter is better off than the former. The second one is absolute mobility which only considers if children are worse or better off than their parents, independently of the relative positions they may take. Although these are different concepts, they are equally important. Without relative mobility, a vicious cycle between low mobility and inequality may exist, because inequality in parental investments in children will be reflected in low mobility between generations which, in turn, perpetuate inequality. Besides, human capital is not efficiently allocated, harming productivity, economic growth, and development. In the absence of absolute mobility, living standards cannot improve, damaging social cohesion, by strengthening the sense of unfairness, feelings of frustration, lack of aspirations for the future and hopelessness, leading to social divisions across groups.

Understanding how to deal with the consequences of low mobility through policy making implies uncovering which factors may contribute for it. Therefore, the first of the three main chapters of this thesis consists on an essay entitled "Using Machine Learning to Unveil the Determinants of Intergenerational Mobility", where we assess the determinants of intergenerational mobility in income and education worldwide. Considering that most existing studies only analyse a single country or a limited set of countries, we try to fill this literature's limitation. Using the Global Database on Intergenerational Mobility from the World Bank, we consider a sample of 137 countries with mobility estimates for income and education for the period 1960-2018. To the best of our knowledge, this is the first work with a worldwide outlook on this topic. We complement it with a large sample of potential mobility determinants, grounded on an extensive literature review.

The authors that aim to find any relationship between mobility and other variables usually use simple pairwise correlations. Others (a few) parametrically model the relationships of interest using a limited set of covariates. The results found in the literature may be the consequence of the authors' arbitrariness associated with an *ad-hoc* model selection, either through variables omission, overfitting or incorrect functional forms, leading to estimation biases. We try to tackle this problem by applying machine learning algorithms on this context, namely the Random Forest and Gradient Boosting methodologies. We take advantage of the fact that they consider out-of-sample replicability and present flexibility regarding the relationships being analysed. Besides, we use them to complement the Rigorous

Least Absolute Shrinkage and Selection Operator approach. In the context of machine learning, which only provide features' importance, we are the first ones using Shapley values to analyse the direction of the relationships.

Our results suggest a positive relationship between income mobility and the share of married individuals, while a negative one appears to exist for the share of children that have completed less than primary education, the growth rate of population density, and inequality. When considering the individuals born in the 1960s, mobility in income is higher. Regarding mobility in education, the evidence is that it should be positively influenced by the adult literacy, government expenditures on primary education, and the migrants' stock. Shapley values are not clear about the direction of the relationship between income mobility and unemployment and poverty rates, although these are important factors. The same is verified 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. Lower income and education mobility appears to exist for countries belonging to the Latin America and Caribbean region.

Knowing the most important factors for intergenerational mobility, we predict income persistence for the countries which miss it, but present estimates for educational persistence (but not for income due to the unavailability of data). We find that predicted income mobility has a strong relationship with the existing estimates of income mobility. Besides, income mobility predictions are weakly positively related with existing estimates for educational mobility. It is also confirmed that developing economies are penalized in both mobility dimensions.

The second and third chapters of this thesis are focused on Portugal. The need for a deeper study of intergenerational mobility in the Portuguese economy becomes of utmost importance when international organizations have been pointing it as presenting a high degree of intergenerational persistence. This view has been supported by some of the scarce academic studies on education mobility and by the even most limited studies on intergenerational income mobility.

Our first contribution to the literature for Portugal is the essay "From Rags to Riches? Using Survey Data to Estimate Intergenerational Mobility in Portugal", which constitutes the second chapter of this thesis. This is the first comprehensive study about the status of intergenerational mobility in the Portuguese context, constructing different measures of absolute and relative mobility that complement each other, not only for education, but for income as well. With this purpose in mind we use the Portuguese components of the European Community Household panel and European Union Statistics on Income and Living Conditions and a sample of individuals born between 1968 and 1988. For income, mobility is measured by the intergenerational earnings elasticity, the intergenerational correlation coefficient, the rank-rank slope, the share of individuals earning more than their parents, and the probability that a child born with a low-income father has of reaching the top income level in his or her generation (bottom to top income level probability), complemented by an ordered logit transition matrix. For education, we compute the intergenerational education correlation and the probability that a child

born with a low educated father has of reaching the highest education level (low to high education level), also complemented by an ordered logit transition matrix, and the share of individuals with more education than their fathers.

Addressing income and education mobility allows us not only to analyse their joint behaviour but also to place Portugal in the international context. In other words, it is possible to examine how the country compares with other countries, which use similar data and apply the same methodologies.

Most of the existing literature, at least for income mobility, is focused on the male gender due to measurement issues of females income. In Portugal, there are clear and strong differences between genders in the labour force participation, the type of occupations and their associated pay, the share of individuals earning the minimum wage, and also retirement age. In this work, the measures computed for income and education mobility cover both genders in a joint analysis as well as separately. This occurs for children, while for parents it is only possible to use fathers.

We conclude that gender differences are evident. With the exception of the bottom to top income level probability, men present higher income persistence than women. Portugal stands amongst the countries with the highest relative mobility in income when considering sons. Regarding daughters, the Portuguese economy fits in the middle of existing works estimates for other countries. There is a strong degree of intergenerational mobility for offspring with fathers at the low-income level and upward probabilities are decreasing in the father's income level.

Women present less relative mobility in education than men whereas the opposite occurs in absolute terms. Besides, the share of individuals with more education than their fathers exceeds 80%. As it happened for relative income mobility, Portugal also stands amongst the most relative mobile countries in education, when analysing sons, but for daughters continues to be in the middle of existing estimates. When raised in a low educated environment, full absolute mobility is verified.

Grounded on existing literature that report that individual's characteristics may be associated with different patterns of mobility, we decompose our measures by children/father education, occupation, income levels, and employment status and compare them with our benchmark estimates. We find that individuals with a high education level present greater mobility. Relative income mobility is also higher for offspring with low educated fathers. Accordingly, income mobility is always higher for legislators, senior officials, managers, and professionals' categories, which are occupations requiring more education attainment. Income mobility is also greater if fathers are in the medium-low-income level. It is lower for self-employed individuals. Offspring of clerks show higher education mobility. The same is verified if fathers belong to the medium-high-income level. Both dimensions of mobility are greater for children in the medium income level.

To further analyse mobility in Portugal, we derive the relationship between relative mobility in income and education using the well-known Mincerian equations, through the use of another Portuguese survey, which is the *Quadros de Pessoal* database. This is grounded on the argument pointed by existing works that income mobility is a consequence not only of the inherited endowments from parents, but

also of the investments parents make in children's education. Besides, in the previous decomposition, there are cases where both mobility in income and education are jointly higher or lower, pointing for the same positive direction. We are able to confirm that although positive, this relationship is weak.

The last essay of this thesis is entitled "Mother! Father! What Have You Done? The Contribution of First Names or Surnames for Generational Mobility". This is the first work using surnames and first names in the study of Portugal's mobility in education. We use the 2021 wave of the European Union Statistics on Income and Living Conditions and the Orbis Database, developing the analysis for men aged between 25 and 64 years old.

The use of names in our context finds its roots in a new branch of research that uses a proportion of the population with rare surnames to study intergenerational mobility, with the argument that if an individual has a particular surname, he or she belongs to a given family, given its rarity. As it occurs with conventional studies that do not use surnames to measure mobility, most works implementing surname-based measures consider different cross sections of data. Unlike what occurs in those works, information on Portuguese surnames is only available in a single cross-section of data. Therefore, the study of mobility through surnames in Portugal is only possible by applying a methodology developed by Güell *et al*. (2015), which avoids data requirements for several generations and operates through a single cross-sectional snapshot of the country.

The authors developed a measure defined as the Informational Content of Surnames (ICS), which captures the proportion of the differences in educational attainment that is due to surnames. If surnames explain the variance of educational attainment, this means that being born into a specific family matters for individuals' socioeconomic outcomes, since there are intrinsic inherited family characteristics influencing them. The higher this degree of dependence on the past, the more socioeconomic status is likely to persist. With this purpose in mind, we look only to the subsamples of individuals which have rare surnames, i.e., the ones held by fewer than 30 individuals (ICC30) to increase the likelihood of family relationships' existence.

Additionally, as highlighted by Levitt and Dubner (2006) in their famous book *Freakonomics*, the first name given to children and chosen by parents can also impact their future. Since in Portugal parents have a much higher degree of discretion when choosing their children's first names in comparison with surnames, the same reasoning as before can be applied: the more first names are important for differences in educational attainment, the more important is the choice of parents regarding their children's naming. We build an analogous to the ICS measure, which is defined as the Informational Content of First Names (ICF).

Results show that rare surnames explain 14% of the educational attainment variance, while first names are responsible for only about 2%, indicating that belonging to a specific family weights more in individual's future than parents' choice about first names.

Moreover, we exploit, for the first time, the Portuguese spatial variability in the informational content of first names and surnames. The ICS30 is lower in the coast in comparison with the interior

regions. We account for the role of internal migration, since migrants can have different outcomes in comparison with natives and compute a regional index for the informational content of surnames, Regional ICS30. This measure covers the 50% most common surnames in a given region. Clear differences emerge, in comparison with the ICS30, meaning that migration plays a role, which points to the relevance of this index when comparing regions. The ICF also presents differences across space but these are not pronounced. However, contrary to the ICS, the ICF is lower in the country's interior when compared to the coast.

Finally, we analyse how the informational content of surnames relates with socioeconomic and political regional variables, considering that its spatial variability is twice the one verified for surnames. The evidence suggests that the Regional ICS30 is positively related with higher retention and dropout rate in primary and lower secondary education, while weakly and negatively related with imports. A robust negative linear relationship is found when considering the P80/P20 ratio as a measure of inequality, while a convex connection exists for the P90/P10 ratio, as well as the Gini Index.

The three essays composing this thesis are independent, although they cover the same topic. This means that concepts and literature may be repeated in different parts of the entire document. Different acronyms may be given to the same word/variable across chapters. The same occurs with figures' formatting which is conditional on the programming language and software used. At the end, conclusions are addressed.

1. Using Machine Learning to Unveil the Determinants of Intergenerational Mobility

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

JEL Classification: C26; E24; I24; J62; O15.
Keywords: Intergenerational Mobility in Income; Intergenerational Mobility in Education; Determinants of Intergenerational Mobility; Rigorous Least Absolute Shrinkage and Selection Operator; Random Forest; Gradient Boosting; Shapley values.

1.1. Motivation and Main Findings

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.

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

In this work we use a database with data for 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, due to authors' arbitrariness, may result in estimation biases. This is particularly important in light of the increasing availability of datasets.

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

Our contribution to the literature is the following. Heterogeneity in mobility is conditional on different features, meaning that these features might themselves be potential 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.

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 essay is organized in the following way. Section 1.2 describes our empirical methodology, by defining the variables in our database as well as our statistical approach. Section 1.3 presents and discusses our results. Section 1.4 explores the contribution of each feature for individual predictions of mobility using Shapley values. In Section 1.5 we use our model to predict income mobility for a specific set of countries and study the relationship between income and education mobility. Section 1.6 concludes.

1.2. Empirical Methodology

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

1.2.1. 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.

1.2.1.1. 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 them 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

[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).

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. Therefore, when married women are included in the analysis, their individual earnings may not accurately reflect their economic status. Also, women may be absent from the labour market due to maternity related issues, and therefore, their activity status may be intermittent, which supports the previous view. GDIM (2018) considers both genders for parents and children. For several countries, and for males and females, persistence in income is measured as persistence in individual earnings using an instrumental variable procedure with the Equalchances[3] methodology of 2018. This is not consistent, because if females are considered in GDIM (2018), individual earnings should not be used in their mobility estimates. Adding to this, the values in the database divided by gender are typically equal. All of this leads us to consider that estimates for females may not be properly estimated, which reinforces the exclusive use of males in our work.

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

1.2.1.1.1. Intergenerational Mobility in Income

Intergenerational persistence of income or elasticity (IGPI), being the estimated coefficient from i) either regressing, through OLS, child's earnings on the parents' earnings around the reference age (both in logarithms), or ii) from the final of three sequential steps using four instrument-related estimation methods, namely the two-sample two-stage instrumental variables estimation (TSTSIV) or two-stage least squares (TSLS): two-sample instrumental variables estimation (TSIV) and instrumental variables estimation (IV) are substitute designations for these methods in the literature. First, one regresses income on a list of variables reflecting parents' characteristics such as parents' age and education on a sample, which represents the current parents' population when younger; second, the coefficients reflecting parental education and experience returns are used to predict parental earnings for a reference age; third, regress children's earnings on the predicted earnings for parents. Persistence in income ranges between approximately 0.1 and 1.1, considering all countries and cohorts, meaning that increases in father's

[3] http://equalchances.org.

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.

1.2.1.1.2. 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 1.1.

Table 1.1 – 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

1.2.1.2. 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 1.2 summarizes those determinants and the effects they are expected to have on income and education mobility.[4]

Table 1.2 – 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

Note: NA - not applicable.

(continues in the next page)

[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 (2021), the Global Debt Database from the International Monetary Fund (2018), and the World Values Survey by Inglehart *et al*. (2018).

Table 1.2 – Determinants of Intergenerational Mobility in Income and Education *(continued)*

Determinants	Expected Effect	
	Income	**Education**
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 (primexp).	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
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

Note: NA - not applicable.

(continues in the next page)

Table 1.2 – Determinants of Intergenerational Mobility in Income and Education *(continued)*

Determinants	Expected Effect	
	Income	**Education**
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.

1.2.2. Methodology

1.2.2.1. 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 1.1.

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).

The summary of the periods on which potential mobility determinants are averaged over time for each country is in Table 1.3.

[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.

Table 1.3 – 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 1.3), 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.

1.2.2.2. 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.

1.2.2.2.1. Rigorous Least Absolute Shrinkage and Selection Operator

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

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

[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 1.3 and treating them as the expected value considering the entire time span, we are assuming that the missing values share those same properties. This causes no problems if stationarity is verified.

[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).

(1.1)

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) considering $i = 1, \dots, N$ cross-sections; βs are the model's coefficients and the error term is given by e_i. One may expect correlates to differ by cohort and therefore $IMD_{i,k}$ also include cohort dummy variables. Our baseline estimation will use the mobility measures defined previously and obtained by using the outcomes of the father-son pair, as in most of the literature.

The LASSO shrinkage estimator is used to select and fit the covariates that constitute the model's regressors among all the K 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.

1.2.2.2.2. 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 bootstrap 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 $\widehat{f_{\iota,b}}$ is obtained. The final prediction for a given observation is computed as $\widehat{f_b}(\cdot) = B^{-1} \sum_{b=1}^{B} \widehat{f_{\iota,b}}(\cdot)$.

1.2.2.2.3. 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 5^{th} fold. The metric used to evaluate the cross-validated model in the testing data is the R^2. The tuned hyperparameters chosen are in Table 1.4.

Table 1.4 – 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.

1.3. Empirical Results

Results for the estimated baseline mobility RLASSO regressions are presented for income in Table 1.5 and for education in Table 1.6 (only the statistically significant determinants are reported). Figures 1.1 and 1.2 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 1.3 and 1.4, 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 1.5 – 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

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

Using the Rigorous LASSO estimator for income mobility, we find that a country being part of the Latin America and Caribbean subsample (LatinAmericaCaribbean), the 1960s (cohort60), and intergenerational persistence in education (IGPE) appear to 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 A1 and A2 in Appendix A, 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 1.1 – Feature Importance for IGPI Determinants Using the Random Forest Algorithm

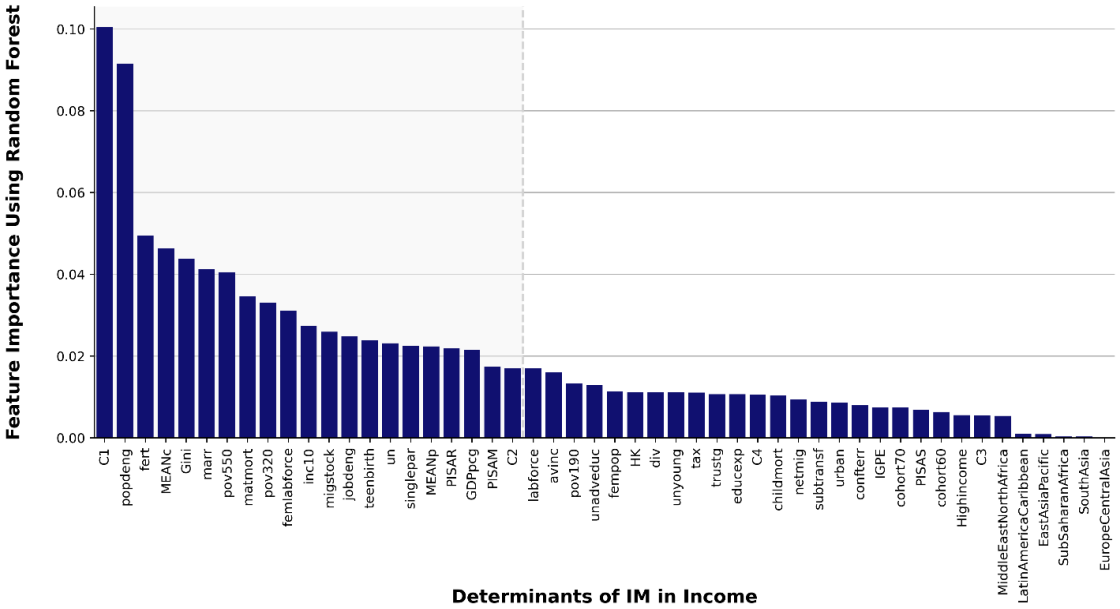
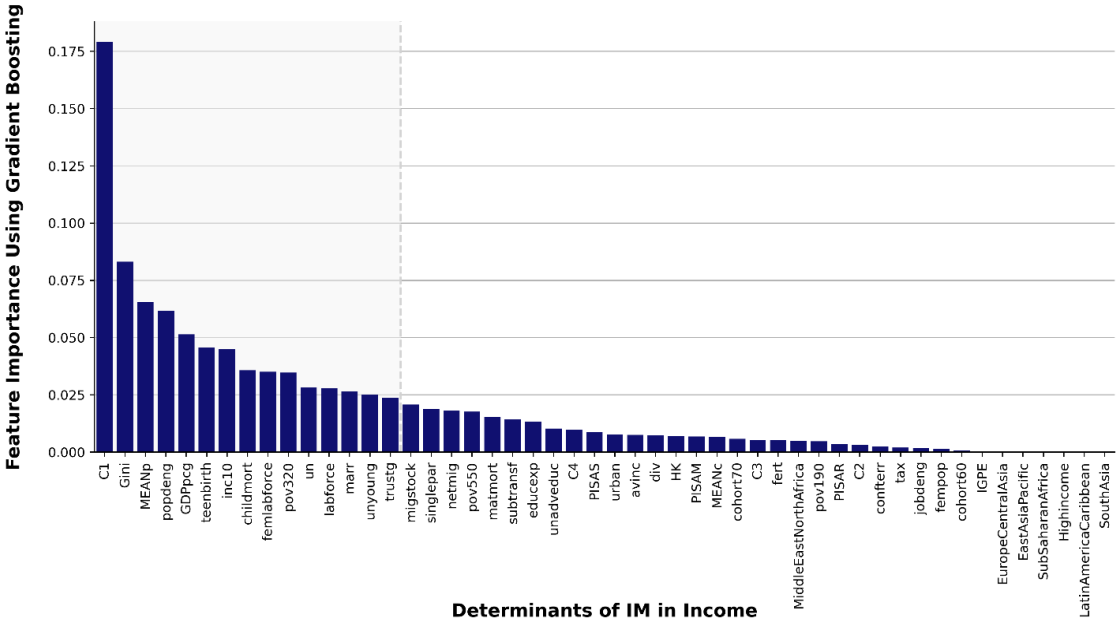


Figure 1.2 – Feature Importance for IGPI Determinants Using the Gradient Boosting Algorithm



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 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 1.6 – Rigorous LASSO Results for Intergenerational Persistence in Education

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

Notes: *, ** and *** 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 1.3 – Feature Importance for IGPE Determinants Using the Random Forest Algorithm

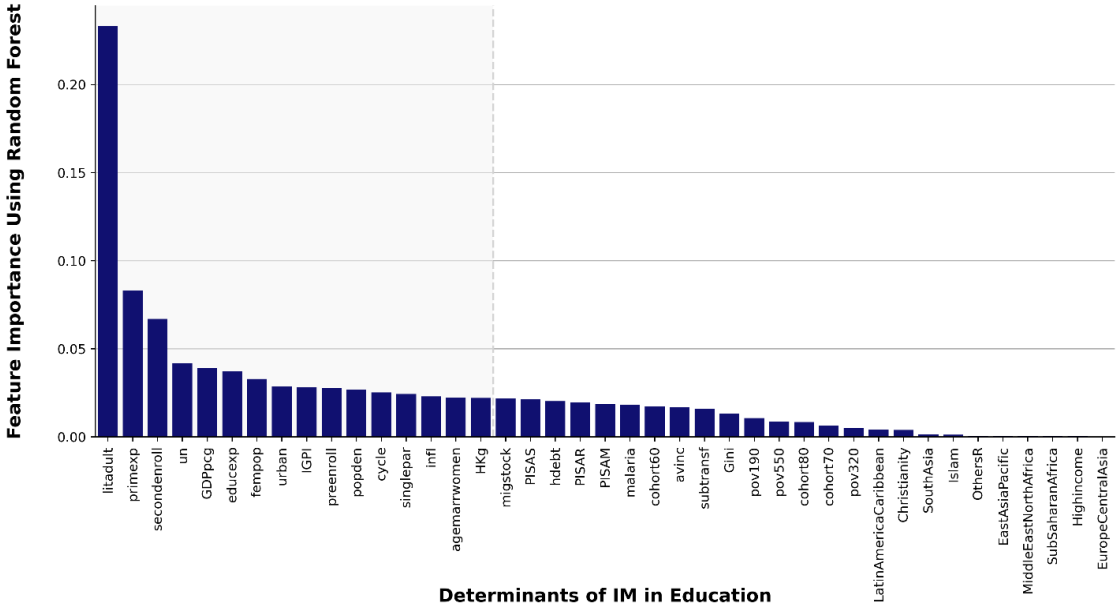
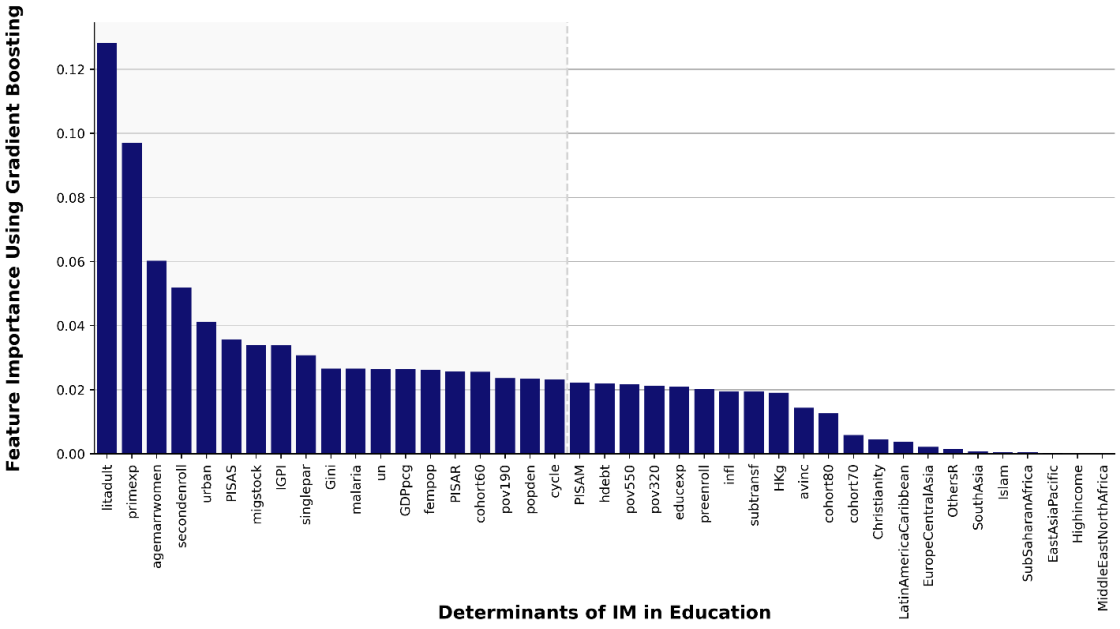


Figure 1.4 – Feature Importance for IGPE Determinants Using the Gradient Boosting Algorithm



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 (2017) 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 1.3, 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 1.7 summarizes the set of robust determinants.

Table 1.7 – 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*

Notes: 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 Random Forest and Gradient Boosting 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.

1.4. 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 reflect 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:

$$\phi_l(g) = \sum_{P \subseteq \{1,\dots,x\} \setminus \{l\}} \frac{|P|!\,(n - |P| - 1)!}{x!} [g(P \cup \{l\}) - g(l)] \tag{1.2}$$

where l is the total number of players, P considers the set of coalitions to which player l can make a marginal contribution, g is the function to obtain the outcome of the game. The same reasoning can be applied to 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 1.7. Hyperparameters are again optimized (see Table A3 in Appendix A), 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 1.5 and 1.6 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 1.7 and 1.8. Our interpretation again relies on the evidence considering both algorithms.

Figure 1.5 – Feature Contribution for Income Persistence Prediction Using the Random Forest Algorithm

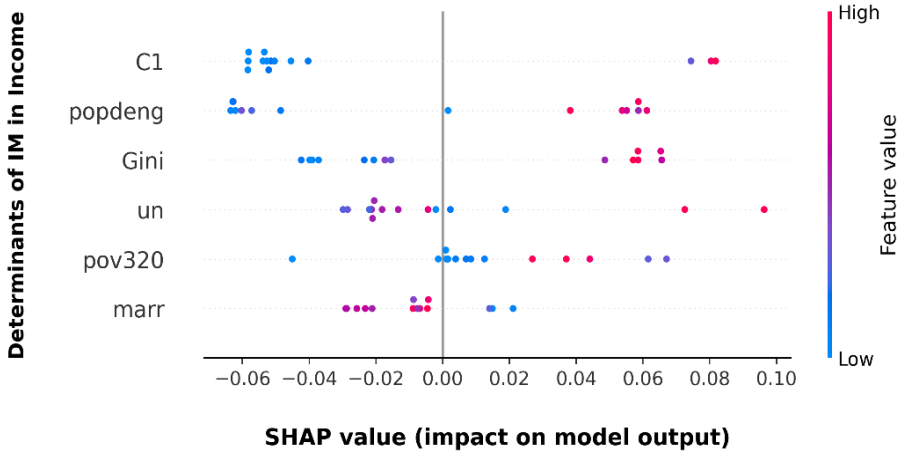
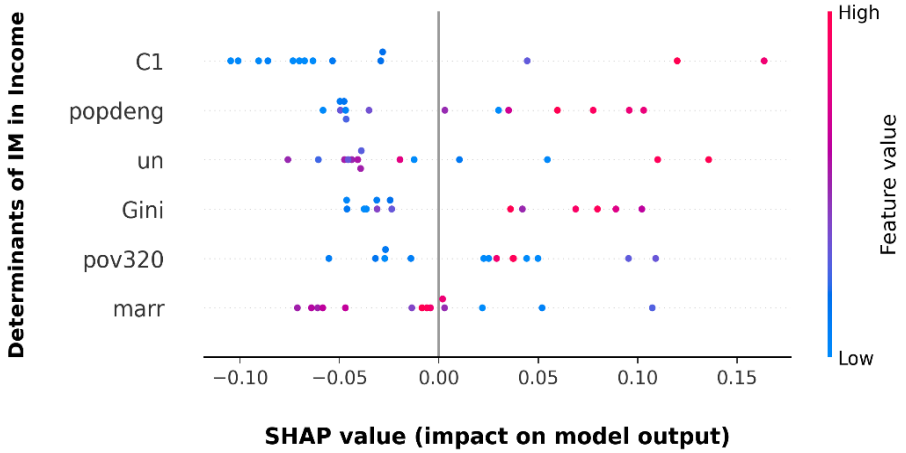


Figure 1.6 – Feature Contribution for Income Persistence Prediction Using the Gradient Boosting Algorithm



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

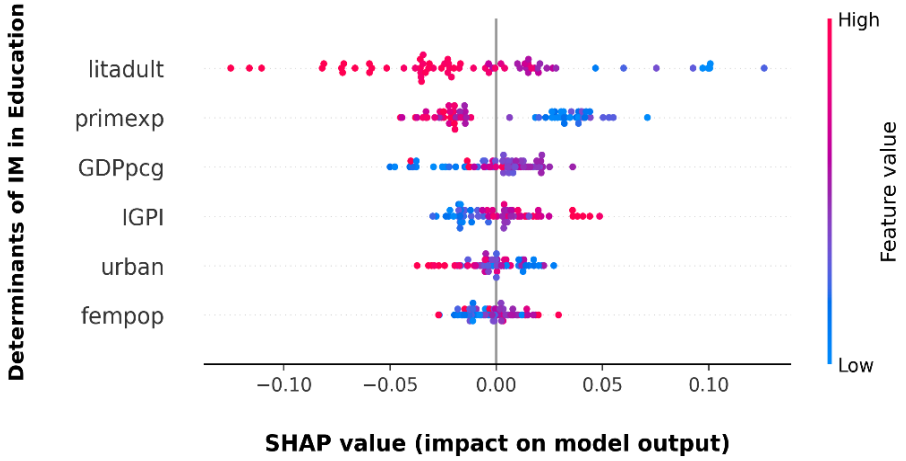
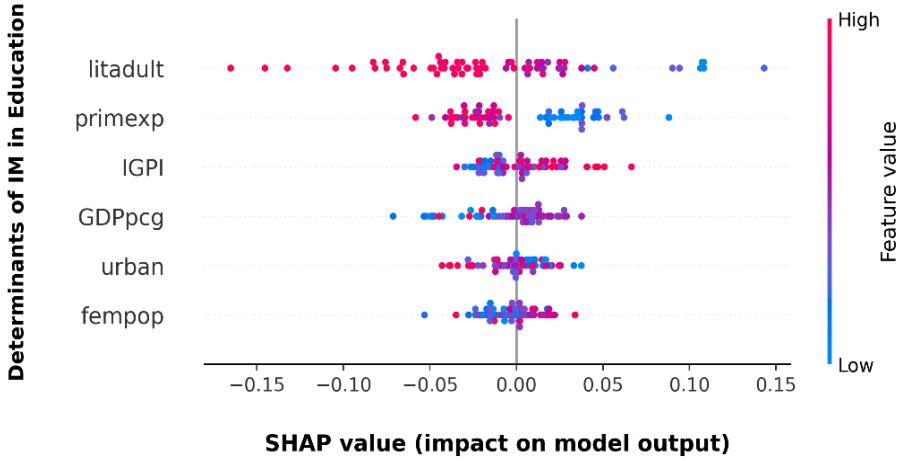


Figure 1.8 – Feature Contribution for Education Persistence Prediction Using the Gradient Boosting Algorithm



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).

1.5. Predicting Income Mobility

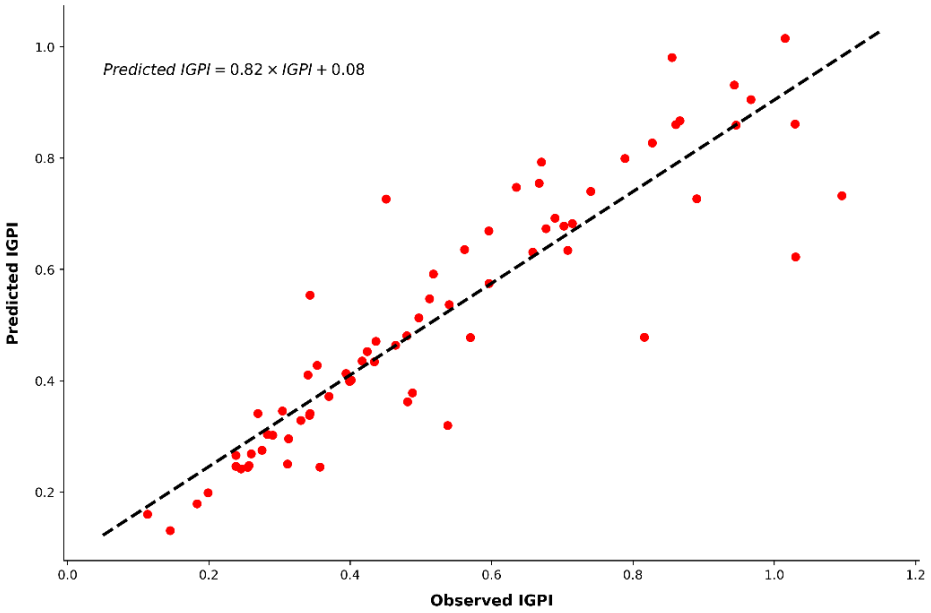
Narayan *et al*. (2018) point out that education mobility is important in its own right and should be an important element of income mobility: incomes persist due to the inherited endowments received from parents and to the investments parents make in children's education. A positive connection may be 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 1.5 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 1.3 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 1.9 plots the observed intergenerational persistence in income values against the predicted ones.

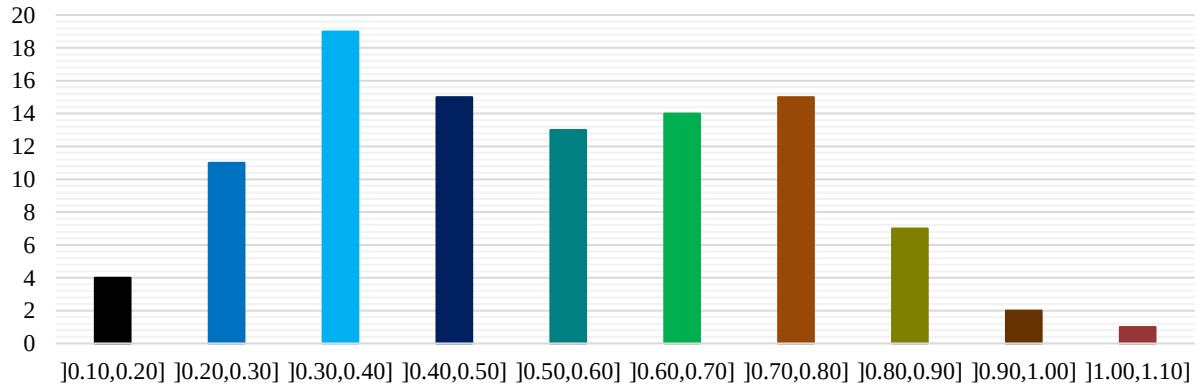
Figure 1.9 – Predicted IGPI vs Observed IGPI



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 1.10, 1.11 and 1.12 show the absolute frequency of countries in a set of intervals of income persistence predictions. The corresponding countries are listed as well.

Figure 1.10 – Predictions of Intergenerational Income Persistence for the 1960 Cohort



Legend:

Belgium, Denmark, Finland and Norway.

(continues in the next page)

Figure 1.10 – Predictions of Intergenerational Income Persistence for the 1960 Cohort *(continued)*

Legend:

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

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

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

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

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

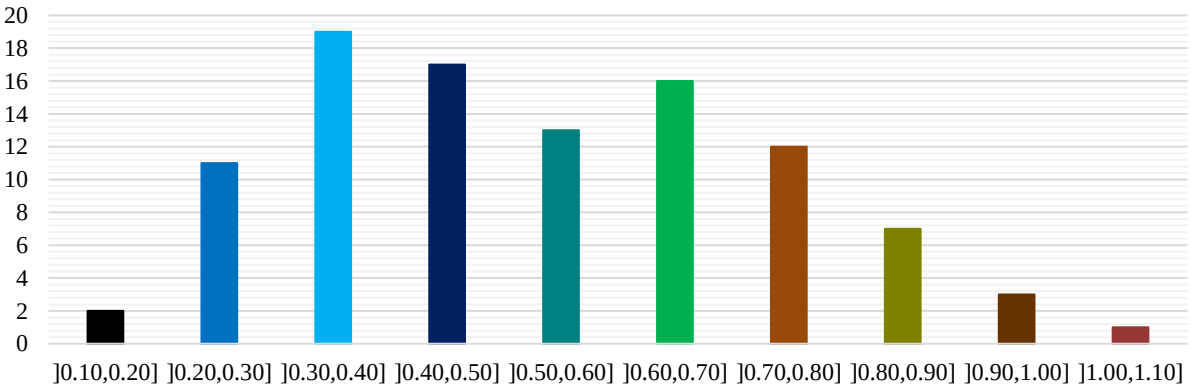
Bolivia, Bosnia and Herzegovina, Brazil, Colombia, Comoros, Ecuador, Egypt Arab Rep., Iraq, Malawi, Pakistan, Peru, Rwanda, Serbia, Timor-Leste and Togo.

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

Morocco and Tunisia.

Guatemala.

Figure 1.11 – Predictions of Intergenerational Income Persistence for the 1970 Cohort



Legend:

Portugal and Slovenia.

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

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

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

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

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

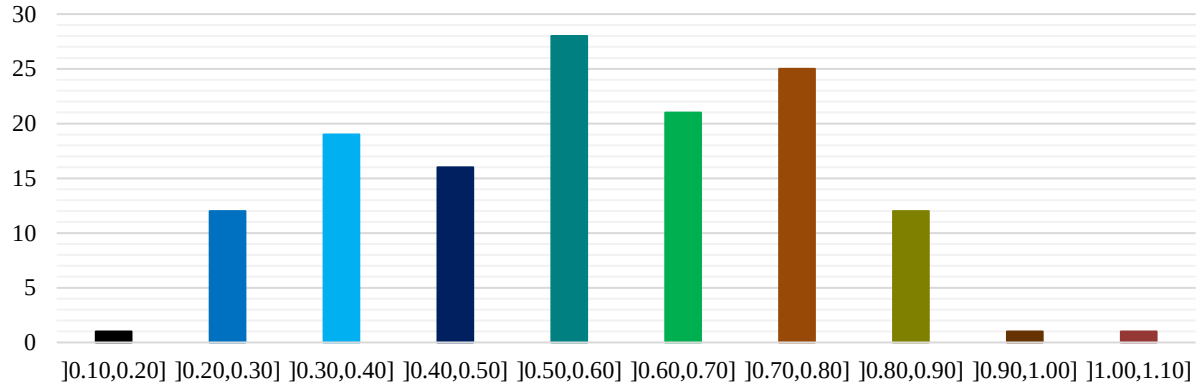
Bhutan, Brazil, Colombia, Comoros, Kenya, Latvia, Malawi, Pakistan, Peru, Rwanda, Senegal and Serbia.

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

Benin, Egypt Arab Rep. and Panama.

Guatemala.

Figure 1.12 – Predictions of Intergenerational Income Persistence for the 1980 Cohort



Legend:

Belgium.

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

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

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

Afghanistan, Bhutan, Djibouti, Guinea-Bissau, Mauritania, Papua New Guinea, Senegal, Sudan, Togo, Tunisia, Uganda and Zambia.

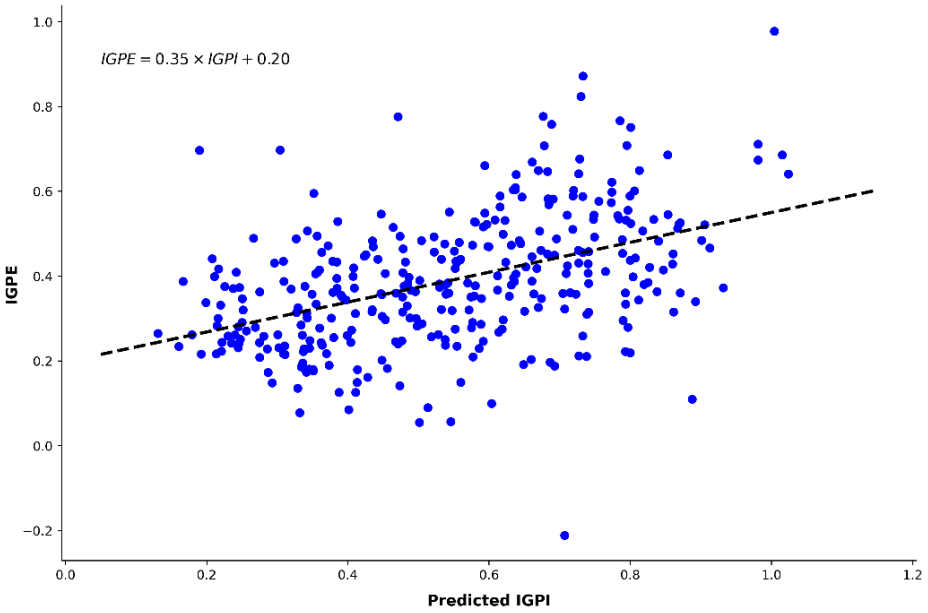
Benin.

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 1.13.

Figure 1.13 – 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

[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.

low-income countries' average student performs more poorly than 95% of high-income economies' average students, when considering literacy and numeracy assessments.

1.6. Concluding Remarks

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 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.

2. From Rags to Riches? Using Survey Data to Estimate Intergenerational Mobility in Portugal

Executive summary: Previous studies about intergenerational mobility for the Portuguese economy find that education and income persistence is very high in comparison with other developed economies. We construct relative and absolute measures of mobility for Portugal, comparing them with existing evidence for this and other countries. We consider the 1968-1988 cohorts and the 1995 and 2019 waves of the European Community Household Panel and the European Union Statistics on Income and Living Conditions, respectively. Overall, women present more mobility in income. Income persistence is high for children with low-income fathers, with upward mobility decreasing in the father income level. Women present a greater absolute educational mobility. More than 80% of individuals have a higher education than their fathers and full upward education mobility exists for children of low educated fathers. Mobility in education is higher for offspring of medium-high-income fathers. Individuals with a high education level, in the medium-high income level or with occupations requiring a higher education level show higher mobility. Additionally, using Mincer equations and the 1994 and 2018 waves of the *Quadros de Pessoal* database (a Portuguese survey), we show a positive but weak relationship between income and education mobility. Policies targeting mobility improvements through quality education, better childhood development, less segregation, and a healthy labour market, together with a robust economic growth should have a positive feedback effect on future generations' outcomes.

JEL Classification: C26; E24; I24; J62; O15.

Keywords: Intergenerational Mobility in Income and Education; Relative and Absolute Intergenerational Mobility Indicators; Two Sample Two Stage Least Squares; Ordered Probit; Mincer Equations.

2.1. Motivation and Main Findings

We construct measures of intergenerational mobility in income and education for Portugal using the 1968-1988 cohorts and the 1995 and 2019 waves of the European Community Household Panel (ECHP) of the European Union Statistics on Income and Living Conditions (EU-SILC), respectively. We study the relationship between income and education mobility, using Mincer equations and the 1994 and 2018 waves of the *Quadros de Pessoal* database, to assess the effects of education mobility on income mobility.

The relevance of the study of these topics for Portugal has been stressed by the findings of national and international organizations for some time now. In a study of the International Monetary Fund (IMF) Clements (1999) identified Portugal as a country where, despite improvements in educational attainment from the 1970s on, the share of individuals having completed upper secondary education in the 25-34

years old age range was 32% in 1996. Portugal was behind the other Organization for Economic Cooperation and Development (OECD) countries in terms of percentage of individuals attaining the secondary education level. Twenty years later, in 2018, the OECD (2019) showed that this rate was 34% for the same age range, below the 41% OECD average and the 44% EU-23 average. A report from the World Bank by Narayan *et al.* (2018) also shows that Portugal is the country with the lowest intergenerational mobility in education from the group of high-income economies, when using survey data for 2014 and considering the 1960s and 1970s cohorts. The Bank of Portugal (2022) used the EU-SILC survey and found that when analysing representative individuals aged 25-59 years old, in 2018 about 73% of those completing tertiary education had parents with that education attainment, and less than 20% of individuals ending up with a tertiary degree had parents with up to basic education.

Overall, studies seem to support the view that low educational attainment is likely to perpetuate and a high education persistence should exist from one generation to another. The same appears to occur when considering income: according to the OECD (2018) a five-generation time window is needed in order for someone who belongs to the 10% poorest population to reach the median income. Additionally, Portugal presents one of the highest rates of female labour force, above the European average (in 2021 it stood at 54%, above the 51.3% of the European Union[9]); and job occupations that pay lower wages and demand lower qualifications are typically occupied by women. The gap in wages between women and men, which in all countries is routinely positive, has been widening in Portugal in comparison with the EU-27 average.[10] Also, women account for the highest share of persons earning the minimum wage, and women typically retire later. Furthermore, the share of the female population completing the highest educational level is higher than amongst men, while the opposite occurs for intermediate levels.

The above evidence presents a dismal prospect for the Portuguese economy in terms of economic growth and development. Lower mobility prevents an efficient allocation of resources, as children with more educated parents are more likely to obtain more education and higher-paying jobs regardless of their innate abilities. A vicious cycle of high persistence and high inequality seems to exist: if higher inequality promotes an unequal distribution of parental investments in children and opportunities, it will harm mobility in the next generation, which promotes inequality. Finally, low mobility negatively influences individuals' perceptions regarding fairness and aspirations, with lower tolerance for inequality and policies to fight it, thereby discouraging growth and social stability.

Our contributions to the literature are clear. First, to the best of our knowledge, we are the first to compute different measures of intergenerational income mobility for Portugal, a developed country in which education and income persistence is still very high. The canonical measure for mobility is the intergenerational income elasticity (IGE) using a two-sample two-stage least squares (TSTSLS) method. Grounded on literature, we predict the unobserved parental income using parents' education, occupation, and managerial position. The other measures we consider are the intergenerational correlation

[9] https://databank.worldbank.org/reports.aspx?source=2&series=SL.TLF.CACT.FE.ZS&country
[10] https://www.pordata.pt/en/DB/Europe/Search+Environment/Table

coefficient, the rank-rank slope, the share of individuals earning more than their parents, and the probability that a child born with a low-income father has of reaching the top income level in his or her generation (we define it as the bottom to top income level probability), complemented by an ordered logit transition matrix.

Second, we compute intergenerational educational mobility measures in relative and absolute terms to complement income mobility measures. We calculate the intergenerational education correlation and the probability that a child born with a low educated father has of reaching the highest education level (denoted by low to high education level probability), also complemented by an ordered logit transition matrix. Additionally, the share of individuals with more education than their fathers is computed.

Third, we disaggregate intergenerational mobility in income and education by own and father's characteristics and uncover which of those may be related with more or less mobility in comparison with the entire sample. These include their education levels, occupation categories, income levels, and status in employment.

Fourth and as final contribution, using Mincerian equations we are also, to the best of our knowledge, the first to assess the effect of relative mobility in education on relative intergenerational persistence in income. This extension to the main benchmark analysis of mobility will enable us to test the argument pointed by existing literature according to which income mobility reflects not only endowments inherited from parents but also the parental investments in children's education, so both income and education mobility should have a positive connection. We check whether or not this relationship exists by using the regression coefficient between children's and parents' years of education and the intergenerational earnings elasticity as measures for education and income persistence.

Our benchmark results make gender differentials evident, since they show that women generally obtain higher mobility in income than men, a finding also reported in the literature. When considering transition probabilities between income levels we observe that there is a strong degree of intergenerational mobility when fathers are at the low-income level, but upward probabilities decrease the higher the father's income level. Our value estimates are according to estimates previously done by other authors. As in the case of income, women have a higher probability of passing from a low to a high education level than men, with previous studies for Portugal reaching lower probabilities than ours. In Portugal the share of individuals with more education than their fathers is greater than 80% and the probability of staying in a low education level, if that is the case of the father, is 0%, a finding that improved relative to other estimates for Portugal. The likelihood that an individual has of reaching or remaining in the high-education level is increasing on the father's education level.

Moreover, when decomposing intergenerational mobility measures by individual (own) and father's characteristics and contrary to what is reported in the literature, we find that individuals with a high education level obtain greater income and education mobility. Also contrary to earlier findings reported in the literature, children whose fathers have a low education level are the ones obtaining higher relative income mobility. Supporting this, we observe that mobility in income is always higher in comparison to

the benchmark for legislators, senior officials, managers, and professionals' categories; occupations that require more education than occupations such as skilled agricultural and fishery workers and plant/machine operators and assemblers, which show lower income mobility. Absolute mobility in education is higher when fathers work as clerks. Mobility in income and education is greater for individuals in the medium-high income level and more absolute mobility in education also occurs when fathers also belong to the medium-high-income level category. However, medium-low-income fathers bring more mobility in income to their offspring. Self-employed individuals obtain lower income mobility when compared to the entire sample.

Finally, complementing the results from the previous decomposition on which there are some cases where income and education mobility are jointly higher or lower, the evidence from the extension using Mincerian equations show that a positive relationship between income and education mobility exists, but we found it to be weak.

This essay is organized as follows. In Section 2.2, the state of the art on intergenerational mobility in Portugal is revised. Section 2.3 details the methodology. In Section 2.4, we describe the data and sample construction. Section 2.5 provides a discussion of the results. Section 2.6 concludes.

2.2. Literature Review

Studies on education mobility in Portugal are scarce. For income, which should have a close relationship with human capital formation, they are almost non-existent. Carneiro (2008) uses transition matrices to show that educational persistence is strong in Portugal: a high share of children who do not complete high school, with fathers who did not complete primary education is verified, while almost no children complete less than high school if their fathers have a university degree. Evidence shows that parental generational differences in educational attainment create differences in opportunities for their children and these differences in educational attainment differ from generation to generation.

Pereira (2010) studies the transmission of higher educational attainment in Portugal through the use of probabilistic regression with data for individuals aged between 18 and 64 years old. The author concludes that parents' education strongly matters for children's higher educational attainment. The likelihood of reaching a higher education degree is greater for individuals born into families with higher education, meaning that low education levels are likely to perpetuate over time. Men generally perform more poorly than women, meaning that they have overall lower mobility.

Bago d'Uva and Fernandes (2017) use multinomial probabilistic models and linear regression analysis to study educational mobility of individuals born from 1940 to 1985. Mobility presented by the 1940 cohort is low. Mobility is generally lower in Portugal when compared to the European Union. Individuals born in the 1970s are more mobile than the 1940s cohort. The increase in mobility was more pronounced for Portugal when compared to the European Union, with the gap between the two being shrinking from the 1970s on. The share of individuals reaching a higher education than their fathers is

greater the younger is the cohort considered. The difference in upward mobility between Portugal and the European Union is mainly due to men, while the share of girls reaching a higher education level than their parents is close to the European average.

Only four works identified include Portugal along with other countries. Comi (2003) uses data on current income for the 1994-1998 period considering 12 European countries. Portugal, Ireland, and the Mediterranean countries are the most relatively persistent in income and education when computing earnings elasticities and the eigenvalues of the educational levels' transition matrices. Relative persistence in income is stronger for the pair father-son when compared to the pair father-daughter. One concern that may appear regarding the author's work is that it does not account for the possible life-cycle effects that may arise from the relationship between current income and life-time income. Also, the marital status of individuals is not considered, as individual income may not reflect individuals' true socioeconomic status. Another study reporting that Portugal is the least mobile country of those belonging to the OECD is Causa and Johansson (2010), who computed wages' persistence (as a proxy of income) as the difference between wage premium and wage penalty. Schneebaum *et al*. (2014) consider 20 European countries and find that for the correlation between parental and children educational attainment, Portugal presents the highest mobility for men, while for daughters is surpassed by France, the Nordic and Anglo-Saxon countries, Greece, Czech Republic, and Poland. In Nybom's (2018) analysis of intergenerational persistence in education from a linear regression on educational outcomes and individuals born around 1980, there is cross-country heterogeneity in high-income countries, with Portugal standing amongst the most persistent, along with Hungary and Uruguay. All these appear to be in line with the international organizations' concerns.

2.3. Methodology

In this section we present the intergenerational mobility measures that are used in this work, for income and education, both in relative and absolute terms. Grounded on methodological fundamentals on mobility measurement, we describe how they should be constructed conditional on the type of data we have. Additionally, we derive a theoretical relationship between relative mobility in income and education, through the use of Mincerian equations.

2.3.1. Measuring Intergenerational Mobility

2.3.1.1. Intergenerational Mobility in Income

The following measures for intergenerational mobility in income are analysed. For relative mobility we have the intergenerational income elasticity, the intergenerational income correlation and the rank-rank slope. For absolute mobility we consider the share of individuals earning more than their fathers and the bottom to top income level probability complemented by an ordered logit transition matrix. The larger

the value of relative mobility measures the lower mobility is, while the opposite occurs with absolute mobility measures.

2.3.1.1.1. Relative Mobility Measures

Intergenerational Income Elasticity (IGE)

The coefficient (β_1) obtained by regressing the log of child i's permanent income (y_i^c) on the log parental permanent income (y_i^p)[11], which is the canonical measure used for relative mobility:

$$y_i^c = \beta_0 + \beta_1 y_i^p + \omega_i \tag{2.1}$$

where $i \in [1; N]$ stands for the pair child-parent, from a total of N pairs. It is an elasticity and therefore interpreted as the child's income percentage change resulting from a one percentage point variation in the parental income. The larger the coefficient is in absolute terms the stronger the impact that parental income has on child's income and vice-versa.

The estimation of equation (2.1) is possible only when at least two generations' lifetime income is available. For this purpose, researchers would need long panels to link parents and children during their entire lives. However, data are usually available in short panels where individuals (parents and children) are observed for a few years only and, therefore, different authors use current income (y_{it}) in period t as a proxy for permanent income (y_i) and assume their relationship to be constant and equal to one. The standard least squares estimator for (2.1) using the current income may have inconsistency problems. In light of the classic errors-in-variables model, this procedure is associated with a measurement error, τ_{it},

$$y_{it} = y_i + \tau_{it}. \tag{2.2}$$

When parental permanent income, i.e., our explanatory variable, is proxied by current income, IGE is subject to an attenuation bias, as pointed out by Solon (1992)[12]. Also, as recent non-classic measurement error research points out, the relationship between permanent and current income changes during the life-cycle of individuals (children and parents). Therefore, grounded on Nybom and Stuhler (2016), equation (2.2) should be generalized to account for the changes in time of this relationship (λ_t), as

[11] Permanent/lifetime income can be defined as the average income during an individual's lifetime (Friedman, 1957).

[12] For the attenuation bias, we have that $plim\ \hat{\beta}_1 = \beta_1 \frac{Var(y_i^p)}{Var(y_i^p)+Var(\tau_{it}^p)} < \beta_1$ and $plim\ \hat{\beta}_1 \to 0$ if $Var(\tau_{it}) \to +\infty$, i.e., beta becomes attenuated (Nybom and Stuhler, 2016).

$$y_{it} = \lambda_t y_i + \tau_{it}.$$

(2.3)

meaning that besides the standard attenuation bias, an associated life-cycle bias should also exist[13].

Our work is no exception in the framework of intergenerational mobility estimates because the survey we use for Portugal contains information only about children's current income. We cannot directly observe parental income as the data are not available, so we use the two-samples two-stage least squares method (TSTSLS). Two samples are needed for this purpose: one for children used in the second step and another for parents used in the first step. In the first step, we predict parental current income ($\hat{y}_{it}^{p}$) by proxying their lifetime income with parental characteristics reported by children: we use parental education, occupation and managerial position. In the second stage, we estimate intergenerational mobility by regressing child's observed income on parental predicted current income. Furthermore, we must account for the uncertainty arising from the regressor used in the second stage (parental income, which is predicted from the first stage, $\hat{y}_{it}^{p}$). Pagan (1984) pointed out that the final steps' coefficients may be in general consistent but the standard errors not. As suggested by Björlund and Jäntti (1997), Piraino (2015), and OECD (2018), we compute second step standard errors by employing a bootstrapping methodology.

For the life-cycle bias, controlling for individuals' age (A) and its square (A^2) to account for life-cycle effects is by itself not sufficient (Jenkins, 1987). One should therefore restrict the sample to the age range in which there should be a stable relationship between current and permanent income and λ_{it} equals one (Haider and Solon, 2006)[14]. The authors found that for the USA economy this should occur between the early thirties and mid-forties (therefore around 40 years old), a result corroborated by Brenner (2010) for Germany, and by Böhlmark and Lindquist (2006) for Sweden. Regarding the attenuation bias, the most common way to deal with it in the literature is to average parents' current income over time (Solon, 1992).

Therefore, the IGE is computed through the following equation:

$$y_{it}^{c} = \beta_0 + \beta_1 \hat{y}_{it}^{p} + \gamma_1^{c} A_{it}^{c} + \gamma_2^{c} {A_{it}^{c}}^{2} + \omega_{it}^{c}$$

(2.4)

Intergenerational Income Correlation

Assuming that $\hat{y}_{it}^{p}$ is orthogonal regarding A_{it}^{c} and ${A_{it}^{c}}^{2}$, we have:

[13] The life-cycle bias (if income profiles change throughout life for both generations) is reflected by $plim\ \hat{\beta}_1 = \beta_1 \lambda_{it}^{c} \lambda_{it}^{p} \frac{Var(y_i^p)}{{\lambda_{it}^{p}}^2 Var(y_i^p) + Var(\tau_{it}^p)}$. Depending on λ_{it}^{c} and λ_{it}^{p}, different results may arise (Nybom and Stuhler, 2016).

[14] If we consider $\lambda_{it}^{c} = \lambda_{it}^{p} = 1$ in $plim\ \hat{\beta}_1 = \beta_1 \lambda_{it}^{c} \lambda_{it}^{p} \frac{Var(y_i^p)}{{\lambda_{it}^{p}}^2 Var(y_i^p) + Var(\tau_{it}^p)}$, we only have to worry about the standard attenuation bias.

$$\beta_1 = \rho_{y_{it}^c, \hat{y}_{it}^p} \frac{sd(y_{it}^c)}{sd(\hat{y}_{it}^p)} \Rightarrow \rho_{y_{it}^c, \hat{y}_{it}^p} = \beta_1 \frac{sd(\hat{y}_{it}^p)}{sd(y_{it}^c)},$$

(2.5)

where $sd(y_{it}^c)$ and $sd(\hat{y}_{it}^p)$ are the standard deviations of the (logged) child's current income and predicted parental current income, respectively, and $\rho_{y_{it}^c, \hat{y}_{it}^p}$ is the partial correlation between those two variables. This correlation is the second measure we compute because since $sd(y_{it}^c) \neq sd(\hat{y}_{it}^p)$, we have an intergenerational income elasticity distinct from the intergenerational income correlation, $\beta_1 \neq \rho_{y_{it}^c, \hat{y}_{it}^p}$. In other words, we adjust the elasticity to changing inequality across generations.

Rank-Rank Slope

Dahl and DeLeire (2008) suggest another measure of relative intergenerational income persistence, which is the rank-rank slope, adopted also by Chetty *et al*. (2014a). It may be computed by first rank children and parents in their respective permanent income percentiles' distribution. Second, for each parental income percentile rank $r(y_i^p)$, obtain the average children's income percentile ranks, $\bar{r}(y_i^c)$. Third, regressing it against parental income percentile ranks, as follows:

$$\bar{r}(y_i^c) = \kappa_0 + \kappa_1 r(y_i^p) + \psi_i.$$

(2.6)

The resulting coefficient (κ_1) measures the relationship between the positions children and parents have in their respective income distributions. As with the intergenerational income elasticity, the greater is the coefficient the greater intergenerational persistence will be, and vice-versa, in absolute terms[15].

We rank the predicted values for parental income, $r(\hat{y}_{it}^p)$. Then, for each one, there is a given number of corresponding children about which we observe their percentile income ranks and compute the average, $\bar{r}(y_{it}^c)$. We should face the same constraints as before in terms of income (mis)measurement. Therefore, we should consider the strategies explained above to overcome both the life-cycle and the attenuation measurement-related issues, to obtain more precise estimates of permanent income through current income, following Chetty *et al*. (2014a). Equation (2.6) will therefore be rewritten as

$$\bar{r}(y_{it}^c) = \vartheta_0 + \vartheta_1 r(\hat{y}_{it}^p) + \varpi_{it},$$

(2.7)

We estimate equation (2.7) through OLS.

[15] Chetty *et al*. (2014a) argue that the rank-rank slope and the intergenerational income correlation have a close relationship, since they are scale invariant. This does not occur with the intergenerational income elasticity, because inequality should be different across generations. When inequality is greater for the child's generation, an increase in parental income may have a greater effect on children's income when compared to a scenario where inequality is lower. In other words, the rank-rank slope and the intergenerational income correlation are not affected by changes in inequality, while the intergenerational income elasticity is.

2.3.1.1.2. Absolute (Upward) Mobility Measures

Besides looking at relative mobility, one should be interested in measuring absolute upward mobility as well[16]. As Chetty *et al*. (2014a) argue, while improvements in relative mobility may occur at the expense of rich people's income being harmed, improvements in absolute mobility for a given level of income, *ceteris paribus*, should result in a welfare improvement according to the Pareto Principle. This is the same as saying that, holding other things constant, absolute upward mobility *de facto* reflects beneficial changes in income of individuals from a given background. We follow their work and compute three main measures of absolute upward mobility.

Share of Individuals Earning More than their Parents

The first measure of absolute upward mobility suggested by Chetty *et al*. (2014a) is the share of individuals whose income exceeds their parents' income in real value.

Bottom to Top Income Level Probability

Following Chetty *et al*. (2014a), the other measure one can use for upward absolute mobility is the bottom to top quintile probability, which is the probability that children whose parents are in the bottom quintile of the parental income distribution have of reaching the top quintile of the children's income distribution when adults. This would be the well-known "American Dream". We measure in this way the bottom to top income level probability because, as mentioned above, we are unable to construct percentile ranks for parents. This also prevents us from transforming data into quartiles or quintiles. Therefore, we consider a specific cell of the Ordered Logit Transition Matrix, which we describe below.

Suppose that we assign each child's income level $inclev_i^c$ in one specific category, i.e., we have $inclev_i^c \in \{1,2,\dots,H\}$ where H denotes the number of possible income categories, which will be defined later in this work: the same is considered for the parental income level categories.

The ordered logit transition probability will be estimated by

$$Pr(inclev_i^c = h|inclev_i^p) = \begin{cases} G(c_1 - \Psi inclev_i^p), if\ h = 1 \\ G(c_h - \Psi inclev_i^p) - G(c_{h-1} - \Psi inclev_i^p), if\ 1 < h \leq H-1 \\ 1 - G(c_{H-1} - \Psi inclev_i^p), if\ h = H \end{cases} \tag{2.8}$$

with the cumulative distribution function of the logistic defined by $G(c_h - \Psi inclev_i^p) = \frac{e^{c_h - \Psi inclev_i^p}}{1+e^{c_h - \Psi inclev_i^p}}$.

Ψ is estimated using the maximum likelihood estimator. The bottom to top income level probability is given by $Pr(inclev_i^c = H|inclev_i^p = 1)$, i.e., it corresponds to the probability that a child with parents classified as low income has of becoming classified as a high-income level earner.

[16] We acknowledge the possibility of downward movements, but the focus should be on the upward direction, as it is connected with higher income growth and shared prosperity (GDIM, 2018).

2.3.1.2. Intergenerational Mobility in Education

We now present the following measures for intergenerational mobility in education. For relative mobility in education we have the intergenerational education correlation. Regarding absolute mobility in education we use the share of individuals with more education than their fathers and the low to high education level probability complemented by an ordered logit transition matrix. As it occurred with income, the larger the value of relative mobility measures the lower mobility is, with the opposite occurring with measures of absolute mobility.

The preferred measure in the literature of relative intergenerational mobility in education is analogous to the relative mobility measure used for income and consists of the coefficient obtained by regressing the total years of educational attainment of children on the total years of education of parents. However, our data characteristics do not allow us to compute it[17].

2.3.1.2.1. Relative Mobility Measures

We rely on the Pearson correlation between parental and child's education levels to measure relative mobility in education:

$$P = \frac{\sum_{i=1}^{N}(e_i^c - \bar{e}^c)(e_i^p - \bar{e}^p)}{\sqrt{\sum_{i=1}^{N}(e_i^c - \bar{e}^c)^2}\sqrt{\sum_{i=1}^{N}(e_i^p - \bar{e}^p)^2}}, \tag{2.9}$$

where e_i^c is a variable for the ordered education levels of children, e_i^p is a variable for the ordered education levels of parents, and the respective average education levels in the sample are $\bar{e}^c = \frac{1}{N}\sum_{i=1}^{N} e_i^c$ and $\bar{e}^p = \frac{1}{N}\sum_{i=1}^{N} e_i^p$.

The coefficient ranges between -1 and 1. From its sign it is possible to infer if we have positive or negative monotonic relationships between the education levels of parents and children, with 0 meaning that no such type of correlation should exist. The closer the coefficient is to the extremes, the stronger the relationships are, while the opposite occurs if it is near zero.

2.3.1.2.2. Absolute Mobility Measures

In order to measure mobility in education in absolute terms, two measures are considered. The first is the share of individuals with a higher education level than their fathers. The second is the probability of low to high education level, which corresponds to the probability children have of reaching the highest

[17] It would only be possible if we had information on both parents' and children's educational attainment, expressed in completed years of education. However, that is not considered in the surveys we use. Instead, educational attainment is reported in categories of completed education levels: the disaggregation is not the same for both generations. Therefore, by making both categorizations comparable and attributing them years of education, we could lose information in the end.

education level conditional on the father's education being the lowest one. This corresponds to a specific cell of the Ordered Logit Transition Matrix described below.

Similar to the case of income levels, we model the probability of children having attained a specific observed category in terms of education, e_i^c, conditional on the observed educational category of their parents e_i^p. Suppose that for the educational levels of children we have $e_i^c \in \{1,2,\dots,M\}$ where M denotes the number of educational categories we have for our dependent variable: the same categories are considered for the case of parents. We have an index model for parental educational attainment described as

$$e_i^{*c} = \Theta e_i^p + \xi_i^p, \tag{2.10}$$

where e_i^{*c} is an unobserved latent measure of the years of education of children and e_i^p is a variable for the ordered education levels of parents. Θ is the regression coefficient associated with the explanatory variable, estimated using maximum likelihood. ξ_i^p is the error term, which follows a logistic distribution. Furthermore, the latent variable crosses specific thresholds, t_m, which are also unknown, such that:

$$e_i^c = \begin{cases} 1, \text{if } e_i^{*c} \le t_1 \\ m, \text{if } t_{m-1} < e_i^{*c} \le t_m. \\ M, \text{if } e_i^{*c} > t_{M-1} \end{cases} \tag{2.11}$$

For each value of the transition matrix, we will estimate

$$Pr(e_i^c = m | e_i^p) = \begin{cases} G(t_1 - \Theta e_i^p), if\ m = 1 \\ G(t_m - \Theta e_i^p) - G(t_{m-1} - \Theta e_i^p), if\ 1 < m \le M-1 \\ 1 - G(t_{M-1} - \Theta e_i^p), if\ m = M \end{cases} \tag{2.12}$$

with the cumulative distribution function $G(t_m - \Theta e_i^p) = \frac{e^{t_m - \Theta e_i^p}}{1 + e^{t_m - \Theta e_i^p}}$. The low to high education probability is given by $Pr(e_i^c = M | e_i^p = 1)$.

2.3.2. Estimating the Relationship Between Relative Mobility in Income and Education

As pointed out by Narayan *et al*. (2018), mobility in education and mobility in income should be related. The authors argue that this relationship is likely to be positive because income persistence is verified due to the endowments that are inherited and to the investments parents make in children (e.g., education). Hence, our benchmark analysis of mobility is extended and we try to formalize their

relationship and expect that improvements in education mobility should be reflected in more income mobility, as follows.

THEOREM 2.1. *Assuming that the logged current income of parents, y_{it}^{p}, is orthogonal with respect to the age of children, A_{it}^{c}, and its squared, ${A_{it}^{c}}^{2}$, the responsiveness of intergenerational relative mobility in income (β_1) to marginal changes in intergenerational relative mobility in education (∂_1) is given by*

$$\frac{d\beta_1}{d\partial_1} = \varrho^{c}\varrho^{p}Var\left(Ed_{it}^{p}\right)\left[Var\left(y_{it}^{p}\right)\right]^{-1} \geq 0, \tag{2.13}$$

considering the model defined by

$$\begin{cases} y_{it}^{c} = \varrho^{c}Ed_{it}^{c} + \chi'^{c}W_{it}^{c} + u_{it}^{c} \\ y_{it}^{p} = \varrho^{p}Ed_{it}^{p} + \chi'^{p}W_{it}^{p} + u_{it}^{p} \\ y_{it}^{c} = \beta_0 + \beta_1 y_{it}^{p} + \gamma_1^{c}A_{it}^{c} + \gamma_2^{c}{A_{it}^{c}}^{2} + \alpha_{it} \\ Ed_{i}^{c} = \partial_0 + \partial_1 Ed_{i}^{p} + \pi_i \end{cases} \tag{2.14}$$

where the first two equations reflect, for children and for parents, respectively, the Mincer (1974) wage equations, which measure the change in logged current income (y_{it}) due to an additional year of current maximum education attained (Ed_{it}), reflected by ϱ, after controlling for other factors (W_{it}), namely the sector of activity/occupation, firm age, and size (proxied by the log of sales' volume), tenure and age (Pereira and Martins, 2004; Campos and Reis, 2018); the third regression corresponds to the standard equation used to estimate the intergenerational income elasticity; the last expression estimates the relationship between the maximum years of education attained of parents and children; and u_{it}, α_{it} and π_i are the error terms.

The proof of this theorem is in the Appendix B.

2.4. Data and Sample Construction

In this section, we present the databases that are used not only to construct the mobility measures but also to estimate the relationship between relative mobility in income and education, through the use of Mincer (1974) equations. Besides, we describe how our sample is constructed.

2.4.1. Data

To estimate our benchmark measures of mobility in income and education, we use two databases. Both are provided by INE (*Instituto Nacional de Estatística*, the Portuguese National Statistics Authority) and are the Portuguese components of two main European Union surveys. The first survey is the *Painel dos Agregados Domésticos Privados da União Europeia*, part of the European Community Household

Panel (ECHP), developed for 14 Member States. The second is the *Inquérito às Condições de Vida e Rendimento das Famílias*, which is a part of the European Union Statistics on Income and Living Conditions (EU-SILC) and was launched in 2003, replacing the first survey. Individuals are between 16-80 years old. Our sample of children is restricted to the latest survey wave, in which there is retrospective data on their parents. We use the 2019 wave of the EU-SILC as it contains a module aimed at providing information on intergenerational transmission of poverty. Individuals considered are between 30 and 50 years old. Here, personal information is used, in particular individuals were asked about their parents' characteristics when they were about 14 years old. The pseudo-parents' samples used in our analysis concern the 1995-1999 waves of the ECHP, since they are the ones closer to the periods in which the adults in our main sample are 14 years old. In the EU-SILC survey, an income reference period is defined as the period that income is related to. In most of the EU-member States it corresponds to the previous calendar year (fixed 12-month period). Hence, the outcomes' periods for specific variables considering the 2019 wave is 2018. The same applies to the 1995-1999 ECHP waves, where the reference period is 1994-1998.

To perform the extension to the benchmark analysis using Mincerian equations, we use another database, which is the *Quadros de Pessoal*. This is because in the ECHP and the EU-SILC databases the education levels of parents are provided in only three categories, which prevents us from transforming them into years of education with a considerable degree of disaggregation and compute an analogous to the IGE measure for education. This third dataset is provided by the Portuguese Ministry of Employment and annually links employers' and employees' information, including characteristics such as wages and education, among others. Individuals are also restricted to the 30-50-year-old range. We assume that the 1994 and 2018 waves of *Quadros de Pessoal* are representative of the same population covered using the other surveys, although household employees, self-employed individuals, and civil servants are excluded, and the public sector is not part of the dataset. Monetary information is provided in euros and the data regard the month of October.

Additionally, although research about intergenerational income mobility is mainly focused on fathers and sons, in this work we consider both genders for children. The reason, as stated above, is because Portugal has some very specifics characteristics regarding the female labour market and educational attainment for women.

2.4.2. Sample Construction

The sample construction is now presented. We describe how we deal with unobserved parental income, lifecycle effects in income measurement, differences between permanent and current income, and income measurement conditional on gender. We also show how we make information comparable across surveys and detail the definitions of income, education, occupation and managerial position related variables.

2.4.2.1. Income

2.4.2.1.1. Predicting Father's Income

We follow the common methodology of a variety of previous studies in which the datasets share the same characteristics as ours and father's income has to be predicted, namely, Björklund and Jäntti (1997), Leigh (2007), Lee and Solon (2009), and Nuñez and Miranda (2010). Our strategy can be formalized as follows. Consider that the log of parents' current income (in *t*) can be defined as the sum of permanent income y_i^p and time-varying characteristics, namely age (A) and its square (A^2) to control for life-cycle effects in income:

$$y_{it}^p = y_i^p + \gamma_1^p A_{it}^p + \gamma_2^p {A_{it}^p}^2 + \mu_{it}^p. \tag{2.15}$$

In the current wave of the survey (main sample) we cannot observe parental current income, y_{it}^p. We also cannot link parents and children across waves. Although this is the case, we can observe in an earlier wave of the survey the current income of individuals, which are assumed to be representative of the same population as the current one. We call it the auxiliary sample of pseudo-parents. Thus, let X_{ij}^p be a vector of dummies for each possible parental characteristic ($j \in J$) which can proxy for lifetime income (again, not observed), such that:

$$y_i^p = {\Phi'}_{ij}^p X_{ij}^p + \varphi_{ij}^p. \tag{2.16}$$

Equation (2.15) becomes:

$$y_{it}^p = {\Phi'}_j^p X_{ij}^p + \gamma_1^p A_{it}^p + \gamma_2^p {A_{it}^p}^2 + \mu_{it}^p + \varphi_{ij}^p. \tag{2.17}$$

We estimate equation (2.17) through an OLS estimator in t (i.e., our results are computed for a cross-section). The resulting coefficients are used to predict the current income of pseudo-parents of children in the main sample, $\hat{y}_{it}^p$:

$$\hat{y}_{it}^p = {\hat{\Phi}'}_{ij}^p X_{ij}^p + \hat{\gamma}_1^p A_{it}^p + \hat{\gamma}_2^p {A_{it}^p}^2. \tag{2.18}$$

We consider as potential proxies of parental permanent income their individual characteristics such as occupation, educational attainment and managerial position.

This approach has some issues attached to it that are worth mentioning. First, we use a sample of pseudo-parents which is not the same as using parents, taken from the population in our main sample. Second, the predicted income is not the same as the observed income. Third, results may be biased due

to the possible lack of validity of the instruments used. As pointed out by Solon (1992), there is the possibility of these instruments not being exogenous and, in turn, having a relationship with children's income that goes beyond the parental income channel. Grounded in Nicoletti and Ermisch (2008) and supported by the evidence presented by Björklund and Jäntti (1997), Cervini-Plá (2015) argues that these instruments may positively influence the children's income even after controlling for the parental income, promoting an upward bias in the estimate of the elasticity. Thus, most authors that use this method assume that the estimates are upper bounds of the true coefficient. We test how sensitive our results are to the use of different combinations of characteristics that proxy for parental permanent income. Fourth, as parental income is predicted using a small number of different instruments that proxy for their permanent income, we have a limited small set of distinct values that these can assume and a lack of variability in parental income[18]. All together these issues may influence the results and conclusions.

Additionally, the *Quadros de Pessoal* database does not contain retrospective information on parents so we cannot predict father's income and estimate its variance. We assume that the 1994 wave of this dataset will have the fathers' income of individuals that are in the 2018 sample. The same occurs for parental education.

2.4.2.1.2. Life-cycle and Attenuation Bias

To account for the life-cycle measurement error we restrict our sample to individuals aged 30-50 years old[19]. Current income is used for both generations. We predict parental income at 40 years old, the age in the middle of the range at which permanent income may be proxied[20,21]. To address the standard attenuation bias, existing evidence shows that a large time range would be needed to make it disappear. According to Mazumder (2005), a father's income averaged for 5 years will still produce attenuated beta (IGE) estimates, which are 30% biased for the USA, and even using a 25-year range period the bias would remain. As Cervíni-Plá (2014) points out for Spain (Spain's data have the same characteristics as ours), when using instruments in the TSTSLS approach to proxy for parental income and then predict parental current income, one is already computing its average. By using a single year for parental income in our benchmark sample, we assume that we are obtaining the most attenuated estimate of relative persistence in income, which means that relative mobility in income may be lower than the one we obtain[22]. Additionally, considering more than a single year implies guaranteeing that individuals are in

[18] This has implications for the rank-rank slope because we cannot rank predicted parental income in percentiles as it is done for children. Nevertheless, parental income is still ranked but in different bins.

[19] Different authors used similar age ranges: e.g., 30-50 in Cervini-Plá (2014), 25-54 in Mendolia and Siminski (2019), and 38-45 in Corak (2019).

[20] We follow authors such as Leigh (2007) and Mendolia and Siminski (2019).

[21] Results for the first stage are presented in Table B3 in the Appendix B.

[22] For the standard attenuation bias, when we average the annual income of fathers from 1 to T and regress y_i^c on $\bar{y}_i^p = \frac{1}{T}\sum_{t=1}^{T} y_{it}^p$, we obtain that $plim\ \hat{\beta}_1 = \beta_1 \frac{Var(y_i^p)}{Var(y_i^p)+\frac{1}{T}Var(\tau_{it}^p)} < \beta_1$. If $t \to +\infty$, beta becomes less attenuated, which reflects more persistence (Björklund and Jäntti, 1997).

the cross-sectional samples for all periods, which reduces the number of observations by a large amount. We perform a sensitivity exercise to assess how sensitive our estimates are when using an average for parental income (i.e., using more than one period to compute it).

2.4.2.1.3. Measurement Issues

In our sample of children, individuals can be either single or married. For the latter some concerns may rise. Regarding married women, Ermisch *et al.* (2006) consider that if the female labour force participation of married women is lower than the male labour force participation of married men, this may not only reflect that in a couple men are more likely to work, but also that women's decision to work is not random. Cervini-Plá (2014) points out that the decision of women to work may also be related to the fact that they belong to households with specific characteristics, namely to those in which a single person working is not enough to support the couple's expenditures. When looking at married women, their individual income may not be a good measure of their true economic status and could yield biased mobility estimates.

Chadwick and Solon (2002) show that in the case of daughters, we should use the couple's income to better proxy for their economic status. Although this may justify the use of couple's income for women, it should not rule out the use of the couples' income as well for men. This is because in our sample women earn on average 45% of the couple's income. Hence, the difference between 55% and 45% of the couple's income is not substantial. This makes us consider the couple's income as well for men, when married.

We restrict the parents' sample to fathers only: as we predict parental individual income and the best option for women is to use family income/couple's income (while for men, concerns are not that clear in the literature), we do not have an intersection between both conditions. We will therefore have estimates for the pairs father-children, father-son, and father-daughter. A father is defined as the individual considered by the interviewed person as his or her father when aged 14, having (or not) a biological relationship, even if the biological father was known and alive.

Additionally, since we are studying intergenerational income mobility, we decide to include only individuals with positive income during the income reference period. For singles we use individual income. For married individuals we use the combined income of the couple, i.e., we add the total income of the couple and divide by two, obtaining an average, following Chadwick and Solon (2002) and Raaum *et al.* (2008). Married individuals who do not work but benefit from the income of his/her spouse are also not considered, as what they earn is not a direct result from being active in the labour market. In a later sensitivity exercise we include the partners with no individual income, but with positive average couple's income and test if results change. We also perform a sensitivity analysis to evaluate the possible differences arising from using individual income instead of average couple's income when individuals are married.

When using the ECHP, we measure labour-related income as the wage and salary income for employees and self-employment income for employers. The corresponding variables available in the EU-SILC are the net employee cash or near cash income and the net cash profits or losses from self-employment. In the second survey, the first variable is defined as the gross cash or near cash income, deducted from tax at source and/or social insurance contributions. In turn, gross cash or near cash income consists of the cash monetary component of employees' compensation paid by an employer, including the value of income taxes and social contributions that are paid either by the employee or by the employer to tax authorities and/or social insurance schemes (on behalf of the employee). The second variable can be defined as the net of tax at source and/or social insurance contributions net operating profit or loss for owners/partners that work in an unincorporated company, with interest on business loans deducted, plus royalties (writing, inventions, among others). To make income comparable across surveys we use the Consumer Price Index (CPI) with a base year in 2010 to obtain income in real values.

We also define income levels for both children and parents. We ground our definition for each income category on the OECD definition for low and high pay workers[23]. We consider the low-income level to be the one in which individuals earn less than two-thirds of the median national income, while the high-income level comprehends individuals earning one and a half the median income. Individuals classified as middle-level earners are those between, and are split into two categories, middle-low and middle-high, according to the intermediate value of the category's possible values' range. We again apply the CPI base year 2010. For parents, the log income' bounds separating classifications are 8.81, 9.29, and 9.62. For children we have 8.96, 9.45, and 9.77.

When the *Quadros de Pessoal Database* is used, we can only obtain individual income because the marital status or the identification of partners is not available. The income variables that follow the definition of gross employee income of the ECHP and the EU-SILC are the base remuneration, regular payments made to employees, and supplementary payments. In the *Quadros de Pessoal* database it is provided in gross terms. To obtain net values we define two scenarios grounded on the factors determining the tax level, considering individuals that are not married: the category that the wages belong to and the number of dependent individuals. In the first scenario, for each range of gross wages, we deduct when applicable the maximum taxation, and in the second case, we apply the minimum taxation possible. Hence, $d\beta_1/d\partial_1$ will have an upper and a lower bound. In each scenario, although social security percentage contributions may differ across individuals, the general case is a deduction of 11%. The values obtained are then transformed into real amounts using the Consumer Price Index (base 2010).

[23] https://data.oecd.org/earnwage/wage-levels.htm

2.4.2.2. Education

Education is classified using the International Standard Classification of Education (ISCED) of the United Nations Scientific and Cultural Organization (UNESCO). There exist two categorizations. The first one, ISCED 1997, considers 7 levels of education. Data in the 1995 wave of the ECHP cover three valid groups, which have correspondence with the ISCED 1997 classification. The second categorization, ISCED 2011, was used in the 2019 wave, covering 9 levels. When asked about their parents' education, children's responses are divided into low, medium, and high educational levels, which have correspondence with ISCED 2011 classification. This means that to estimate intergenerational mobility in income – in which we predict parental income grounded on educational attainment – as well as in intergenerational education mobility, we must match children's own education levels in the pseudo-fathers' education categories, which is presented in Table 2.1 below.

Table 2.1 – Correspondence between ISCED Classifications Across Surveys

ECHP 1995 (ISCED 1997)	EU-SILC 2019 (ISCED 2011)	Retrospective question about parents
Less than second stage of secondary education	Primary	Low level
	Lower secondary	
Second stage of secondary education	Upper secondary	Medium level
Recognized third level education	Short cycle tertiary	High level
	Bachelor or equivalent	
	Master or equivalent	
	Doctorate or equivalent	

Notes: Adapted from Eurostat online tables (correspondence between ISCED 2011 and 1997 levels). Source: http://www.uis.unesco.org/Education/Pages/internationalstandard-classification-of-education.aspx.

Information about education in the pseudo-parents' sample is only used to proxy for their permanent income and then predict their current income (which is not available) in a first stage, which is the estimation of intergenerational mobility in income. For the estimation of intergenerational mobility in education we only need to use the children's samples where retrospective information about education is directly available. As we aim not only to analyse mobility in education, but also to identify patterns regarding its joint behaviour with mobility in income, we should consider the same individuals in both analyses, which implies that the age range we first chose is the same. We also have to ensure that individuals are not enrolled in school. Therefore, we include in the analysis only individuals between 30-50 years old, which have finished school and are not enrolled in any type of education at the time of the survey, following Urbina (2018). In 2018, 5% of the Portuguese individuals aged 30-34 where still enrolled in school, 4% for the age range of 34-39, and 2% for 40-64 years old[24].

When computing the relationship between income and education mobility in relative terms through Mincerian equations, we are able to classify individuals according to ISCED-11 and then convert the

[24] https://stats.oecd.org/Index.aspx?DataSetCode=EAG_ENRL_RATE_AGE

education levels, which have the same degree of disaggregation across generations, into years of education[25] (not possible in the benchmark analysis).

2.4.2.3. Occupation

The International Standard Classification of Occupations (ISCO) from the International Labour Organization (ILO) is considered in our work. For the 1995 wave of the ECHP the ISCO-88 classification is used, while for the 2019 wave of the EU-SILC the ISCO-08 classification is considered. The correspondence is in Table 2.2.

Table 2.2 – Correspondence Between ISCO Classifications Across Surveys

ECHP 1995 (ISCO-88)	EU-SILC 2019 (ISCO-08)
Legislators, senior officials and managers	Managers
Professionals	Professionals
Technicians and associate professionals	Technicians and associate professionals
Clerks	Clerical support workers
Service workers and shop and market sales workers	Services and sales workers
Skilled agricultural and fishery workers	Skilled agricultural, forestry and fishery workers
Craft and related trades workers	Craft and related trades workers
Plant and machine operators and assemblers	Plant and machine operators and assemblers
Elementary occupations	Elementary occupations

Source: International Labour Organization (https://www.ilo.org/public/english/bureau/stat/isco/).

2.4.2.4. Managerial Position

Another characteristic we use to proxy for father's permanent income is his managerial position. The parent can be either in a supervisory or non-supervisory position. We create a dummy variable that is equal to 1 in the first case, if the individual has formal responsibility for an employees' group, with direct supervision of the work, and 0 otherwise. Expectedly, for the same occupation category and education level, an individual in a superior managerial position should have higher income than one in a lower managerial stage.

Summary statistics are presented in Table B2 in the Appendix B.

2.5. Empirical Results and Discussion

In this section we present our benchmark results for the measures of intergenerational mobility in income and education for the Portuguese economy[26].

[25] We use the minimum cumulative time required to complete a specific education level presented in the work of Narayan *et al*. (2018). The correspondence for each education level is as follows: primary – 6 years; lower secondary – 9 years; upper secondary – 12 years; short cycle tertiary – 15 years; bachelor or equivalent – 16 years; master or equivalent – 18 years; doctorate or equivalent – 21 years.

[26] The surveys we use provide individual weights that are computed accounting for the sample design and individuals' characteristics. They reflect the structure of the population: the greater the weight the stronger the

2.5.1. Intergenerational Mobility in Income

Table 2.3 presents the benchmark results for the intergenerational mobility in income for all children regardless of gender and also for male and female children separately.

Table 2.3 – Benchmark Results for Intergenerational Mobility in Income

	Elasticity	Corr.	Rank-rank	Prob.	Share
All individuals	0.26***	0.20***	0.45***	7.15***	53.11
n = 2,549 | N = 980,083	(0.04)		(0.01)	(0.01)	
Males	0.3***	0.24***	0.48***	8.21***	52.90
n = 1,027 | N = 431,849	(0.05)		(0.02)	(0.01)	
Females	0.22***	0.17***	0.42***	6.42***	53.28
n = 1,522 | N = 548,234	(0.06)		(0.02)	(0.01)	

Notes: Standard errors are presented in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals earning more than their fathers are expressed in %. The share of individuals earning more than their fathers does not have an associated significance level. n stands for the number of observations in the sample and N for the total population represented by those observations using survey weights.

Our results make gender differences evident. Women show more intergenerational income mobility than men, with the exception of the bottom to top income level probability. A distorted labour market may be contributing for these results. In comparison with poorer parents, richer and more educated parents are able to invest in their children's education and offspring's characteristics unrelated to education. Their social status provides children the access to better opportunities, all of which are valued in the labour market, so incomes tend to persist. According to Narayan *et al*. (2018) if a labour market values individuals' attributes over which they have no control rather than their abilities, income persistence may become stronger. Accordingly, if the labour market discriminates individuals grounded on gender, i.e., if women's income is penalized for the same education level of men, it may be the case that males' income positions are more tied to the ones of their fathers. The Portuguese economy is in line with this argument, since Reis and Campos (2017) show that from 1986 until 2013 there is a constant gender wage gap favouring men.

Our evidence is also verified in the literature regarding other countries. Borisov and Pissarides (2019) show that mobility is higher in correlation ranks for females in Russia. For this measure as well as for the intergenerational income elasticity, Helsø (2020) finds daughters to be more mobile than sons for Denmark and USA, while ambiguous findings are reported by Kyzyma and Groh-Samberg (2020) for Germany. Acciari *et al*. (2022) show that mobility is higher for women when considering the rank-rank slope for Italy. Considering the work of Comi (2003), for an older generation and with differences

representativeness an individual has on the population, which cannot be ignored. We therefore use population weights in our analysis.

in variables' definitions and sample construction, the same finding is presented for Portugal regarding the intergenerational income elasticity, according to which girls show more mobility.

When analysing intergenerational mobility measures, a main goal is to stress how high or low mobility is. This is done through comparisons between countries. We must be careful in the comparisons because estimates are sensitive to measures of income, estimation methods, and sample selection, among others. This means that we try to choose works that make choices to ours in terms of sample and methods.[27] Most of these studies address mainly the case for intergenerational income elasticity for a single gender (usually men).

By country and for sons, we have elasticities being around: 0.1-0.3 (Blanden *et al*., 2004), 0.20-0.25 (Nicoletti and Ermisch, 2008), and 0.56-0.59 (Dearden *et al*., 1997) for the UK[28]; 0.19-0.22 for Canada (Fortin and Lefebvre, 1998); 0.28 for Sweden (Björklund and Jäntti, 1997); 0.2-0.3 (Leigh, 2007), 0.35 (Mendolia and Siminski, 2016) and 0.59-0.74 (Nuñez and Miranda, 2010) for Australia; 0.4 for France (Lefranc and Trannoy, 2005); 0.42 for Spain in Cervini-Plá (2015); 0.45-0.53 (Solon, 1992), 0.34-0.49 (Lee and Solon, 2009), and 0.52 (Björklund and Jäntti, 1997) regarding the USA; 0.5 for Italy (Piraino, 2007; Mocetti, 2007); 0.58 (Ferreira and Veloso, 2006) and 0.69 (Dunn, 2007) for Brazil. Our estimated value for the elasticity of males, 0.3, is similar to some of the estimates for the UK, Sweden, and Australia, but higher than the estimates found for Canada, and lower than the ones for France, Spain, the USA, Italy, and Brazil.

For daughters we have elasticities ranging about: 0.05-0.46 in the USA (Lee and Solon, 2009); 0.1-0.3 (Blanden *et al*., 2004) and 0.63-0.70 (Dearden *et al*., 1997) in the UK; 0.3 in France (Lefranc and Trannoy, 2005). Our estimate of 0.22 fits in the interval of some of the estimates made for the USA and the UK, but lower than the estimate made for France.

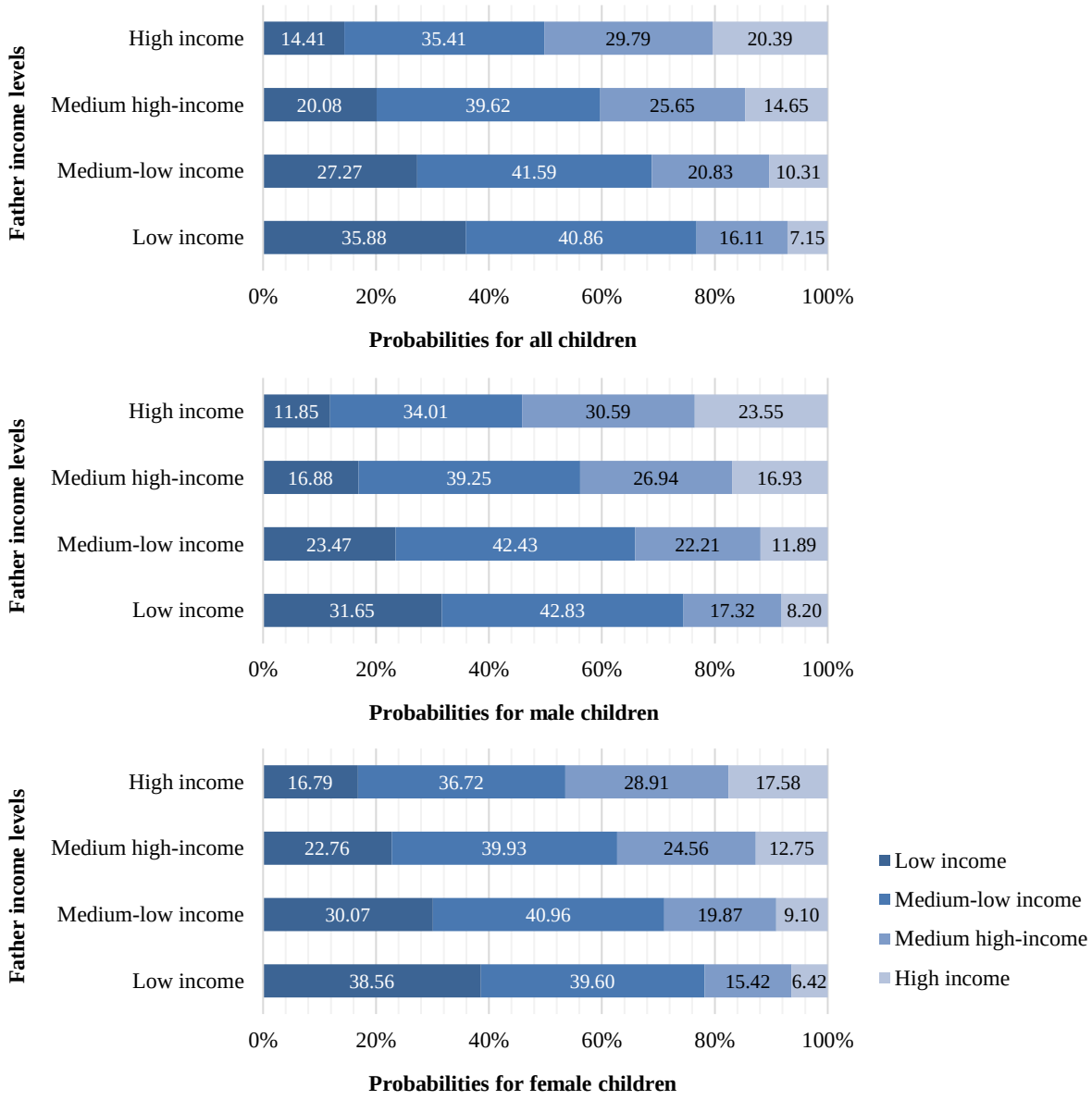
Mendolia and Siminsky (2019) also compute the intergenerational income correlation, which is around 0.233 for men in Australia, similar to our findings. We can only compare our estimates for men and women with those in the literature. None of the authors instrumenting and predicting parental income compute the other mobility measures. For sons, Portugal may stand amongst the most relative mobile countries in income, being similar to the UK, Australia and Sweden. Regarding daughters, it fits in all the ranges for the countries described.

Figure 2.1 presents the transition probabilities between father and children (sons and daughters) income levels in the respective generations, which complements the previous measures.

[27] Slight differences between ours and the following studies, and also across studies, may lead to wrong conclusions (see Solon, 2002). This is also true for education mobility estimates.

[28] Large differences for the UK may be due to differences in the cohorts or the surveys used by the authors.

Figure 2.1 – Intergenerational Transition Probabilities in Income Using an Ordered Logit



Notes: Probabilities obtained using an ordered logit are expressed in % and are all statistically significant at 1%. Parental individual income (in logs) is predicted at the age of 40 years old, with results for the first stage presented in Table B3 in the Appendix B and using father's education, occupation, and managerial position as instruments for permanent income. Children's income (in logs) correspond to the average of the couple's income when married and to individual income when not married. Results can be found in Table B4.

There is a strong degree of intergenerational mobility when the father is classified as low-income earner: the majority of individuals are likely to arrive at higher income levels when adults. The probability of keeping a high income level is lower than the one of reaching a lower income level (downward mobility is high for children of high income fathers). Besides, the upward probabilities decrease the higher the fathers' income levels. The chances of reaching a high-income level are lower for all fathers' income levels. The likelihood of departing from a low-income level and reaching the highest is lower than the opposite movement. The chances of ending up in the medium-low income level

are the highest. These are higher for females with fathers in the medium high and high-income levels (39.93% and 36.72% for women when compared to the 39.25% and 34.01% for men, respectively), and higher for males with fathers in the low and medium low-income levels (39.60% and 40.96% for women when compared to the 42.83% and 42.43% for men with fathers, respectively).

2.5.2. Intergenerational Mobility in Education

We present the benchmark results for intergenerational mobility in education in Table 2.4.

Table 2.4 – Results for Intergenerational Mobility in Education

	Correlation	Prob.	Share
All individuals n = 2,549 | N = 980,083	0.26***	44.35*** (0.01)	84.47
Males n = 1,027 | N = 431,849	0.24***	36.93*** (0.02)	82.54
Females n = 1,522 | N = 548,234	0.29***	49.98*** (0.02)	85.99

Notes: Standard errors are presented in parentheses. *** stands for statistically significant at 1% levels. Probabilities obtained using an ordered logit and the share of individuals with more education than their fathers are expressed in %. The share of individuals with more education than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Results show that men are more relatively mobile than women (0.24 compared to 0.29 for the intergenerational correlation) while more persistent in absolute terms (36.93% compared to 49.98% for the low to high education level probability and 82.54% compared to 85.99% for the share of individuals with more education than their fathers). Gender differences in relative mobility may be related with differences in school drop-out rates. Narayan *et al*. (2018) show that lower out-of-school rates tend to be associated with higher relative education mobility. This is because opportunities are equalized across individuals from different educational backgrounds. In Portugal, women appear to have higher dropout rates than men in more than 90% of times, considering the primary education level between 1968 and 2018. For the lower secondary level, this share is close to 89%[29]. Besides, regarding the primary level of education, the reduction in school dropouts in the same period was more pronounced for men than for women (99% for men and 90% for women), while these were similar between genders for the lower secondary education level. All together may have made women present more relative education persistence than men.

[29] Data on dropout rates for the primary education level for males and females are available in https://data.worldbank.org/indicator/SE.PRM.UNER.MA.ZS and https://data.worldbank.org/indicator/SE.PRM.UNER.FE.ZS, respectively. For the lower secondary level, these are presented in https://data.worldbank.org/indicator/SE.SEC.UNER.LO.MA.ZS and https://data.worldbank.org/indicator/SE.SEC.UNER.LO.FE.ZS, respectively.

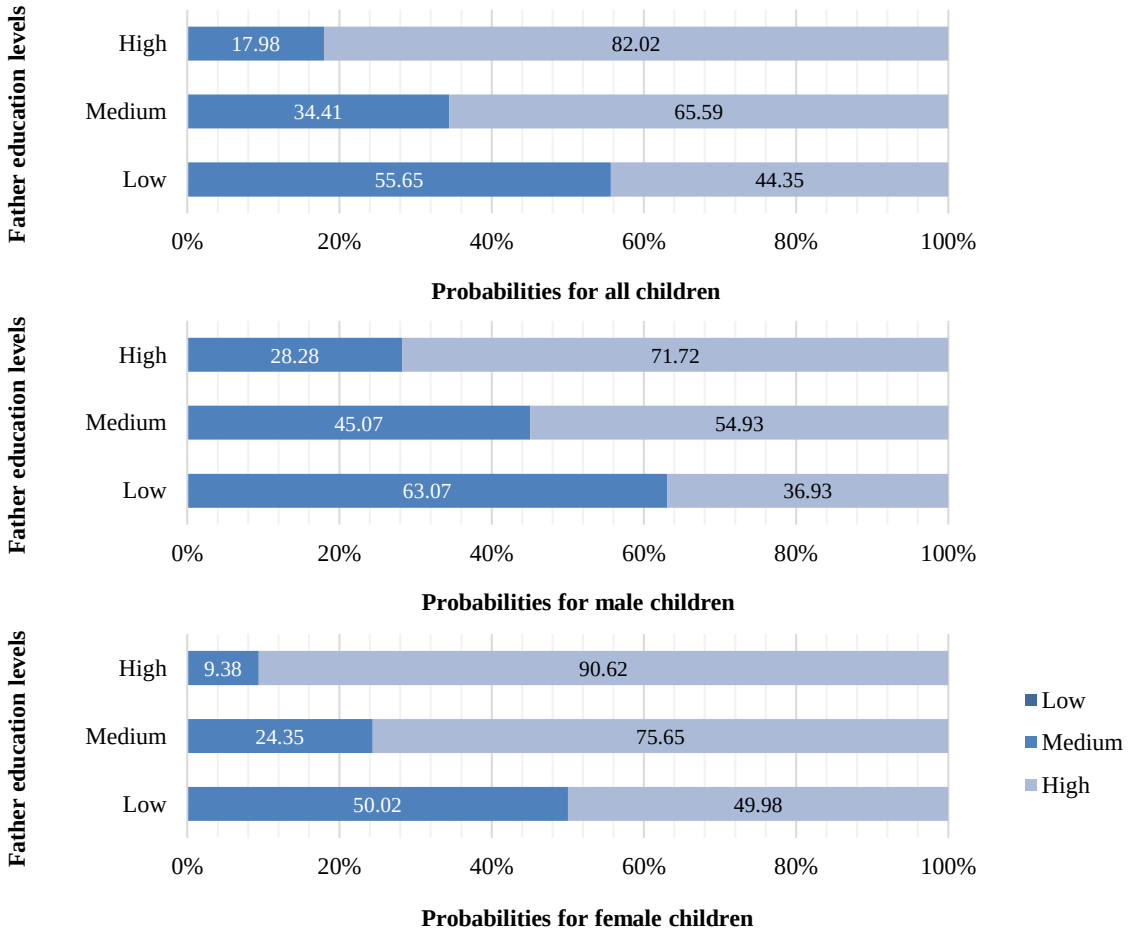
Some authors compute an analogous measure to our probability measure. Lam and Liu (2019) find that for primary and lower secondary educated fathers (both in our low level of education), the chances children have of reaching the high education level are in the 26.63-33.07% range, for both-generation Hong Kong born individuals, 32.78-40.61% for second-generation Mainland immigrants, and 16.11-20.35% for both-generation Mainland immigrants. Schneebaum *et al.* (2016) show that for Austria, this likelihood is around 8% for males and 7% for daughters. For Portugal, Bago d'Uva and Fernandes (2017) find that this is around 20% considering male children, which is about 17 percentage points below our 36.93% estimate. Although cohorts used are similar to ours (1970-1985), differences should be noted in the methodology. They use a multinomial logit and their calculations involve the 2005 and 2011 waves of the EU-SILC. All these are below our estimates and the differences between genders are the opposite to what we obtain. The share of individuals with more education than their fathers is higher than 80%. This value is larger than the one found in Lam and Liu (2019) for Hong Kong-born children with Hong Kong-born fathers (78.06%), while lower than the one for Hong Kong-born children of Mainland immigrant fathers (89.47%). Both generation Mainland immigrants fall in the middle (86.54%). Due to the lack of comparability in the literature that, for our measures, is scarce, we cannot infer if Portugal has high absolute mobility in education (or not) in the World.

Education correlations are the most studied measure in the literature and mainly use years of education instead of education levels. Considering that there may be a strong link between years of education and education level attained, we abstract from this last issue. Urbina (2018) is the only investigator studying the pair father-children and finds a correlation that is between 0.45 and 0.51 regarding Mexico. As before, reported studies often confront the analysis for each gender separately. Schneebaum *et al.* (2014) consider 20 European countries. In general, mobility is lower for sons (0.33) when compared to daughters (0.26). They include Portugal in their analysis, finding values similar to ours: the intergenerational correlation for Portugal for the pair father-son is equal to 0.24 (the same as we obtain), while the pair father-daughter is equal to 0.26 (lower than our estimate). This country presents the highest mobility when considering men. Regarding daughters, Portugal is surpassed in terms of mobility by France (0.24), all the Nordic countries (average correlation of 0.20), the Anglo-Saxon countries (average correlation of 0.23), Greece (0.22), Czech Republic (0.20), and Poland (0.21). The highest persistence value is found for Italy (0.40). Latif (2018) shows for Canada that boys are on average less mobile than girls with the education correlation being equal to 0.33 for boys and 0.32 for girls. Schneebaum *et al*. (2016) found that persistence appears to be greater for girls, 0.43, than for boys, 0.41, for Austria. Azam and Bhatt (2015) find that the correlation between father and son's education is around 0.64 for India. To sum up, Portugal is the most relatively mobile country in education for sons when considering the intergenerational correlation in education whereas for daughters it is in the middle of known World's estimates.

The fact that Portugal is the one presenting a larger relative change in the government expenditures as a share of GDP may be leading our evidence (an increase around 186% between 1968 and 2018[30]). According to Narayan *et al*. (2018), higher public spending in education is associated with larger relative mobility in education in richer countries, by compensating the inequality in private investments in education between poor and rich parents. Besides, the role of school dropouts can be again considered to explain differences between countries. Portugal is the country with the highest decrease in the school drop-out rate for men, considering primary education, while for women it lays in the middle of the group of countries' estimates. However, women do not maintain the same position when compared to other countries in terms of primary school dropout rates' decrease as the one they have regarding relative education mobility. This reinforces the argument of Clements (1999) that early education is one of the main drivers of educational achievement in the Portuguese economy.

Figure 2.2 presents the transition probabilities for education levels considering both generations.

Figure 2.2 – Intergenerational Transition Probabilities in Education Using an Ordered Logit



Notes: Probabilities obtained using an ordered logit are expressed in % and are all statistically significant at 1%. Results can be found in Table B5.

[30] https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS

An interesting result emerges when we analyse the transition probabilities for intergenerational mobility in education. The probability of staying in the same low education level as the father is equal to 0%, i.e., individuals present full absolute mobility when raised in a low educated environment. This result appears to be stronger than the one found by Bago d'Uva and Fernandes (2017): noting the same differences in methodology mentioned before, sons with low educated fathers appear to have almost 50% chance of reaching a higher education level. When the father is classified as medium educated, children's chance of surpassing that level is higher than the one they have of obtaining the same level. The probability of remaining in the same education level of the father is higher for men regarding the medium education level and for women regarding the high education level (45.07% compared to 24.35% for the first case and 90.62% compared to 71.72% in the second case). Moreover, the chances of completing the highest education level is always higher for females when compared to males for all the father's education levels. Finally, the likelihood that an individual has of reaching or remaining in the high-education level is increasing on the father education level, which reflects a high persistence at the top of the education classification: this finding is similar to that presented by Bago d'Uva and Fernandes (2017), regarding sons born from 1950 on.

2.5.3. Sensitivity Analysis

In this section, we check how sensitive our benchmark estimates of intergenerational mobility are to changes in variables definitions and sample construction.

2.5.3.1. Income Definitions

We start by considering different income definitions. We use individual income as opposed to the benchmark estimation in which the average couples' income is used. We are able to do this exercise just for children, not only because the characteristics used to proxy for father's permanent income pertain to individual income, but also because the father's marital status is not known. Results are presented in Table 2.5.

Table 2.5 – Sensitivity of Intergenerational Mobility in Income to Alternative Income Definitions for the Benchmark Sample

	Income definitions for children	Elasticity	Correlation	Rank-rank slope	Prob.	Share
All individuals n = 2,549 N = 980,083	**Average total family income**	**0.26*** (0.04)**	**0.20*** **	**0.45*** (0.01)**	**7.15*** (0.01)**	**53.11**
	Individual income only	0.27*** (0.05)	0.18***	0.39*** (0.01)	9.36*** (0.01)	52.30
Males n = 1,027 N = 431,849	**Average total family income**	**0.30*** (0.05)**	**0.24*** **	**0.48*** (0.02)**	**8.21*** (0.01)**	**52.90**

(continues in the next page)

Table 2.5 – Sensitivity of Intergenerational Mobility in Income to Alternative Income Definitions for the Benchmark Sample *(continued)*

	Income definitions for children	Elasticity	Correlation	Rank-rank slope	Prob.	Share
Males n = 1,027 N = 431,849	Individual income only	0.27*** (0.07)	0.18***	0.31*** (0.02)	16.1*** (0.02)	59.30
Females n = 1,522 N = 548,234	**Average total family income**	**0.22*** (0.06)**	**0.17*** **	**0.42*** (0.02)**	**6.42*** (0.01)**	**53.28**
	Individual income only	0.23*** (0.06)	0.16***	0.40*** (0.02)	5.14*** (0.01)	46.49

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals earning more than their fathers are expressed in %. Correlations have the associated significance level of the elasticities used to compute them. The share of individuals earning more than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

The intergenerational income elasticity and the share of individuals earning more than their fathers are the ones for which there is only a slight increase in persistence compared to the benchmark (and therefore they may be considered as reasonably robust to the income definition). On other hand, when analysing the intergenerational income correlation and the rank-rank slope, one may conclude that there is a change of about 10% and 13%, respectively, meaning that persistence is higher in the first scenario. The bottom to top income level probability shows a 31% increase between the two cases (7.15% in the benchmark compared to 9.36% when using an individual measure of income), which is the biggest change. This is in line with relative persistence increases when using household related measures in the work of Murray *et al*. (2018). Although the most obvious reason for the correlation change is related with the increase in the variability in children's income, when considering individual income, it is likely that assortative mating had play its role for both measures: this is the process according to which individuals select a partner with similar backgrounds. Torche (2015) argues that if the characteristics of individuals with which one shares a life are approximately the same, it is therefore expected that persistence will be higher in those cases, when compared to the scenario for which this type of mating does not occur. A simple exercise allows us to have a clue on the likelihood this has of occurring in our estimation sample. About 32.19% of individuals who are married and have fathers in the medium-high and high-income levels, have selected individuals with fathers in those same levels. The scenario is more evident when considering married individuals with parents in the low and medium-low income levels, with that share being approximately 43.74%.

When individual income is considered, men have the intergenerational income correlation and the rank-rank slope decreasing more than women. Persistence increases when the couples' income is considered, with a higher percentage change for men. This may reflect the fact that men are more likely to be married to individuals with similar backgrounds than women. From the medium-low parental income level on, the shares of men in this situation are approximately 51.37, 31.29, and 34.12% against

48.95, 29.14, and 33.43% for women. The exception is the low level, where women surpass men by 2 percentage points, with a share equal to 4%. Interestingly, absolute persistence for men increases when the couples' income is used in comparison with individual income (the opposite occurs for women, who benefit in terms of mobility when average couples' income is considered).

All in all, the exercise of using individual income instead of average couple's income provides different results from the benchmark analysis. This reinforces our decision to consider average total income instead of individual income only, since the marital status of individuals plays a role. In other words, the point made by Chadwick and Solon (2002) is clear: there is some degree of intergenerational persistence in household structure which cannot be ignored.

2.5.3.2. Alternative Specifications for Parental Income

We now test how sensitive our results are to different combinations of instruments used to predict parental income. Table 2.6 presents our estimates (excluding the cases where only one instrument is used as predictor).

Table 2.6 – Sensitivity of Intergenerational Mobility in Income to Alternative Instruments for Father Income for the Benchmark Sample

	Instruments	Elasticity	Correlation	Rank-rank slope	Prob.	Share
All individuals n = 2,549 | N = 980,083	Education, managerial position	0.26*** (0.04)	0.19***	3.75*** (0.06)	6.99*** (0.01)	55.66
	Occupation, managerial position	0.25*** (0.04)	0.28***	1.12*** (0.02)	7.13*** (0.01)	50.85
	Education, occupation	0.27*** (0.04)	0.20***	0.99*** (0.02)	6.95*** (0.01)	56.04
	Education, occupation, managerial position	**0.26*** (0.04)**	**0.20*** **	**0.45*** (0.01)**	**7.15*** (0.01)**	**53.11**
Males n = 1,027 | N = 431,849	Education, managerial position	0.31*** (0.06)	0.23***	3.98*** (0.12)	7.51*** (0.01)	59.07
	Occupation, managerial position	0.28*** (0.05)	0.22***	1.16*** (0.05)	8.17*** (0.01)	51.08
	Education, occupation	0.31*** (0.06)	0.23***	1*** (0.05)	8.05*** (0.13)	57.19
	Education, occupation, managerial position	**0.3*** (0.05)**	**0.24*** **	**0.48*** (0.02)**	**8.21*** (0.01)**	**52.9**
Females n = 1,522 | N = 548,234	Education, managerial position	0.22*** (0.07)	0.15***	3.45*** (0.07)	6.53*** (0.01)	52.98
	Occupation, managerial position	0.21*** (0.06)	0.17***	1.06*** (0.03)	6.4*** (0.01)	50.67
	Education, occupation	0.24*** (0.06)	0.17***	0.96*** (0.04)	6.15*** (0.01)	55.12

(continues in the next page)

Table 2.6 – Sensitivity of Intergenerational Mobility in Income to Alternative Instruments for Father Income for the Benchmark Sample *(continued)*

	Instruments	Elasticity	Correlation	Rank-rank slope	Prob.	Share
Females n = 1,522 | N = 548,234	**Education, occupation, managerial position**	**0.22*** (0.06)**	**0.17***	**0.42*** (0.02)**	**6.42*** (0.01)**	**53.28**

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals earning more than their fathers are expressed in %. Correlations have the associated significance level of the elasticities used to compute them. The share of individuals earning more than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Results are robust to this sensitivity exercise with the exception of the rank-rank slope, which is unstable when different combinations of instruments are considered. This makes us unable to guarantee the direction of the bias which is very likely to be present and may be a consequence of not observing parental income. However, we also recognize that the case with more instruments makes the rank-rank slope more efficient, with lower standard errors.

2.5.3.3. Inclusion of Individuals with No Individual Income

There may exist mobility mismeasurement in our benchmark analysis by including only individuals that work. Hence, we tested the sensitivity of the results by including individuals with no individual income derived from work. Results are presented in Table 2.7.

Table 2.7 – Sensitivity of Intergenerational Mobility in Income to the Inclusion of Individuals with no Individual Income

		Elasticity	Correlation	Rank-rank slope	Prob.	Share
All individuals	**Benchmark sample** n = 2,549 | N = 980,083	**0.26*** (0.04)**	**0.20***	**0.45*** (0.01)**	**7.15*** (0.01)**	**53.11**
	Considering individuals with no individual income n = 2,665 | N = 1,024,434	0.27*** (0.04)	0.20***	0.45*** (0.01)	6.93*** (0.01)	51.84
Males	**Benchmark sample** n = 1,027 | N = 431,849	**0.3*** (0.05)**	**0.24***	**0.48*** (0.02)**	**8.21*** (0.01)**	**52.9**
	Considering individuals with no individual income n = 1,059 | N = 441,624	0.31*** (0.06)	0.24***	0.48*** (0.02)	8.11*** (0.01)	52.1
Females	**Benchmark sample** n = 1,522 | N = 548,234	**0.22*** (0.06)**	**0.17***	**0.42*** (0.02)**	**6.42*** (0.01)**	**53.28**

(continues in the next page)

Table 2.7 – Sensitivity of Intergenerational Mobility in Income to the Inclusion of Individuals with no Individual Income *(continued)*

		Elasticity	Correlation	Rank-rank slope	Prob.	Share
Females	Considering individuals with no individual income n = 1,606 | N = 582,810	0.23*** (0.06)	0.17***	0.41*** (0.02)	6.14*** (0.01)	51.64

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals earning more than their fathers are expressed in %. Correlations have the associated significance level of the elasticities used to compute them. The share of individuals earning more than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Results are almost unchanged for all measures, both for all individuals and for each gender. Percentage changes in the estimates are no higher than 5%.

2.5.3.4. Co-residents Bias

Following Azam and Bhatt (2015), the co-resident bias may exist in our context. The idea is that if parents are part of the same household as children, they can still influence their offspring's decisions about education. The authors point out that the use of samples with co-residents (children-parents) may lead to problems related to sample selection, as co-resident individuals may not represent the adult population. We consider this to be true also for work-related decisions and therefore income, although published research is mainly related to education. Our benchmark sample includes not only co-resident fathers and children, but also individuals who do not live with their fathers. Now, we compare the original estimates to a sample with no co-residents and see whether the results change significantly. Results are presented in Table 2.8 for income mobility and Table 2.9 for educational mobility.

Table 2.8 – Sensitivity of Intergenerational Mobility in Income to the Exclusion of Co-residents

		Elasticity	Correlation	Rank-rank slope	Prob.	Share
All individuals	**Benchmark sample** n = 2,549 | N = 980,083	**0.26***** **(0.04)**	**0.20*****	**0.45***** **(0.01)**	**7.15***** **(0.01)**	**53.11**
	Without co-resident fathers n = 2,279 | N = 901,644	0.27*** (0.04)	0.21***	0.47*** (0.02)	7.42*** (0.01)	55.08
Males	**Benchmark sample** n = 1,027 | N = 431,849	**0.3***** **(0.05)**	**0.24*****	**0.48***** **(0.02)**	**8.21***** **(0.01)**	**52.9**
	Without co-resident fathers n = 902 | N = 395,392	0.31*** (0.06)	0.25***	0.5*** (0.03)	8.41*** (0.01)	55.17
Females	**Benchmark sample** n = 1,522 | N = 548,234	**0.22***** **(0.06)**	**0.17*****	**0.42***** **(0.02)**	**6.42***** **(0.01)**	**53.28**

(continues in the next page)

Table 2.8 – Sensitivity of Intergenerational Mobility in Income to the Exclusion of Co-residents *(continued)*

		Elasticity	Correlation	Rank-rank slope	Prob.	Share
Females	Without co-resident fathers n = 1,377 | N = 506,252	0.23*** (0.07)	0.18***	0.43*** (0.02)	6.69*** (0.01)	55.01

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals earning more than their fathers are expressed in %. Correlations have the associated significance level of the elasticities used to compute them. The share of individuals earning more than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Overall, we can observe a slight change in the income mobility measures when comparing the original (benchmark) sample and the one without co-residents, i.e., an increase in persistence. However, although differences exist, the sizes of the potential biases may be considered negligible, as mobility measures are around the same values with and without co-residents in the sample. Previous literature on the topic, (e.g., Nicoletti and Francesconi, 2006) found a lower intergenerational income elasticity when using a sample of co-residents, only in comparison with a sample of parents and children who do not co-reside.

Table 2.9 – Sensitivity of Intergenerational Mobility in Education to the Exclusion of Co-residents

		Corr.	Prob.	Share
All individuals	**Benchmark sample** n = 2,549 | N = 980,083	**0.26*** **	**44.35*** ** **(0.01)**	**84.47**
	Without co-resident fathers n = 2,279 | N = 901,644	0.26***	44.69*** (0.02)	84.66
Males	**Benchmark sample** n = 1,027 | N = 431,849	**0.24*** **	**36.93*** ** **(0.02)**	**82.54**
	Without co-resident fathers n = 902 | N = 395,392	0.24***	37.19*** (0.02)	82.80
Females	**Benchmark sample** n = 1,522 | N = 548,234	**0.29*** **	**49.98*** ** **(0.02)**	**85.99**
	Without co-resident fathers n = 1,377 | N = 506,252	0.28***	50.36*** (0.02)	86.11

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals with more education than their fathers are expressed in %. The share of individuals with more education than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

For education, marginal differences are also verified. The work of Muñoz and Siravegna (2021), who use the first two measures, confirms this behaviour.

Summing up, both income and education mobility sensitivity analysis contain similar results to the benchmark estimates. Almost unchanged results may result from the small disparity regarding the sizes of the samples represented in the two scenarios analysed. This happens because there is not a high degree of co-residency. Our evidence is consistent with individuals leaving their parents' home, on average, before their 30s. According to the Eurostat[31], in 2018 the estimated age at which young people leave their parents' home is 26.3 years (27.2 for males and 25.2 for females) for the EU-27 – below Portugal, for which the age is around 28.2 years old (29.9 for men and 28 for women). In turn, the influence parents might exert on children is residual and this bias can also be ignored.

2.5.3.5. Attenuation Bias

For the parental income measure, we now compare our benchmark estimates using a single year to an estimate obtained by using a 4-year period average. Here the estimates based on income measured using a single year will be different from the ones presented in Table 2.3. This is because we must ensure that the pseudo-parental sample remains constant from year 1 to year 4 for results to be comparable. In other words, we have to guarantee that the same individuals remain in the different survey waves used to compute the average incomes. This allows us to make some inference about what might happen to our main estimates if we were able to keep the entire initial pseudo-parents sample, which would guarantee that the differences are mainly due to the number of years used to compute parental average income, instead of changes in the sample composition (Murray *et al*., 2018). According to Solon (1992), the larger the number of periods used to compute the parental average income, the more reduced the attenuation bias should be regarding the intergenerational income elasticity. The same can be considered for the rank-rank slope as shown in Chetty *et al*. (2014a). Results are presented in Table 2.10[32].

Table 2.10 – Sensitivity of Intergenerational Mobility in Income to Attenuation Bias

	Number of periods	Elasticity	Correlation	Rank-rank slope	Prob.	Share
All individuals n = 2,549 N = 980,083	**1 year**	**0.23*** (0.04)**	**0.17*** **	**0.52*** (0.02)**	**8.13*** (0.01)**	**58.37**
	4 years	0.28*** (0.04)	0.21***	0.64*** (0.02)	6.55*** (0.01)	48.73
Males n = 1,027 N = 431,849	**1 year**	**0.28*** (0.06)**	**0.22*** **	**0.56*** (0.04)**	**8.69*** (0.02)**	**58.65**
	4 years	0.33*** (0.06)	0.25***	0.67*** (0.03)	7.35*** (0.01)	49.41
Females n = 1,522 N = 548,234	**1 year**	**0.19*** (0.07)**	**0.13*** **	**0.47*** (0.03)**	**7.70*** (0.01)**	**58.16**

(continues in the next page)

[31] https://ec.europa.eu/eurostat/databrowser/view/ILC_LVPS08$DV_1041/default/table?lang=en

[32] For children, we could use longitudinal samples to also address the attenuation bias, but these do not contain retrospective information on parents. Besides, we cannot link longitudinal to the cross-sectional waves where that information would be available.

Table 2.10 – Sensitivity of Intergenerational Mobility in Income to Attenuation Bias *(continued)*

	Number of periods	Elasticity	Correlation	Rank-rank slope	Prob.	Share
Females n = 1,522 N = 548,234	4 years	0.24*** (0.06)	0.17***	0.59*** (0.02)	5.95*** (0.01)	48.19

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals earning more than their fathers are expressed in %. Correlations have the associated significance level of the elasticities used to compute them. The share of individuals earning more than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Regarding the intergenerational income elasticity, when we increase the time on which parental individual income is averaged to four years, we obtain an estimate that is 22% higher (from about 0.23 to 0.28). Mazumder (2005) simulates by how much intergenerational elasticity is attenuated when considering several periods on which the author averages parental income. The author shows that when four years are used, the estimates may be downward biased by about 31.3%. Our evidence suggests that attenuation bias plays a considerable role in our estimates. If we perform the same exercise as Mazumder (2005) on our benchmark estimates using the corresponding attenuation factor for a single year (0.526), the intergenerational income elasticity in Table 2.3 should be approximately 0.49 instead of 0.26. For men this elasticity would be around 0.57 instead of 0.3 and for girls 0.42 instead of 0.22. This attenuation bias would change the benchmark correlations as well: for children, males, and females, they would be equal to 0.38 instead of 0.20, 0.46 instead of 0.24, and 0.32 instead of 0.17, respectively[33].

Nybom and Stuhler (2017) and Murray *et al*. (2018) find the rank-rank specification to be more robust to attenuation bias than the log-log specification. Our estimates for the elasticity and rank slope may have also increased more than in some works (23% from 0.52 to 0.64 in our case), which use family income as a measure for parental income. This is the case of Chetty *et al*. (2014a), who consider that in this context using individual measures of economic status, such as individual income, may lead to larger differences when comparing estimates for different period averages because individual income fluctuates more across years. Income measured on a single year should also be noisy for the measures not directly influenced by the attenuation bias: we have a pattern of mobility declining for the remaining measures. Concerning the bottom to top income level probability, there is a fall of 19% (from 8.13 to 6.55%), and in the case of the share of individuals earning more than their parents this measure falls 17% (from 58.37 to 48.73%). Gender patterns are similar to the findings for the benchmark sample.

Overall, our analysis shows that results are robust to most of the sensitivity exercises. The rank-rank slope may be upward or downward biased, as found when performing the sensitivity analysis for different instruments for parental income. The other measures are likely to be attenuated. We also show

[33] We cannot state that, in opposition to our previous finding, results for Portugal would now be part of the most persistent countries in the literature, since the studies used for comparison can also suffer from attenuation bias: if this is true, the relative positions of the countries should remain the same.

that it is a fair choice to consider average total household income for married individuals instead of individual income, because household structure persistence influences the transmission of socioeconomic status.

2.5.4. Decomposing Intergenerational Mobility

Literature reports that individual's characteristics are associated with more or less mobility. We decompose intergenerational mobility in income and education by different individual characteristics to assess these previous findings. Hence, the benchmark analysis is extended by children/father education, occupation, income levels, and employment status. This analysis allows us to understand which characteristics are associated with more or less mobility.

2.5.4.1. Education

We present the disaggregation by own and father's education levels, in Table 2.11 and Table 2.12, respectively.

Table 2.11 – Results by Own Education Level

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Own Education Level	**Elasticity**	**Corr.**	**Rank-rank**	**Prob**	**Share**	**Corr.**	**Prob.**	**Share**
Medium n=1,291 | N = 477,439	0.19*** (0.05)	0.13***	0.33*** (0.04)	4.85*** (0.01)	49.44	-	-	87.51
High n=1,258 | N = 502,644	0.18*** (0.05)	0.20***	0.3*** (0.02)	11.51*** (0.02)	56.6	-	-	87.00

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities obtained using an ordered logit and the share of individuals earning more/with more education than their fathers are expressed in %. Correlations and the share of individuals earning more/with more education than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

In the majority of the indicators, individuals with a high education level present the highest relative and absolute mobility in income, pointing to the possibility that there is an absolute advantage in income of completing the highest education level. This is not in line with the findings of Blanden *et al*. (2005), who find a connection between higher educational attainment and more income persistence.

Table 2.12 – Results by Father Education Level

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Father Education Level	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Low n=2,040 | N = 745,593	0.2*** (0.05)	0.12***	0.37*** (0.02)	6.74*** (0.01)	44.38	-	-	100.00
Medium n=275 | N = 124,035	0.36** (0.15)	0.19**	0.49*** (0.08)	9.84* (0.03)	49.44	-	-	66.34

(continues in the next page)

Table 2.12 – Results by Father Education Level *(continued)*

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Father Education Level	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
High n=234 | N = 110,455	0.22 (0.35)	0.06	1.66*** (0.22)	-	17.69	-	-	0.00

Notes: Standard errors are in parentheses. ** and *** stand for statistically significant at 5% and 1% levels, respectively. Probabilities obtained using an ordered logit and the share of individuals earning more/with more education than their fathers are expressed in %. Correlations and the share of individuals earning more/with more education than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

The share of individuals earning more than their fathers is the measure for which results appear to have opposite findings, as reported in the work of Causa and Johansson (2010) for the OECD. The authors show that highly educated households are associated with more relative mobility, while we observe that children whose fathers have a low education level are the ones with higher relative mobility. The opposite happens for indicators of absolute mobility in comparison with the entire sample. Regarding education, when fathers have a low education level, children have more absolute mobility in comparison to the entire sample. Also, children of high educated fathers show more persistence in income than the sample for which we consider all individuals.

2.5.4.2. Occupation

Results by own occupation and father occupation categories are in Table 2.13 and Table 2.14 respectively.

Table 2.13 – Results by Own Occupation Category

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Own occupation	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Legislators, senior officials, and managers n=183 | N = 82,768	0.25 (0.15)	0.18	0.29*** (0.08)	17.87*** (0.06)	59.87	0.2***	61.71*** (0.05)	82.68
Professionals n=858 | N = 328,837	0.15** (0.07)	0.12**	0.21*** (0.03)	14.87*** (0.02)	59.5	0.03	95.27*** (0.01)	80.46
Technicians and associate professionals n=410 | N = 157,272	0.2** (0.09)	0.17**	0.36*** (0.07)	6.03*** (0.02)	52.21	0.2***	33.78*** (0.04)	81.45
Clerks n=298 | N = 110,206	0.02 (0.11)	0.02	0.11 (0.07)	4.51** (0.02)	53.34	0.24**	19.85*** (0.03)	88.38
Service workers and shop and market sales workers n=474 | N = 163,753	0.09 (0.08)	0.07	0.18** (0.07)	2.62** (0.01)	41.91	0.04	14.67*** (0.02)	89.81

(continued in the next page)

Table 2.13 – Results by Own Occupation Category *(continued)*

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Own occupation	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Skilled agricultural and fishery workers n=12 | N = 3,318	0.13 (0.59)	0.11	1.7** (0.75)	-	48.5	0.68*	3.25 (0.04)	84.88
Craft and related trades workers n=103 | N = 44,842	0.31 (0.2)	0.19	0.32*** (0.09)	3.21 (0.03)	53.14	0.33	9.78** (0.04)	95.34
Plant and machine operators and assemblers n=94 | N = 48,608	0.13 (0.09)	0.13	0.22 (0.19)	1.82 (0.01)	50.56	0.17	10.01** (0.04)	86.27
Elementary occupations n=103 | N = 34,717	-0.03 (0.12)	-0.02	-0.01 (0.11)	1.2* (0.01)	36.21	0.38**	9.14** (0.05)	81.92

Notes: Standard errors are in parentheses. *, ** and *** stand for statistically significant at 10%, 5%, 1% levels, respectively. Probabilities obtained using an ordered logit and the share of individuals earning more/with more education than their fathers are expressed in %. Correlations and the share of individuals earning more/with more education than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Mobility in income is always higher than in the benchmark sample when considering the (significant) subsamples of individuals with occupations in the following categories: legislators, senior officials, and managers, and professionals. The opposite occurs for skilled agricultural and fishery workers, plant and machine operators and assemblers, and elementary occupations. Subsamples where individuals have occupations classified in the technicians and associate professionals and also service workers and shop and market sales workers categories are the ones presenting more relative mobility but less absolute mobility in income in comparison with the entire sample. Regarding education, relative mobility is higher than in the benchmark sample, except when considering the subsamples where individuals work as skilled agricultural and fishery workers or have elementary occupations. Absolute mobility is also lower than in the sample with all individuals for the technicians and associate professionals' category and also for the elementary occupations.

Table 2.14 – Results by Father Occupation Category

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Father occupation	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Legislators, senior officials, and managers n=126 | N = 64,302	0.46** (0.22)	0.23**	0.75*** (0.06)	7.85* (0.05)	41.59	0.16*	71.86*** (0.06)	67.36
Professionals n=213 | N = 98,992	0.06 (0.27)	0.02	0.69*** (0.07)	4.42 (0.14)	14.58	0.34***	39.18*** (0.11)	20.51

(continues in the next page)

Table 2.14 – Results by Father Occupation Category *(continued)*

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Father occupation	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Technicians and associate professionals n=403 | N = 170,812	0.27* (0.15)	0.12*	0.2*** (0.06)	15.33* (0.08)	33.14	0.12*	54.01*** (0.04)	82.66
Clerks n=186 | N = 65,185	0.17 (0.23)	0.06	-0.47*** (0.16)	11.92* (0.07)	45.07	0.11	54.65*** (0.06)	94.21
Service workers and shop and market sales workers n=359 | N = 114,019	0.36** (0.15)	0.16**	0.69*** (0.13)	5.55*** (0.02)	41.14	0.1	42.18*** (0.04)	93.40
Skilled agricultural and fishery workers n=122 | N = 26,053	-0.17 (0.35)	-0.05	-1.39*** (0.25)	7.03 (0.05)	80.67	-0.07	31.83*** (0.07)	97.24
Craft and related trades workers n=593 | N = 246,746	0.12 (0.17)	0.04	0.13 (0.1)	6.2*** (0.02)	77.5	0.11*	43.24*** (0.03)	96.17
Plant and machine operators and assemblers n=362 | N = 143,457	0.49** (0.23)	0.01**	0.71*** (0.04)	5.37** (0.02)	62.59	0.04	35.45*** (0.03)	99.01
Elementary occupations n=185 | N = 50,517	-0.61 (0.6)	-0.12	-1.32*** (0.41)	5.54** (0.02)	87.96	0.18	-	100

Notes: Standard errors are presented between parentheses. *, ** and *** stand for statistically significant at 10%, 5%, 1% levels, respectively. Probabilities obtained using an ordered logit and the share of individuals earning more/with more education than their fathers are expressed in %. Correlations and the share of individuals earning more/with more education than their fathers do not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

In comparison with the entire sample, when fathers' occupations are classified in the legislators, senior officials, and managers as well in the professional's category, children always have lower relative mobility in income. In turn, the absolute income mobility is always higher than in the benchmark case when fathers belong to clerks and skilled agricultural and fishery workers occupations. Regarding the rank-rank slope, most of the professional categories show lower mobility, with exceptions being for technicians and associate professionals. Relative mobility in education presented in each subsample is higher than in the benchmark case for legislators, senior officials, and managers, technicians, and associate professionals, and craft and related trades workers, but lower for professionals. Absolute mobility in education is higher than in the benchmark subsample when children have fathers working as clerks, but lower than in the entire sample when considering the subsample of children whose fathers are classified as professionals.

2.5.4.3. Income Level

Table 2.15 and Table 2.16 present the results by own and father income levels, respectively.

Table 2.15 – Results by Own Income Level

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Own income level	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Low n=609 | N = 230,374	-0.09 (0.1)	-0.07	-0.03*** (0.01)	-	11.55	0.26***	28.11*** (0.03)	87.15
Medium-low n=1,007 | N = 390,927	0.03** (0.01)	0.10**	0.07*** (0.02)	-	55.00	0.22***	37.31*** (0.02)	87.28
Medium-high n=602 | N = 228,948	0.03*** (0.01)	0.14***	0.06*** (0.01)	-	71.68	0.22***	64.62*** (0.03)	85.23
High n=331 | N = 129,834	0.14*** (0.05)	0.21***	0.06*** (0.01)	-	88.43	0.19***	69.05*** (0.04)	69.93

Notes: Standard errors are in parentheses. *, ** and *** stand for statistically significant at 10%, 5%, 1% levels, respectively. Probabilities obtained using an ordered logit and the share of individuals earning more/with more education than their fathers are expressed in %. Correlations and the share of individuals earning more/with more education than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Mobility in income is in most cases higher than the one verified in the entire sample. For the majority of measures of intergenerational mobility in education, relative mobility is also higher in each income level partition when compared to the entire sample, while absolute mobility is always higher than in the benchmark case when individuals' income belongs to the medium-high category.

Table 2.16 – Results by Father Income Level

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Father income level	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Low n=275 | N = 68,264	5.33 (4.48)	0.18	6.18* (3.17)	-	89.27	0.07	35.68*** (0.05)	98.95
Medium-low n=1262 | N = 488,814	0.12 (0.12)	0.03	0.22*** (0.02)	-	67.86	0.13***	40.68*** (0.02)	97.84
Medium-high n=451 | N = 182,071	0.41 (0.37)	0.08	0.82*** (0.17)	-	43.8	0.04	58.67*** (0.04)	91.08
High n=561 | N = 240,934	0.22 (0.17)	0.08	0.63*** (0.08)	-	19.99	0.33***	47.62*** (0.04)	48.25

Notes: Standard errors are in parentheses. * and *** stand for statistically significant at 10% and 1% levels, respectively. Probabilities obtained using an ordered logit and the share of individuals earning more/with more education than their fathers are expressed in %. Correlations and the share of individuals earning more/with more education than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Only when fathers are classified as medium-low income earners do children present more (relative and absolute) mobility in income than the one in the entire sample. When individuals have fathers in medium-high- and high-income levels, (relative and absolute) mobility in income appears to be lower than in the benchmark case. Relative mobility in education is higher than the entire sample when parents have a medium-low income level and is lower than in the entire sample when parents have a high-

income level. In terms of absolute mobility in education, for children of parents with medium-high- and high-income level, mobility is higher.

2.5.4.4. Status in Employment

In this section we analyse children by their status in employment (either self-employed or employees), with results in Table 2.17.[34]

Table 2.17 – Results by Status in Employment

	Intergenerational Mobility in Income					Intergenerational Mobility in Education		
Own status	**Elasticity**	**Corr.**	**Rank-rank**	**Prob.**	**Share**	**Corr.**	**Prob.**	**Share**
Self-employed n = 155 | N = 70,871	0.44** (0.18)	0.23**	0.62*** (0.11)	6.76** (0.03)	53.11	0.29***	53.18*** (0.06)	80.64
Employee n = 2,260 | N = 851,447	0.23*** (0.04)	0.19***	0.41*** (0.02)	6.61*** (0.01)	53.89	0.25***	42.66*** (0.02)	85.70

Notes: Standard errors are in parentheses. ** and *** stand for statistically significant at 5% and 1% levels, respectively. Probabilities obtained using an ordered logit and the share of individuals earning more than their fathers are expressed in %. Correlations have the associated significance level of the elasticities used to compute them. The share of individuals earning more than their fathers does not have an associated significance level. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Overall, children who are self-employed present higher persistence in income and education when compared to the entire sample, while the opposite occurs for the subsample of children who are employees. For the first group the exception is the low to high education level probability. For the last group, lower mobility is verified for the bottom to top income level probability and the low to high education level probability.

Although Portugal has a positive framework in terms of mobility in education, the above exercises allow us to conclude that differences across groups in the Portuguese economy still exist. Hence, policies targeted at directly promoting mobility in education should also be considered[35]. These should close the gap between advantaged and disadvantaged individuals by improving the outcomes of the latter. Among the ones summarized by Narayan *et al.* (2018), we first have early childhood development with measures aimed at filling the intra-uterine deprivations, eliminating nutritional and health problems and making childcare accessible to all children. Second, one should consider improving the access to education as well as its quality, complemented with an efficient investment of public resources in education. Our findings suggest that more mobility is associated with attaining a high education level, so this policy direction is of utmost importance in the Portuguese context for that stage of education. Third, decreasing

[34] This exercise cannot be performed by fathers' employment status, because we cannot distinguish in the sample of children the cases for which their fathers were employees only, employers only, or both.

[35] We note that the adoption of mobility-enhancing policies depends on the mobility drivers which are country specific. Uncovering them was not the goal of our work, although it is a suggestion for future research. Also, for the same intervention different implementations may occur and these also depend on the context. Even though this is the case, we address some policy directions that are found to be important in existing research.

the segregation across space, strengthening institutions (through social capital), housing, and infrastructure and promoting safety may have a positive effect on mobility as well (Brown and Richman, 1997; Chetty *et al*., 2014a, 2016; Kline and Moretti, 2014; Chetty and Hendren, 2018a, 2018b).

Policy makers can also target education and income mobility to promote a feedback effect, having a long-term perspective. This is because if the current generation sees its mobility in income improved, their ability to invest in their children is promoted, which impacts education mobility and therefore income mobility again. Easing the access to capital markets through credit to finance not only education but also entrepreneurial undertakings should be a way to do so.

2.5.5. The Relationship Between Relative Mobility in Income and Education

In the previous disaggregation there were cases in which, when compared to the benchmark scenario, mobility in income and education were jointly higher or lower: these occurred for the subsamples of own and father high education levels, own elementary occupations category, and medium-high own income level. To measure the impact that intergenerational persistence in education has on intergenerational persistence in income we use the OLS point estimates for ϱ^c and ϱ^p. We also use two samples moments: $Var\left(Ed_{it}^{p}\right)$ and $Var\left(y_{it}^{p}\right)$. Table 2.18 presents the results.

Table 2.18 – The Relationship Between Relative Intergenerational Mobility in Income and Education (in years)

	Father-children (max)	**Father-children (min)**	**Father (max) - children (min)**	**Father (min) - children (max)**
ϱ^c	0.028 [0.028; 0.029]	0.032 [0.031; 0.033]	0.032 [0.031; 0.033]	0.028 [0.028; 0.029]
ϱ^p	0.03 [0.029; 0.031]	0.027 [0.026; 0.027]	0.030 [0.029; 0.031]	0.027 [0.026; 0.027]
$Var\left(Ed_{it}^{p}\right)$	5.822	5.822	5.822	5.822
$Var\left(y_{it}^{p}\right)$	0.172	0.146	0.172	0.146
$d\beta_1/d\partial_1$	0.029 [0.027; 0.030]	0.034 [0.032; 0.035]	0.033 [0.031; 0.034]	0.030 [0.029; 0.031]

Notes: max and min stand, respectively, for the maximum and minimum taxation possible for each generation. The lower and upper bounds of the 95% confidence interval are in square brackets.

While β_1 is an elasticity, ∂_1 is an absolute impact. We log education to make the interpretations easier as ∂_1 will be considered as an elasticity. Table 2.19 shows robust results to either specification.

Table 2.19 – The Relationship Between Relative Intergenerational Mobility in Income and Education (in logs)

	Father-children (max)	**Father-children (min)**	**Father (max) - children (min)**	**Father (min) - children (max)**
ϱ^c	0.281 [0.273; 0.288]	0.317 [0.310; 0.325]	0.317 [0.310; 0.325]	0.281 [0.273; 0.288]

(continues in the next page)

Table 2.19 – The Relationship Between Relative Intergenerational Mobility in Income and Education (in logs) ***(continued)***

	Father-children (max)	Father-children (min)	Father (max) - children (min)	Father (min) - children (max)
ϱ^p	0.254 [0.248; 0.261]	0.227 [0.221; 0.233]	0.254 [0.248; 0.261]	0.227 [0.221; 0.233]
$Var(Ed_{it}^p)$	0.078	0.078	0.078	0.078
$Var(y_{it}^p)$	0.172	0.146	0.172	0.146
$d\beta_1/d\partial_1$	0.032 [0.031; 0.034]	0.038 [0.037; 0.040]	0.037 [0.035; 0.038]	0.034 [0.032; 0.036]

Notes: max and min stand, respectively, for the maximum and minimum taxation possible for each generation. The lower and upper bounds of the 95% confidence interval are in square brackets.

There is a positive relationship between intergenerational relative persistence in education and income, although the link between the two is modest: a one percentage point increase in educational persistence makes income persistence increase by approximately 0.03-0.04 percentage points. One could think about the scenario in which for the current generation there would be a unit elasticity scenario. Fixing $\varrho^p, Var(Ed_{it}^p)$ and $Var(y_{it}^p)$ at the values in Table 2.19 that would only occur if $\varrho^c \in \{8.2; 8.7\}$, which does not seem reasonable. Summing up, although limited, there is a positive effect on relative income mobility improvements from relative mobility in education improvements, in Portugal. This relationship is supported by evidence presented by Fletcher and Han (2019) for the USA, which finds a positive correlation between relative income and education mobility.

In the World Bank report by Narayan *et al.* (2018), the labour market is referred to as one of the main transmission mechanisms between education mobility and income mobility. The relationship between education mobility and income mobility should be positive but probably not very strong, as in our case, because of labour market weaknesses. These fragilities are likely to be stronger in recession periods, as it occurred in the 2008 financial crisis, which may have affected individuals in our sample. Research points out that there is a long-term disadvantage for individuals entering the labour market in a recession period, where the unemployment rate at that time can influence future labour-related outcomes (e.g., Altonji *et al.*, 2016). During the 2008-2013 period, the unemployment rate in Portugal increased about 8.5 percentage points (from 7.2 to 15.7%)[36]. The idea is that the higher unemployment is, the lower the likelihood of matching between labour demand (firms) and labour supply (workers). In turn, reservation wages become lower for workers, who accept suboptimal offers, and then lifetime earnings may be affected in a negative way. Moreover, high unemployment during a specific period is detrimental for experience accumulation and leads to skills depreciation, scarring effects, and psychological discouragement (Pissarides, 1992; Arulampalam *et al.*, 2001; Clark *et al.*, 2001; Raaum and Røed, 2006). These effects are especially important for young individuals, who try to get their first job in such conditions: for individuals aged less than 25 years old, the unemployment rate in Portugal

[36] https://www.pordata.pt/en/Portugal/Unemployment+rate+total+and+by+age+group+(percentage)-553-5397

in 2008 was about 16.7% and in 2013 was around 38.3%[37]. Hence, even if educational mobility appears to be higher, its potential positive effect on income mobility may be offset by the negative long-term consequences on unemployment. Besides, results may be shaped by the data as the public sector and household employees, self-employed individuals, and civil servants are not considered.

Narayan *et al.* (2018) identify unemployment benefits and access to a better access to social protection systems as policy measures that should reduce the negative lagged effects of unemployment when it takes place. Moreover, to mitigate the increased penalization of youth and facilitate their integration in the labour market for the first time when recessions take place, existing works point to wage subsidies and taxes reductions, incentives to hire young individuals, and subsidized training/employment (Katz, 1998; Coenjaerts *et al*., 2009; Betcherman *et al*., 2010; Groh *et al*., 2016; Chen *et al*., 2018) as possible measures to be adopted or reinforced. This topic needs further research attention.

2.6. Concluding Remarks

Published work on intergenerational mobility in Europe has been focused on Scandinavian countries while research on Southern Europe is still limited. In this group, literature is scarce for Portugal in terms of income mobility, although some developments have been made regarding the study of educational mobility. Our work analyses intergenerational mobility in income and education for this country by constructing several relative and absolute measures of intergenerational mobility. For income mobility we compute the intergenerational income elasticity, the intergenerational correlation coefficient, the rank-rank slope, the share of individuals earning more than their parents, and the bottom to top income level probability. For education mobility we compute the intergenerational education (Pearson) correlation, the low to high education level probability, and the share of individuals with more education than their fathers. Both income and education mobility measures are complemented by ordered logit transitions matrices. We uncover the patterns that exist and which individual characteristics present more or less mobility, for individuals born in 1968-1988. Both genders are considered. Two Portuguese components of European datasets are used: the European Community Household Panel and the European Union Statistics on Income and Living Conditions.

Our benchmark results reveal gender differences, showing that women generally present higher mobility in income than men, a finding also for Russia, Denmark, the USA, Italy, and in other studies that included Portugal. When considering transition probabilities between income levels, we observe that there is a strong degree of intergenerational mobility when fathers are at the low-income level but upward probabilities decrease the higher the father's income level. Our value estimates are according to estimates previously done. As in the case of income, women have a higher probability of passing from a low to a high education level than men, with previous studies for Portugal reaching lower probabilities

[37] https://www.pordata.pt/en/Portugal/Unemployment+rate+total+and+by+age+group+(percentage)-553-5398

than ours. In Portugal the share of individuals with more education than their fathers is higher than 80% and the probability of staying in a low education level, if that is the case of the father, is 0%, a finding that improved relative to other estimates for Portugal and is higher than previous estimates for Hong Kong and Austria. The likelihood that an individual has of reaching or remaining in the high-education level is increasing on the father's education level, confirming published findings.

We decompose our mobility measures to check which own and father's characteristics are associated with more or less income and education mobility when compared to the benchmark sample. We assess characteristics such as education level, occupation, income level, and status in employment. Contrary to what is found in the literature, individuals with a high education level show higher income and education mobility. This is a further advantage of having more education and corroborates the findings for occupations in the legislators, senior officials, and managers and professionals' categories for which mobility in income is higher: they require a higher education level than occupations as skilled agricultural and fishery workers, and plant and machine operators and assemblers, which show lower income mobility when compared to the benchmark sample. Also, individuals with elementary occupations always present lower mobility and absolute mobility in education is higher when fathers work as clerks. Also, a finding against previous literature, children whose fathers have a low education level are those presenting higher relative income mobility. Mobility in income and education is higher for individuals in the medium-high income level and more absolute mobility in education occurs for these individuals when fathers also belong to the medium-high-income level category. However, medium-low income fathers bring more mobility in income to their offspring. Self-employed individuals present lower income mobility when compared to the entire sample. *Vis-à-vis* these results, policies such as the ones proposed in Narayan *et al*. (2018), which promote early childhood development, access to quality education, an efficient public investment in education, the reduction of segregation, strong institutions, and infrastructures should help to close the gap between advantaged and disadvantaged individuals. The ease of access to capital markets and a robust economic growth should have a feedback effect of education and income mobility in future generations.

Based on some of our earlier results, we performed a final exercise using Mincer (1974) equations and the *Quadros de Pessoal* Database. We show that there is a positive effect of mobility in education on income mobility, in relative terms, although not very strong. We consider that this relationship may be mediated by a weak labour market. We argue that there may be lagged harmful effects of the unemployment verified during recessions (as the 2008 financial crisis), which offset the increase policy makers may be interested in making in relative education mobility. Narayan *et al.* (2018) point to unemployment benefits, social protection, wage subsidies, and tax reductions as possible policies to make the previous relationship stronger, although further work on this topic needs to be done.

We have also performed sensitivity analyses to some of our initial methodological hypotheses, namely in terms of income definition (average total income and individual income, parental income, inclusion of individuals with no individual income, co-resident bias, and attenuation bias) to determine

if our benchmark results hold. There is some degree of intergenerational persistence in household structure, as reported in previous literature. Furthermore, if we consider four years, instead of one year, to estimate parental income, the designated attenuation bias, this changes the results. However, neither the inclusion of individuals with no individual income or taking in consideration two generations living in the same home affect our benchmark results.

Some shortcomings can be pointed to our work. First, the datasets we use do not provide direct information on father's income when children were around 14 years old. Following the literature, we predict parental income which is not observed, through education, occupation, and managerial position information, i.e., by using a pseudo-parent's sample, which has implications for results. Second, our datasets provide a set of retrospective questions about parental characteristics, which allow us to predict their income. But the range of available characteristics is insufficient: the higher the number of instruments to proxy for parental permanent income, the more unique values parental income could assume, which increases the heterogeneity of the pseudo-parents' sample. Regional proxies are a simple example that would fill this need. Third, it is possible to follow individuals in both generations over time, but for the children's subsample this is done at the expense of no retrospective questions about parents, and hence mobility could not be computed. This implies that analysis that require addressing temporal behaviours for the measures we compute to complement our cross-sectional framework cannot be performed. Although we try to make the possible adaptations, the final analysis of the effect of education mobility on income mobility is not performed using the same sample as in the other analysis, which can influence the results. Overcoming these problems would improve our work, although it is a difficult task, since the majority of problems are due to the nature of the data supplied by the existing surveys. Finally, the biases we address are mainly studied in the literature for relative mobility measures, but the restrictions to avoid them should influence absolute mobility measures as well. This topic needs further research. We also recognize the need for future research to investigate what drives mobility in Portugal to uncover which specific policy actions should take place to improve mobility in this country.

3. Mother! Father! What Have You Done? The Contributions of First Names or Surnames for Generational Mobility

Executive summary: We construct two indexes based on a measure by Güell *et al.* (2015); the Informational Content of First Names (ICF) and the Informational Content of Surnames (ICS30) to understand the impact of the choice of first names and the weight of family names on children's educational outcome. Our purpose is to measure generational mobility in education for Portugal, a country that previous literature has emphasized as exhibiting a high persistence in education. Surnames explain about 14% of the observed variability of educational attainment, a significant proportion, and first names are responsible for only 2%. Additionally, we explore mobility patterns across space. While first names present more informational content in the country's coast than in its inner regions, the opposite occurs for surnames, even when accounting for internal migration. The informational content of surnames is greater if the retention and desistance rate in primary and lower secondary education are higher. The opposite occurs for inequality, when considering the P80/P20 ratio. A convex relationship between the informational content of surnames and the P90/P10 ratio and the Gini Index is found. Policies targeting information provision, social protection, a healthy labour market, and economic growth should mitigate the positive relationship between inequality and mobility.

JEL Classification: E24; I24; J62; O15.
Keywords: Generational Mobility in Education; Inequality; Informational Content of First Names; Informational Content of Surnames.

3.1. Motivation and Main Findings

We study the relevance of information contained in Portuguese first names and surnames regarding generational mobility in education. We use the 2021 wave of the European Union Statistics on Income and Living Conditions (EU-SILC) to obtain information about education and the Orbis Database to collect information on individuals and their first and family names. Our analysis is developed for men aged between 25 and 64 years old.

The study of intergenerational mobility in Portugal is of utmost importance, as this country has consistently been identified as being highly persistent in education. In 1999, in an International Monetary Fund (IMF) study, Clements showed that Portugal was lagging behind the other Organization for Economic Cooperation and Development (OECD) countries regarding the share of individuals reaching a secondary education level. Narayan *et al.* (2018) from the World Bank showed 20 years later that among the high-income countries, Portugal stands out as the one with the highest persistence in education level. In 2022, the Bank of Portugal confirmed the existence of a persistent framework for education. Persistence in education jeopardizes a country's economic development and growth (Narayan

et al., 2018). Resource allocation is inefficient, a vicious cycle with high inequality may exist, and individuals' aspirations and perceptions about fairness and justice may be harmed, which has consequences for social stability.

Several authors have studied intergenerational mobility in education for Portugal. Most of them reinforce the international organizations' concerns, confirming the high degree of persistence for this country. Carneiro (2008) finds that the share of individuals not completing high school when fathers did not attain primary education is high, while it is low considering people not completing high school when father have a tertiary degree. Pereira (2010) shows that the likelihood of being highly educated when parents are highly educated is greater in comparison with children of less educated parents. Bago d'Uva and Fernandes (2017) point out that Portugal still stands behind the European Union in terms of mobility, although improvements over time have been verified, namely regarding the share of children with more education than their fathers. Comi (2003) finds that persistence in education is higher in Portugal, Ireland, and the Mediterranean countries. The share of highly educated parents with highly educated children is low. In Nybom's (2018) work, Portugal is also amongst the countries with a higher degree of persistence, considering high-income economies, comparable with Uruguay and Hungary. Only Schneebaum *et al.* (2014), who studied 20 European countries, found that mobility is the highest in Portugal when considering men, although for women, France, the Anglo-Saxon and Nordic countries, Poland, Czech Republic, and Greece surpass the Portuguese mobility.

A new branch of literature in the context of mobility studies uses surnames to measure how the socioeconomic status, either through income or education, persists throughout generations. Surnames are markers that reflect individuals' belonging to a specific family. This literature is still scarce, and none of it examines Portugal.

Clark and Cummins (2012) study mobility in England during the period 1800-2011. Persistence is verified throughout time and mobility rates are lower when compared to the conventional estimates. Clark and Cummins (2014) follow the previous work and draw the same conclusions for England and Wales for the period 1858-2012. Clark and Ishii (2012) show that persistence is high for the descendants of modern as well as former elites, when studying Japan for the period 1868-2012. Mobility is lower than in comparable studies for the USA, Sweden, and UK and also lower than the conventional mobility estimates. Clark and Landes (2013) study India's mobility, which appears to be low from 1860 to 2012 due to marital endogamy. Clark *et al.* (2017) find that persistence in socioeconomic status was strong in Australia between 1870 and 2017, and similar to that found for England and the USA. Collado *et al.* (2013) analyse mobility in Spanish regions from the end of the 19th century to the start of the 21st century, and show that there is a strong connection between the socioeconomic class of children and one of their great-grandfathers and great-great grandfathers. Chetty *et al.* (2014a) analyse intergenerational mobility for the USA, considering the 1980-1982 birth cohorts. The authors find that there is spatial variability in mobility and that mobility measures using surnames yield results similar to the conventional estimates. Barone and Mocetti (2021) examine intergenerational mobility in the long run

in Florence, by looking at surnames of individuals living in the city between 1427 and 2011. The authors are able to show that intergenerational correlations persist throughout the six century time-span and are higher than suggested by other works not using name related information. Hao (2021) estimates intergenerational mobility in China during the period 1644-1949. Mobility is low in all Late Imperial, Republican, and Communist eras. For the last, the author's estimates are lower than the conventional ones computed for the UK, USA, Scandinavia, and China. Bukowski *et al.* (2022) study mobility in Hungary from 1949 to 2017. The authors find that persistence was high for low- and high-class families, and similar between communist and capitalist regimes. Privilege was still verified for eighteenth century noble-class descendants.

While the above works use several cross-sections of data, information for Portugal regarding surnames is available in a single cross-sectional snapshot. Therefore, we apply a methodology developed by Güell *et al.* (2015) to measure intergenerational mobility in education based on surnames, which they have applied to Catalonia, Spain. We follow their claim that an individual's surname explains their socioeconomic outcomes, which is equivalent to arguing that being born into a specific family (with that specific surname) matters for their socioeconomic outcomes, since there are intrinsic family characteristics that individuals inherit that are responsible for part of their socioeconomic outcomes. The more surnames explain about differences in socioeconomic outcomes of a cross-section of individuals the more those outcomes depend on the past and greater persistence is likely to be verified. With this purpose, they created the Informational Content of Surnames (ICS) measure, which reflects the share of the variability of socioeconomic outcomes that may be attributable to surnames as a measure of intergenerational persistence. Güell *et al.* (2015) use census data, the telephone directory, and tax returns to find that education mobility decreased throughout the 20th century due to the increase in the degree of assortative mating in Catalonia (a Spanish region) considering the year of 2011. Güell *et al.* (2018) extend the previous work for Italian provinces in 2005, correlating it with social and economic outcomes: significant relationships are found for economic activity, education, social capital, and inequality.

In Portugal, parents have very limited discretion in selecting surnames for their children which does not occur with their offspring first names. As pointed out by Levitt and Dubner (2006) in their well-known book *Freakonomics*, the name chosen by parents and given to children may influence their life as well. The decision of first naming is connected with parental economic status at the time of their child's birth (Clark *et al.*, 2015), their preferences for the future conditional on their status, and also influenced by several factors including cultural and racial features.

The influence the naming decision has on individuals' future finds support in existing studies. Figlio (2005) uses data from a Florida school district testing the hypothesis that individuals with names associated with a poor socioeconomic status are treated differently by school administrators, leading to low performance in school. The author finds that the hypothesis is empirically verified and reflected in differences in test scores. Kalist and Lee (2009) study the relationship between first name popularity

and crime in an American State, showing that unpopular names are connected with living in a low socioeconomic status county, having a disadvantaged home environment (single parents or female headed households), which increases the likelihood of committing crime. Aura and Hess (2010) show that first names may be connected with individuals' education, relative financial position, social class, and occupational prestige. Bertrand and Mullainathan (2004) find that in Massachusetts, due to the racial information conveyed by the name, in order to get a job interview people with names associated with black ethnicity need to send their application to job openings 50% more times than people with a name associated with white ethnicity. Fryer and Levitt (2004) find relationships for individuals in California between a person's name and the probability of being black, the number of children born, babies' low birth weight, being an unmarried parent, and the absence of health insurance (which can affect labour productivity). They show that the blackness of a name reveals the parental background. For a Florida school district, Figlio (2007) shows that suspension from school is more likely to occur for boys with names that are usually given to women, especially for black individuals. Aura and Hess (2010) also find that non-white non-black individuals with blacker names perform financially more poorly than blacks with predominantly whiter names.

Following the same reasoning, we argue that the more responsible first names are for differences in socioeconomic outcomes in a cross section of individuals, the more important should be the choice of parents regarding their children's first names. We will therefore understand the explanatory power of first names regarding differences in individuals' outcomes. With this purpose, we use an analogous to the ICS measure. We call it the Informational Content of First names, ICF.

Our main contribution to the literature is clear. By using surnames we compute intergenerational persistence in education for Portugal without needing intergenerational links. We therefore overcome the need for data availability for at least two generations. When this type of information exists it is often presented with a different degree of disaggregation between parents' and children's educational levels. Although one can transform it into years of education, when making both generations comparable in terms of education categories, information will be lost. Through a single cross section of data, this is no longer a problem. We add to the existing works applying this method. We follow Güell *et al.* (2018) and analyse surnames belonging to less than 30 individuals (ICS30) because they form the partition where they should belong to the same family. We also consider that first names may be responsible for individuals' socioeconomic status. We therefore extend our analysis to assess the informational content of first names, ICF, studying if parents' choices about naming have a role in the differences in educational attainment for Portugal.

Results show that when considering rare surnames, i.e., the ones held by fewer than 30 individuals, these explain 14% of the variability observed regarding educational attainment. The rarer surnames are, the more informative those surnames are regarding persistence, as found by Güell *et al.* (2018). When analysing first names, the evidence suggests that first names are responsible for only about 2% of the differences between individuals' educational attainment.

The second contribution is about performing an analysis for small geographies. We are also, to the best of our knowledge, the first to exploit the Portuguese spatial variability in intergenerational persistence.[38] We do the same for the ICF measure.

We find that the ICS30 is higher in the country's interior in comparison with its coast, meaning that intergenerational persistence is greater in the first case, in line with findings for African countries (Alesina *et al.*, 2021). Since internal migration may influence the informational content of surnames (migrants can have different outcomes in comparison with natives), we compute a regional index for the informational content of surnames, Regional ICS30. This considers the partition of the 50% most common surnames in a given region, conditional on being rare in the country. Clear differences emerge in comparison with the ICS30, meaning that when making comparisons across space this is the measure used. When analysing the ICF by regions, differences across space are not as pronounced as with surnames. Besides, the informational content of first names appears to be, on average, higher in the country's coast than in its interior, which is the opposite to the case with surnames.

Finally, and given that the variability of the informational content of surnames across space is twice as much as the one for first names, we consider the relationship between intergenerational persistence and different socioeconomic and political regional outcomes, which change across the territory: these are related with economic activity, education, inequality, social capital, labour market, trade openness, life expectancy, suicides, and crimes.

Our evidence is that economic activity, social capital, labour market outcomes, life expectancy, and suicide rates are not associated with the regional differences in the information provided by surnames. Results suggest that higher retention and dropout rate in primary and lower secondary education have a positive relationship with the informational content of surnames, resembling the evidence found for Latin American countries and the USA (Daude and Robano, 2015; Hilger, 2016), respectively, while a negative weak relationship appears to exist for imports. Considering other socio-political outcomes, there is no pattern for the relationship when looking at crimes. Interestingly, the informational content of surnames presents a robust negative linear relationship with inequality when analysing the P80/P20 ratio, while a convex relationship exists for the P90/P10 ratio as well as the Gini Index.

This essay is organized as follows. In Section 3.2 we present the variables and data. In Section 3.3 our results for the informational content of surnames are discussed. In Section 3.4 we compute and analyse results for the informational content of first names. Section 3.5 concludes.

[38] This was also done in the works of Daude and Robano (2015) for Latin America, Causa and Johansson (2010) for the OECD, Schneebaum *et al*. (2014) for Europe, Neidhöfer and Stockhausen (2018) for Germany when compared to other countries, Fletcher and Han (2019) for the USA, Emran and Shilpi (2015) for India and Latin America, Azam and Bhatt (2015) for India, Choudhary and Singh (2017) for India and China, and Geng (2021) for China, the USA, Nordic countries, and Europe.

3.2. Databases and Variables

In this section we present a description of the databases and variables used in our work.

3.2.1. Databases

We use two main datasets. The first is Orbis, from Moody's Analytics, a database that contains information on public and private companies operating in Portugal, namely on individuals associated with those companies. Full names of these individuals are taken from this dataset, along with the economic activity classification in which they are employed. These individuals are the companies' current and previous managers and directors. The second database is the Portuguese component of the European Union Statistics on Income and Living Conditions (EU-SILC) survey, the *Inquérito às Condições de Vida e Rendimento das Famílias*. We use the 2021 wave of this panel, which contains information for 2020, and extracted data for men between 25 and 64 years old. It provides information on socioeconomic and demographic characteristics, such as education and economic activity classification.

3.2.2. Variables and Sample Restrictions

3.2.2.1. Economic Activities Classification

Our work considers the second revision of the Statistical Classification of Economic Activities in the European Community, NACE. Both databases we use contain this classification for each individual according to the company in which they work. It consists of a two-digit code, which categorizes data according to economic statistics (fields such as production and employment, amongst others) and other areas designed within the European Statistical System. Data on this variable are available by numerical code ranges, which are then transformed into string codes. This classification can be seen in Table C1 in the Appendix C.

3.2.2.2. Education

Data on education are not available in the Orbis Database. We therefore need to proxy this variable through the NACE information. We can find the average education years in the EU-SILC, for each of the statistical classifications of the NACE, and attribute them to the cases for which individuals work in those economic activities in the Orbis Database. Education is classified according to the 2011 categorization of the International Standard Classification of Education (ISCED) of the United Nations Scientific and Cultural Organization (UNESCO). We make a correspondence between education categories and the minimum years of education required in cumulative terms for each level. These are presented in Table 3.1, grounded on Schneebaum *et al.* (2014) and Narayan *et al.* (2018).

Table 3.1 – Correspondence between ISCED Classifications and Minimum Required Years of Education

EU-SILC 2019 (ISCED 2011)	Minimum Years of Education
Primary	6 years
Lower secondary	9 years
Upper secondary	12 years
Post secondary non tertiary	13 years
Tertiary	15 years

We consider only individuals who are older than 25 years, as by this age they should have finished full-time education. Since in 2020 the enrolment rates in schools were 10% of the Portuguese individuals between 25 and 29 years old, 5% for those aged between 30 and 34, 3% for the 35-39 age range, and 1% for 40-64 years old[39], we also ensure that the majority of individuals in the EU-SILC wave are not pursuing any type of education and have finished schooling when the survey took place. Also, we exclude individuals older than 64 to prevent a bias from different survival rates across distinct social backgrounds (Behrman *et al.*, 2001; Urbina, 2018). Table 3.2 has the average years of education in the EU-SILC, for each of the statistical classifications of economic activities and considering men aged between 25 and 64 years old in 2020.

Table 3.2 – Average Education Years *per* NACE Category in the EU-SILC

NACE string codes	Average years of education
A	7.75
B-E	9.62
F	8.08
G	9.81
H	9.71
I	9.58
J	13.18
K	13.33
L-N	12.03
O	10.71
P	14.11
Q	12.61
R-U	10.99

3.2.2.3. The Informational Content of Surnames and the Distribution of First Names and Surnames

[39] https://stats.oecd.org/Index.aspx?DataSetCode=EAG_ENRL_RATE_AGE

3.2.2.3.1. The Informational Content of Surnames (ICS)

Designed by Guëll *et al*. (2015), the informational content of surnames (ICS) is the primary intergenerational mobility measure adopted in our work. We describe it as follows. Consider that each individual i in a cross-section of N persons has a surname s and that his or her economic well-being regarding income and education is defined as y_{is}. In this way, we may define

$$y_{is} = \partial' X_{is} + b'D + \text{error} \tag{3.1}$$

and

$$y_{is} = \partial' X_{is} + b'F + \text{error} \tag{3.2}$$

where X_{is} is a vector of his or her demographical characteristics, D is a vector of name-dummy variables, with $D_s = 1$ if that individual has a particular surname s and $D_s = 0$ otherwise, and F is a vector of fake surname-dummy variables, on which surnames are reshuffled/reassigned in a random way (so they cannot be informative) but maintaining their original marginal distribution.

In order to attribute a fake surname to an individual, we proceed as follows. We save the original surname's variable in an auxiliary dataset. We then generate a random variable which is sorted so that surnames are reshuffled. Reshuffled surnames become fake surnames. Finally, we merge the two datasets. We end up in the original dataset in which individuals now have an assigned fake surname and their true surname.

The first equation's R^2 is denoted as R_L^2. Since the second regression results from a random reshuffling, we replicate it 10 times, following Güell *et al*. (2018), and compute an average of the regressions' R-squared, denoted by $\bar{R}_F^2$, so that the ICS is redefined as:

$$ICS \equiv R_L^2 - \bar{R}_F^2 \tag{3.3}$$

Because the number of surname-dummy variables is large in the regressions, this definition avoids spurious information to be attributed by surnames. The same methodology is applied to compute the informational content of first names, ICF.

In the Orbis Database, names are coded in a single string variable[40]. To separate first names and surnames we proceed as follows. First, we look at the cases for which only two separate words are contained in the string: this means that the first word is the person's given name and the second is his or her surname. Second, we consider strings with more than two words. We used an auxiliary dataset – the DicPRO (Baptista *et al*., 2006) – which is a dictionary containing Portuguese names, 1,935 of which are classified as first names, and 4,200 classified as family names (surnames). Considering that the initial name is always a first name and that usually a person does not have more than two first names, we

[40] Connection words between main words of the string are eliminated (e.g., *de*, *da*, *do*, *e*).

checked whether the second names were classified as possible given names in the DicPRO. If not, we classified it as a surname. For each person, we kept up to two first names and up to four family names, which covers all the possible names that have been manually analysed.

We are only interested in studying Portuguese individuals. However, the Orbis Database provides information about the nationality for just a few individuals, but companies in Portugal may also have non-Portuguese managers. Individuals with foreign citizenship are excluded and individuals with Portuguese citizenship are considered in the analysis. Also, individuals with non-defined nationality are subjected to the following criterion. We analyse the individuals' first names and consider that someone is Portuguese if at least one of the given names is contained in the auxiliary datasets. If not, that individual is dropped from the sample[41]. This criterion was applied because the chances of a given name that appears in a dictionary with Portuguese names has of belonging to a foreign individual are low. Besides, this avoids the potential bias on the ICS estimates if foreign individuals are considered: the educational attainment of migrants may differ from that of the Portuguese natives, meaning that surnames of migrants can introduce additional information that does not regard native families.

Men and women in Portugal are allowed to change their surnames upon marriage by adding their spouses' surnames to their own surnames. We believe this is more likely to occur for women, meaning that when looking for married women's surnames, the likelihood that we are analysing their husband's or wife's surname is high. Therefore, it will not tell us much about family linkages of that specific female. Considering that we cannot identify the marital status of women, and in the event that they are married, we cannot know whether or not the surnames' aggregation took place, we prefer not to consider women.

3.2.2.3.2. The Surname Distribution in Portugal

The reasoning behind using the informational content of surnames is that persons with a specific surname belong to the same family, i.e., surnames reflect family linkages. However, it is not true that in Portugal (or in any country) all individuals with the same surname belong to the same family. As it is in most Western societies, while there is a small number of surnames, each being held by a large number of individuals, there are many surnames that are held by only a few individuals (Güell *et al.*, 2015, 2018). This is due to the birth-death process of surnames, whereby surnames are born when name mutation arises or via immigration and perish when its holder bears no children, which generates a commonality-ordered surname frequency distribution. This process also avoids a convergence of the surname distribution to a uniform distribution across surnames such that they would not provide any information. Considering high frequency surnames, individuals holding them are not likely to have

[41] According to our criterion, it is obvious that an individual will be Portuguese if his or her middle name was first considered as a second birth name when separating first names from surnames, because it belonged to the Portuguese names dictionary. The need for this criterion becomes more important when looking at the first word of a name, specifically in the cases for which an individual has a single first name and the remaining are surnames.

family linkages, while individuals with low frequency surnames are likely to share family relationships. This means that our baseline analysis should rely on rare surnames because they are the ones forming a partition of the population in which, probably, there are family linkages and are the ones from which information on intergenerational mobility can be extracted. This is only possible if we have a skewed surnames' distribution. The distribution of surnames and the surname-incidence Lorenz curve for our dataset are presented in Figures 3.1 and 3.2.

Figure 3.1 – Distribution of Surnames

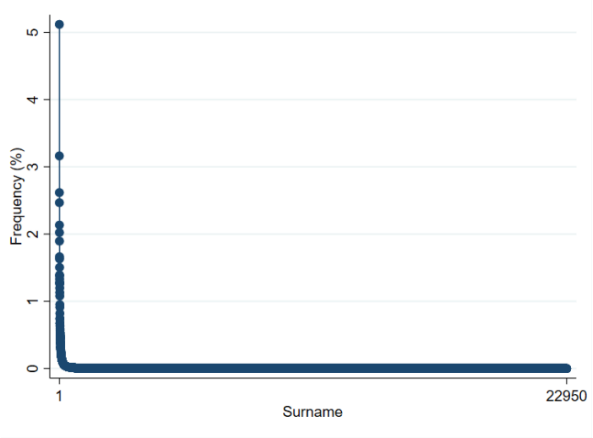
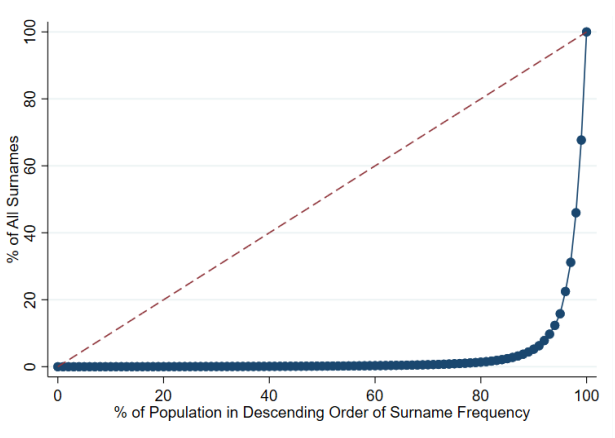


Figure 3.2 – Lorenz Curve for Surnames



Our final sample has 741,157 men and 22,950 surnames, with 32.29 as the average people *per* surname. As seen in Figure 3.1, the ordered frequency distribution is highly skewed, with few surnames being held by a large number of individuals and a large number of surnames held by few people. This is coherent with the assumption we need for our work. According to the Lorenz curve presented in Figure 3.2, about 1.4% of the most frequent surnames cover 80% of the population. The Gini index for the surnames' distribution is equal to 93.70, reflecting a high degree of asymmetry. We will therefore follow Güell *et al.* (2018) so that our estimation considers rare surnames, namely those held by fewer than 30 people: by dropping the most frequent surnames we increase the likelihood of the existence of family links and the ICS provides more accurate information. Our measure is defined as the ICS30. We then test how the measure is sensitive to changes in the number of surnames considered. Note that no ICS measure is defined for surnames associated with only a single person or if there is only a single surname for the entire population. This implies that when the informational content of surnames exists, there must be family linkages.

3.2.2.3.3. The First Name Distribution in Portugal

We are interested in comparing both the informational content of surnames and first names, meaning that our analysis will be performed for the same samples used to analyse the role of surnames in intergenerational transmission of socioeconomic status. The first name distribution for the entire population of 741,157 men may be seen in Figure 3.3 and its associated Lorenz curve in Figure 3.4.

Figure 3.3 – Distribution of First Names

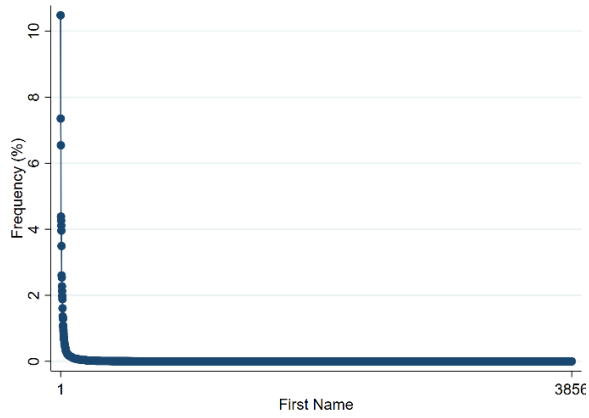
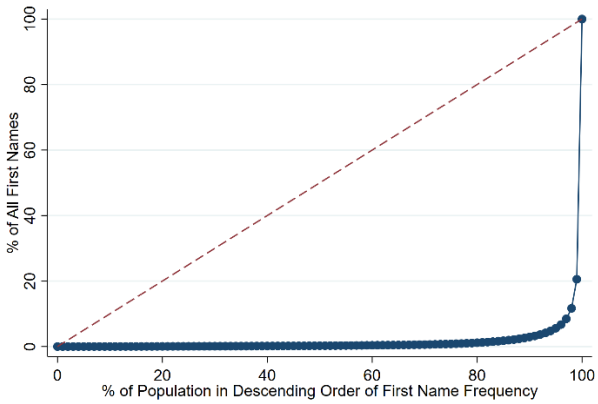


Figure 3.4 – Lorenz Curve for First Names



There are 3,856 different first names and the mean number of individuals *per* first name is 192.21, almost six times the average size for surnames. 80% of the population is covered by approximately 1.17% of the most frequent first names and the Gini index equals 96.98%. Therefore, the first name distribution is even more asymmetric than the surnames' distribution. While we required surnames to reflect family linkages and therefore imposed a restriction in the number of holders of each surname being equal to at most 30 people, first names will not have any restriction, and the ICF is also not defined, by construction, for first names with only a single holder.

3.2.2.4. Regional Correlates of Mobility

Following Güell *et al.* (2018), we gather data on regional macroeconomic variables, considering the Nomenclature of Territorial Units for Statistics 3 (NUTS 3) disaggregation (25 different regions). We divide the variables into three categories, which contain different subcategories: key outcomes, which cover the economic activity level, education, inequality, and social capital; other economic outcomes, which include information about the labour market and trade openness; and other socio-political outcomes, which are related to life expectancy, suicide rates, crime rates, and the public sector. Information is provided by INE, the Portuguese National Statistics Institute, and Pordata, from the *Fundação Francisco Manuel dos Santos*. We do not have a large time span of data for most variables, but we averaged them over the available years, making sure that the years used are the same for each subcategory. Since our ICS measures are for the year 2020, this is the last year considered in the analysis. Variables' information is presented in the Appendix C. Descriptive statistics for the averages computed over time are in Table C2, Table C3, and Table C4.

Looking at the most important variables, i.e., the key economic outcomes, we find that the retention rates on primary and lower secondary education as well as in the upper secondary education show clear regional disparities. Alto Minho has the lowest values for both rates: 3.91% and 12.19%, respectively. The highest are for Região Autónoma dos Açores, equal to 10.76%, and Algarve, equal to 20.33%,

respectively. Although no specific pattern occurs between the country's coast and interior[42] for this variable, the same is not true for real GDP *per capita* and inequality. Those measures tend to be lower in the Portuguese interior regions and higher on the country's coast. The lowest value for real GDP *per capita* occurs for the Tâmega e Sousa region, equal to 9,857.29€, and the highest for the Área Metropolitana de Lisboa, the country's capital, equal to 23,619.38€. Most inequality related measures present modest disparities between maximum and minimum, as opposed to retention rates and real GDP *per capita*, with the exception of the Gini Coefficients. The Área Metropolitana de Lisboa is again the one scoring high in the inequality measures in most measures. The exception occurs for the Douro region and the P90/P10 ratio for taxable persons, equal to 8.17. The P80/P20 and the P90/P10 ratios for taxable households and taxable persons and the Gini coefficient for taxable persons have their lowest values in the Ave region (2.92, 2.53, 6.47, 5.15, and 37.18, respectively), while for the Gini coefficient for taxable households this occurs for the Tâmega e Sousa region (equal to 41.42). Both island regions (Região Autónoma da Madeira and Região Autónoma dos Açores) are amongst the most unequal regions as well. Alto Tâmega, Douro, and Terras de Trás os Montes, all belonging to the country's interior, are exceptions in the group of less unequal regions for most variables, since they fit in the ranges of coast regions' values for inequality.

3.3. Empirical Results and Discussion for the Informational Content of Surnames

Here we analyse the regional disparities for this measure and whether migration influences the information contained in surnames. Additionally, we seek to determine if there are regional characteristics that may be connected with differences in the informational content of surnames in mobility across space.

3.3.1. Informational Content of Surnames

Table 3.3 presents the results for the ICS30 measure. This variable considers the informational content of surnames defined in equation (3.3) for surnames held by fewer than 30 individuals.

Table 3.3 – The Informational Content of Surnames for Surnames Held by Fewer Than 30 Individuals (ICS30)

Dep. Variable: Years of Education	
Real-surname regression R^2	30.66%***
Fake-surname regression mean R^2	17.11%
ICS30	13.55%
No. of Observations	58,252

(continues in the next page)

[42] We classify coast regions as those with genuine ocean shoreline as opposed to the interior ones.

Table 3.3 – The Informational Content of Surnames for Surnames Held by Fewer Than 30 Individuals (ICS30) *(continued)*

Dependent variable: years of education	
No. of Surnames	9,955

Notes: *** stands for surnames being jointly statistically significant at 1% level, respectively. Unique surnames are excluded from the estimations. Regressions with fake surnames are replicated 10 times and an average of their R^2 is computed, represented by the fake-surnames regression mean R^2.

Our evidence shows that when estimating equation (3.1), approximately 31% of the variability observed in the years of education is explained by individual surnames, which are not jointly equal to 0. When considering fake surname dummies to prevent any spurious information to be attributed to real surnames due to the large amount of regressors as formalized in equation (3.2), the adjusted R^2 falls to about 17%, with these independent variables not being jointly significant in each of the 10 rounds that we use to compute this regression. From this we conclude that the incremental information associated with surnames regarding our target variable is around 14%.

3.3.1.1. Rare Surnames

We argued above that rare surnames are those that possess the highest informational content of all surnames, because these are the ones for which, most likely, a family relationship occurs. We now present evidence supporting this assumption. We test what happens if we restrict our sample to having a higher degree of rarity in surnames (they are held by fewer people who are even likelier to be part of the same family) or if we consider the entire sample with all surnames (including those that are not likely to share family links although having the same surname). We also analyse a subsample in which we include only siblings, since we consider only individuals who have the same surnames in the same order, considering that they have more than one surname, i.e., we look to the subsample of surnames held by only two individuals. Table 3.4 presents the results.

Table 3.4 – Sensitivity of the ICS to Different Degrees of Surnames' Rarity

Dep. Variable: Years of Education	30	25	20	15	Siblings	All individuals
Real-surnames regression R^2	**30.66%*** **	32.24%***	34.53%***	37.58%***	53.83%***	3.44%***
Fake-surnames regression mean R^2	**17.11%**	18.19%	19.95%	22.36%	35.77%	1.54%
ICS	**13.55%**	**14.06%**	**14.59%**	**15.22%**	**18.06%**	**1.89%**
No. of Observations	**58,252**	53,997	48,231	41,019	12,191	729,487
No. of Surnames	**9,955**	9,802	9,549	9,146	4,352	11,208

Notes: *** stands for surnames being jointly statistically significant at 1% level. Unique surnames are excluded from the estimations. Regressions with fake surnames are replicated 10 times and an average of their R^2 is computed, represented by the fake-surnames regression mean R^2.

As we were expecting, the ICS increases with higher degree of rarity of surnames in our sample, a result confirmed by the estimations performed with the subsamples with surnames held by fewer than 25, 20, and 15 people, and decreases with the inclusion of more common surnames. When considering siblings (the most restrictive case), the ICS increases about 4 p.p. to 18% when compared to the ICS30. When all individuals are considered the measure falls 12 p.p. to approximately 2%. The informational content of surnames estimated for the entire sample is similar to the one computed by Güell *et al*. (2015) for Catalonia in Spain (close to 3%).

3.3.1.2. ICS30 Comparability Across Regions

The literature reports studies that have constructed intergenerational mobility measures across areas in order to inspect spatial patterns. We compute the informational content of surnames for each one of the 25 small regions defined in NUTS 3, according to the regions defined for each individual's company in the Orbis Database. We also consider surnames held by fewer than 30 individuals, making it comparable to the results for the entire sample. Surname distributions can be proxied by a Pareto distribution that is defined by the Gini coefficient and the average number of people *per* surname (Fox and Lasker, 1983), which we plot for each region in Figures 3.5 and 3.6.

Figure 3.5 – Comparison of Surname Distributions Across Regions (NUTS 3) Considering All Surnames

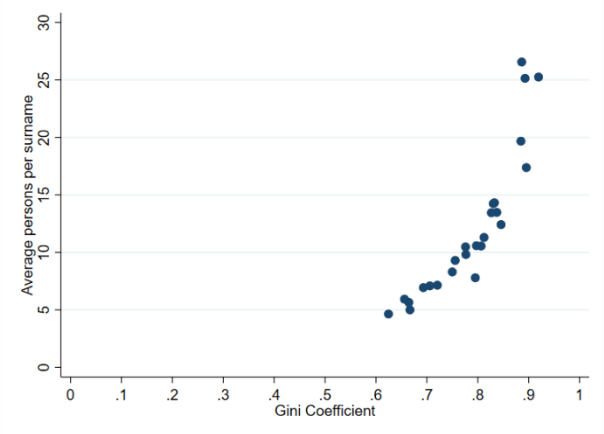
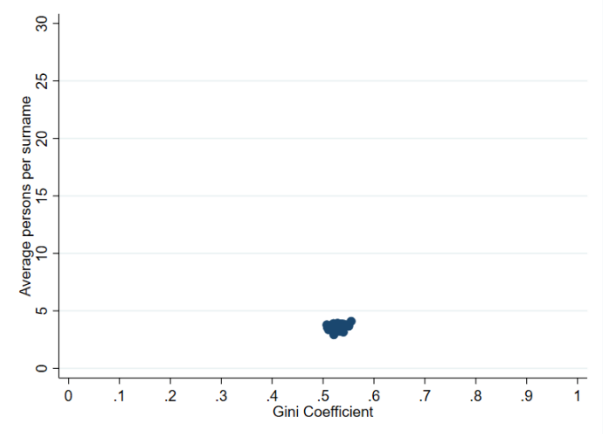


Figure 3.6 – Comparison of Surname Distributions Across Regions (NUTS 3) Considering Rare Surnames (< 30 holders)



The dots in Figure 3.5 do not coincide when common surnames are included, meaning that distributions across regions are not identical, unlike what occurs in the case for which only rare surnames are considered: in Figure 3.6, dots overlap, and distributions are identical.

Since we are interested in comparing regions using the surnames (families) from each region, we have to account for internal migration. This is because migrants may have different outcomes when compared to natives, promoting a bias in the estimates of the ICS30. In order to check whether migration plays a role, we follow Güell *et al*. (2018). In our data we do not have information on the individual's region of origin. The authors define an index for the regional dimension of surnames as follows:

$$Regional\ Index\ ICS30\ (s,r) = \frac{No.of\ Individuals\ with\ surname\ s\ in\ region\ r}{No.of\ Individuals\ with\ surname\ s\ in\ Portugal} \times 100. \quad (3.4)$$

The larger the index, the less likely it is that we are dealing with migrants. Sufficiently regional surnames are considered to be those forming the partition of the 50% most regional surnames (Regional ICS30) and therefore are the ones that are likely to not belong to migrants. As pointed out by Güell *et al*. (2018), this will not solve the migration related problems because we cannot account for the individuals who move away from their regions of origin. Therefore, our analysis addresses this issue only partially.

Results for the ICS30 and the Regional ICS30 are presented in Table 3.5 and graphically represented in Figures 3.7 and 3.8, respectively. Absolute differences between the two measures are also reported in Table 3.5.

Table 3.5 – ICS30 and Regional ICS30 by NUTS 3

NUTS 3	ICS30 (%)	Regional ICS30 (%)	| ICS30 - Regional ICS30 | (p.p.)
Alto Minho	10.79	10.2	0.59
Cávado	10.85	17.77	6.92
Ave	12.97	21.55	8.58
Área Metropolitana do Porto	12.86	17.68	4.82
Alto Tâmega	9.36	17.86	8.5
Tâmega e Sousa	10.85	26.86	16.01
Douro	15.47	22.66	7.19
Terras de Trás-os-Montes	8.2	13.25	5.05
Oeste	8.97	14.7	5.73
Região de Aveiro	9.01	14.2	5.19
Região de Coimbra	11.9	17.58	5.68
Região de Leiria	11.45	14.02	2.57
Viseu Dão Lafões	9.92	17.13	7.21
Beira Baixa	12.38	23.64	11.26
Médio Tejo	11.93	19.44	7.51
Beiras e Serra da Estrela	9.75	19.26	9.51
Área Metropolitana de Lisboa	12.83	16.12	3.29
Alentejo Litoral	10.65	17.31	6.66
Baixo Alentejo	16.08	25.9	9.82
Lezíria do Tejo	12.97	20.06	7.09
Alto Alentejo	17.54	22.86	5.32
Alentejo Central	16.05	23.39	7.34
Algarve	11.36	15.53	4.17
Região Autónoma dos Açores	9.93	18.33	8.4
Região Autónoma da Madeira	11.49	13.07	1.58

Figure 3.7 – ICS30 by Region (NUTS 3)

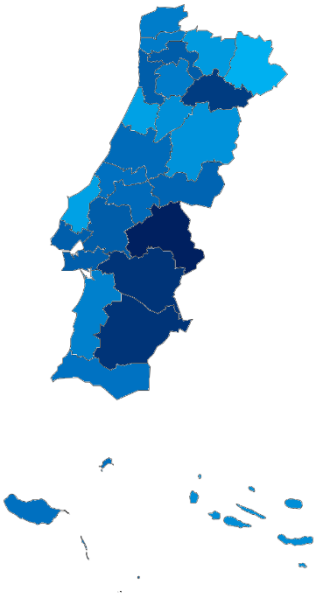
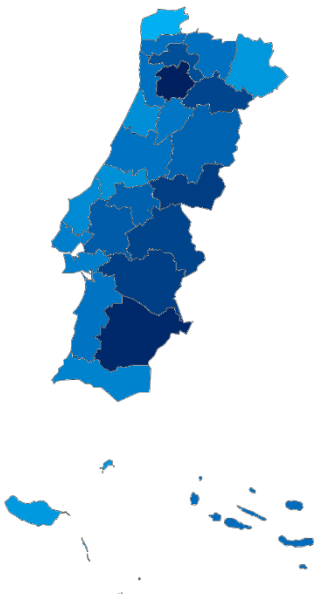


Figure 3.8 – Regional ICS30 by Region (NUTS 3)



Notes: Darker blue implies a higher ICS30/Regional ICS30, i.e., lower education mobility.

Overall, the country's interior presents a higher informational content of surnames than the coast regions. The ICS30 is on average around 12.57% in the interior regions while this value drops to 11.07% when considering the coast. These values for the Regional ICS30 are approximately 21.07% and 15.51%, respectively. The variability of the ICS30 and Regional ICS30 is given by their standard deviations, being around 2.35% and 4.12%. Note that although the general results are maintained, there are clear differences when comparing the ICS30 and the Regional ICS30, which indicates that internal migration probably promotes a bias in the first measure (last column of Table 3.5). This is confirmed by the smooth positive Pearson correlation between both, equal to 61.71%, and appears to be more pronounced in the Beira Baixa and the Tâmega e Sousa regions, with a negligible effect in the Alto Minho and the Região Autónoma da Madeira. With the exception of Terras de Trás os Montes and Viseu Dão Lafões, Regional ICS30 is higher in all of the interior regions. This result is in line with the findings of Alesina *et al*. (2021), who conclude that mobility presents a negative correlation with the distance to the coast, when considering African countries.

3.3.1.3. Correlating Mobility and Socioeconomic Outcomes

Güell *et al*. (2018) study the relationship between their estimates of regional ICS and socioeconomic and political outcomes. We consider the regional measure of mobility in this framework, as we noticed that migration appears to play a role in the spatial dispersion of persistence. We pool the data for each region $r = 1, \dots, R$, with $R = 25$, and perform several regressions between the $Regional\ ICS30_r$ and

each covariate, x_r (described in Subsection 3.2.2.4.) at a time, controlling for real GDP *per capita*, $C_{1,r}$ as well as for migration flows, $C_{2,r}$ (interchangeably and together). This can be formalized as:

$$Regional\ ICS30_r = \beta_0 + \beta_1 x_r + \beta_2 C_{1,r} + \beta_3 C_{2,r} + \omega_r \tag{3.5}$$

where ω_r is the error term.

Our primary goal is to understand which regional characteristics may be associated with differences in mobility across regions, as this may enlighten future research about mobility determinants in Portugal and goes beyond finding causation. For the sake of simplicity, results for the variables, which are statistically significant in at least one specification, are presented in Table 3.6, where the first column reports the cases in which no controls are used, the second column the scenario in which we control for real GDP *per capita*, and the third column considers the case using migration flows as a control variable, as in Güell *et al*. (2018). The last column considers both GDP *per capita* and migration flows as controls.

Table 3.6 – The Relationship Between Regional ICS30 and Regional Outcomes

Variables	(1)	(2)	(3)	(4)
Key Economic Outcomes				
Education | (2009-2020)				
Retention and drop-out rate in primary and lower secondary education (%)	0.59 (0.51)	0.99* (0.52)	0.85* (0.49)	1.04* (0.51)
Inequality | (2015-2020)				
P80/P20 ratio for TH	-7.61** (2.80)	-7.42** (3.28)	-8.14** (3.77)	-7.99* (3.93)
P80/P20 ratio for TP	-7.50** (3.05)	-6.85** (3.29)	-6.82* (3.59)	-6.69* (3.67)
P90/P10 ratio for TH	-1.88** (0.87)	-1.68* (0.98)	-1.64 (1.06)	-1.58 (1.09)
P90/P10 ratio for TP	-1.77** (0.84)	-1.73** (0.83)	-1.52* (0.89)	-1.66* (0.90)
Gini coefficient for TH (%)	-0.93** (0.42)	-0.87* (0.50)	-0.85 (0.54)	-0.83 (0.58)
Gini coefficient for TP (%)	-0.83** (0.38)	-0.75* (0.40)	-0.71 (0.44)	-0.71 (0.45)
Other Economic Outcomes				
Trade openness | (2009-2020)				
Imports/GDP	-12.64* (6.30)	11.14 (6.59)	-10.15 (7.04)	-9.90 (7.17)
Other Socio-Political Outcomes				

Crime Rates (*per* 100,000 residents) | (2015-2020)				
Crimes against cultural identity	-2.47*	-2.37	-2.29	-2.35
	(1.32)	(1.37)	(1.53)	(1.58)

(continues in the next page)

Table 3.6 – The Relationship Between Regional ICS30 and Regional Outcomes *(continued)*

Variables	(1)	(2)	(3)	(4)
Crimes against pets	0.13	0.26*	0.14	0.24*
	(0.13)	(0.13)	(0.12)	(0.14)
Burglary in residence	-0.03*	-0.03	-0.03*	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)
Driving a motor vehicle with a blood alcohol level equal to or above 1.2g/l	-0.03*	-0.03	-0.04*	-0.05**
	(0.02)	(0.02)	(0.02)	(0.02)
Controls				
Real GDP *per capita*	No	Yes	No	Yes
Migration Flows	No	No	Yes	Yes

Notes: Standard errors are presented in parentheses. * and ** stand for statistically significant at 10% and 5% levels, respectively. TH and TP stand for taxable household and taxable person, respectively. Column (1) regards the cases in which no controls are used, column (2) column the scenario in which we control for real GDP *per capita*, column (3) considers the case using migration flows as a control variable, and column (4) considers both GDP *per capita* and migration flows as controls.

Considering the variables that are statistically significant in at least one specification, results show that economic activity, social capital, labour market outcomes, life expectancy, and suicide rates appear not to be associated with differences in mobility. A higher retention and drop-out rate in primary and lower secondary education is associated with more persistence, when controlling for real GDP *per capita* or migration flows. The same was shown by Narayan *et al*. (2018) according to which economies with lower out-of-school rates have higher mobility in education, since opportunities become more equal for children of rich and poor parents. This resembles the results of Daude and Robano (2015) and Hilger (2016), who find for Latin America and the USA, respectively, that enrolment rates on pre-school level and high school enrolment influence intergenerational mobility in education in a positive way. Imports as a share of GDP are significant only when not controlling for other factors, having a positive relationship with mobility. There is no clear pattern for the relationship between Regional ICS30 and other socio-political outcomes, where statistical significance is verified for crimes against cultural identity, pets, burglaries in residences, and driving a motor vehicle with a blood alcohol level equal to or above 1.2g/l, but with different effects. This is also verified by Güell *et al*. (2018). As they point out, the existence of weak or mixed evidence may be a consequence of the connection between mobility and these variables going beyond the interactions performed in terms of complexity and unpredictability.

Interesting evidence emerges when considering the variables proxying for inequality, which are statistically significant across specifications, but with a negative relationship with intergenerational persistence. The connection between educational persistence and income inequality is usually described in the literature as a two-way relationship. First, inequality in the parental generation may restrict

mobility in education of children. Since we average inequality around the time when education mobility is measured, i.e., inequality concerns the youngest's generation, we abstract from this first direction[43]. Second, the lack of educational mobility may contribute to greater income inequality in the same generation. If the relative positions of children in the education distribution resembles those of their parents and the parents' generation is characterized by an unequal income distribution, children from poor families will continue to have lower labour market returns, while the opposite occurs for children from rich households. This promotes the disparities in the relative wages of low-skilled workers in comparison with high-skilled workers, i.e., inequality is maintained. Also, a low education mobility environment is associated with individuals who have more innate abilities being less inclined to obtain more education and productive jobs (Owen and Weil, 1998), and for that reason there may be efficiency losses that harm growth (Galor and Tsiddon, 1997; Hassler and Mora, 2000). Low mobility may influence in a negative way the perceptions that individuals have about fairness, which in turn damages social stability and growth (Narayan *et al*., 2018). Less inclusiveness and lower economic growth lead to the strengthening of capital market imperfections, as credit becomes more limited to poorer and lower-skilled workers (Maoz and Moav, 1999; Owen and Weil, 1998) and also to fewer public resources devoted to equalizing the income distribution, namely through education. Hence, we would expect the relationship between persistence in education and inequality to be positive. This occurs in the works of Hilger (2016) for the USA, considering the 1940-2000 period, Lee and Lee (2020) regarding 30 OECD countries and the generations born in 1947-1990.

Our evidence suggests that in places where persistence in education is lower, inequality is higher. Different explanations may exist for this to occur. The first regards the labour market. According to arguments presented by Narayan *et al*. (2018), regions where the unemployment is high should present a mismatch between the demand and supply of workers, making workers accept suboptimal wage offers that reflect the decrease in reservation wages. Besides, when unemployment is high, human capital investments are less easily monetized. This is particularly important if we consider a recession period, as we do, because the Covid-19 pandemic crisis is present on our inequality measures (the yearly average values for our unemployment measures increased from 2019 to 2020). Considering the NUTS 3 division, for the P90/P10 and P80/P20 ratios and the Gini coefficient (for taxable persons), there are positive and statistically significant Pearson correlations with our unemployment measures (see Table C5 in the Appendix C). Additionally, it suffices that the jobs' distribution is grounded on social connections or specific individual characteristics. Individuals may be less dependent on their parents but will not have

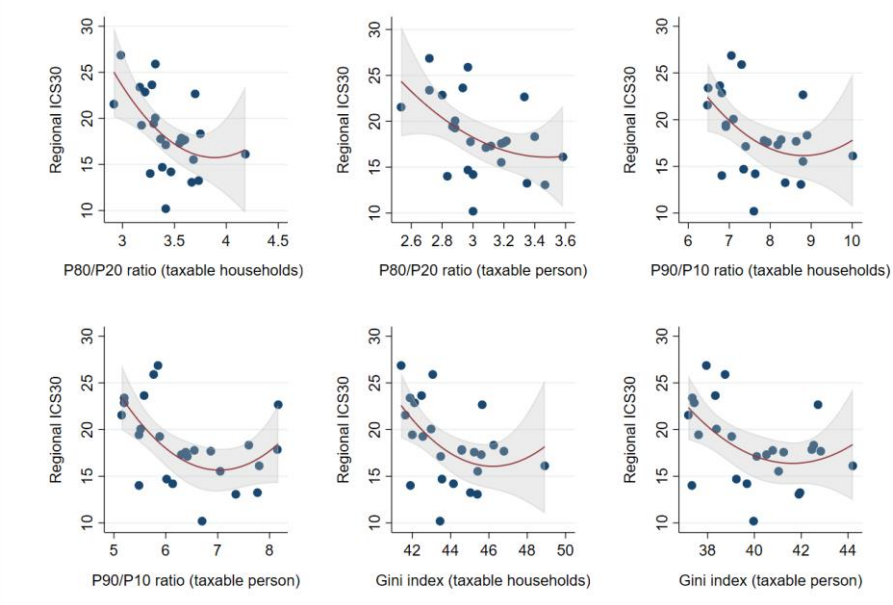
[43] Inequality raises the disparities in investments made by poor versus rich parents (Becker and Tomes, 1979, 1986; Becker *et al*., 2018; Burtless and Jencks, 2003; Andrews and Leigh, 2009; Neidhöfer, 2019; Duncan and Murnane, 2011; Özalp, 2019), it influences children's health, early aptitudes, social and cognitive progress, schools' quality (Corak, 2013; Knudsen *et al*., 2006), distorts incentives and opportunities, which make poorer families' hard-working and talented individuals not receive the deserved schooling (Causa and Johansson, 2010; Lee and Lee, 2020), makes individuals from disadvantaged families disbelieve that they may rise on the economic ladder through education (Leone, 2022; Kearney and Levine, 2016), and promotes the severity of credit constraints through credit market imperfections (Becker and Tomes, 1986), contributing to persistence.

a reward in the labour market. These may be signalling a weak labour market, which harms an inclusive and efficiently based growth, breaking the expected connection between mobility and inequality.

The second explanation is related to perceptions and aspirations. Narayan *et al*. (2018) also point out that when mobility is perceived by individuals as being higher, it can be translated into more tolerance for inequality, if this is associated with merit (Fleib, 2015). This argument supports behavioural works showing that if inequality is considered to be fair, people will not be averse to it (Fehr and Fischbacher, 2003; Cappelen *et al*., 2013). Therefore, income inequality is accepted as an inevitable result of the effort a society has to make in order for individuals and their children to move up on the educational ladder, which should promote higher prosperity in the long-run. When mobility is higher and there are expectations that mobility will also be higher in the future, and people will accept the *status quo*, while the opposite occurs for immobile regions (this is the designated tunnel effect as in Hirschman and Rothschild, 1973). We consider that this second possible reason may in fact reinforce the first one because if inequality exists due to a weak labour market, it may be maintained, since individuals find it fair and do not require governmental intervention to reduce it.

In a final exercise, we check whether nonlinearities exist regarding this relationship. This is grounded on existing evidence about the relationship between inequality and income mobility. As Kourtellos (2021) point out, linearity may proxy the relationship of interest around a particular segment of data, but not globally. Kourtellos (2021) finds, in the context of the relationship between education mobility and education inequality, that these two variables present a linear relationship with a kink. Figure 3.9 represents the relationship between the Regional ICS30 and inequality measures, in which we also adjusted a non-linear (quadratic) fit.

Figure 3.9 – Relationship Between Regional ICS30 and Inequality with Quadratic Prediction



Notes: the red curve represents the non-linear fit prediction for Regional ICS30, from a regression of Regional ICS30 on an inequality measure; the shaded (grey) area represents a 95% confidence interval.

The graphical representations suggest that there may be turning points in the negative relationship after which it becomes positive. We therefore regress the Regional ICS30 on inequality measures using linear and quadratic terms, interchangeably. Table 3.7 presents the results.

Table 3.7 – The Relationship Between Regional IC30 and Inequality Accounting for Nonlinearities

Variables	**(1)**		**(2)**		**(3)**		**(4)**	
Inequality | (2015-2020)	**Linear**	**Quadratic**	**Linear**	**Quadratic**	**Linear**	**Quadratic**	**Linear**	**Quadratic**
P80/P20 ratio for TH	-76.53	9.84	-79.77	10.40	-105.25*	14.32*	-104.76*	14.27
	(45.36)	(6.47)	(46.86)	(6.72)	(55.86)	(8.22)	(57.31)	(8.43)
P80/P20 ratio for TP	-63.24	9.06	-68.66	10.08	-76.49	11.43	-75.81	11.34
	(61.84)	(10.05)	(62.99)	(10.26)	(65.36)	(10.71)	(66.68)	(10.93)
P90/P10 ratio for TH	-20.09	1.14	-24.82*	1.46	-39.81**	2.46**	-40.42**	2.50**
	(13.41)	(0.54)	(14.12)	(0.89)	(16.46)	(1.06)	(16.77)	(1.08)
P90/P10 ratio for TP	-30.83**	2.19**	-28.19**	1.99**	-30.22**	2.16**	-29.10**	2.08**
	(12.18)	(0.92)	(12.71)	(0.96)	(12.23)	(0.92)	(13.03)	(0.98)
Gini coefficient for TH (%)	-26.27	0.28	-30.12*	0.33*	-56.95**	0.64**	-57.44**	0.64**
	(15.83)	(0.18)	(16.69)	(0.19)	(20.56)	(0.23)	(21.04)	(0.24)
Gini coefficient for TP (%)	-25.21	0.30	-28.49*	0.34	-42.10**	0.52**	-41.14**	0.51**
	(16.22)	(0.20)	(16.33)	(0.20)	(18.24)	(0.23)	(18.84)	(0.24)
Controls								
Real GDP *per capita*	No		Yes		No		Yes	
Migration Flows	No		No		Yes		Yes	

Notes: Standard errors are presented in parentheses. * and ** stand for statistically significant at 10% and 5% levels, respectively. TH and TP stand for taxable household and taxable person, respectively. Column (1) regards the cases in which no controls are used, column (2) column the scenario in which we control for real GDP *per capita*, column (3) considers the case using migration flows as a control variable, and column (4) considers both GDP *per capita* and migration flows as controls.

Overall, results are sensitive to the addition of a squared term. The linear and non-linear parts of inequality are both significant for all measures when controlling for migration only (column (3)), except for the P80/P20 ratio for taxable persons, and when controlling for real GDP *per capita* and migration (column (4)), except for the P80/P20 ratio. Grounded on the graphical representations in Figure 3.9, we argue that outliers may be somehow driving these results. Therefore, through the use of box plots we check whether outliers exist in our data. An observation is considered an outlier if it is beyond the lower or upper adjacent values: these respectively correspond to the lower quartile minus 3/2 the interquartile range and the upper quartile plus 3/2 the interquartile range. Boxplots are represented in Figures C1 and C2 in the Appendix C.

We find only one outlier for the P80/P20 ratio for taxable households, which regards the Área Metropolitana de Lisboa (around 4.183). We winsorize it, i.e., change its value to the nearest observation that is not an outlier, i.e., for the Região Autónoma dos Açores (equal to 3.75) and perform the same regressions as before for this measure: the one with only a linear term and another adding a squared term. Results are presented in Table 3.8.

Table 3.8 – The Relationship Between Regional IC30 and the P80/P20 Ratio (TH) After Winsorization

Variables	(1)		(2)		(3)		(4)	
P80/P20 ratio (TH) | (2015-2020)	**Linear**	**Quadratic**	**Linear**	**Quadratic**	**Linear**	**Quadratic**	**Linear**	**Quadratic**
- Accounting for Non-linearities	-99.89 (83.52)	13.34 (12.33)	-98.53 (85.37)	13.20 (12.59)	-100.58 (85.08)	13.55 (12.56)	-99.92 (87.34)	13.46 (12.89)
- Not Accounting for Non-linearities	-9.54*** (3.17)	NA	-9.07** (3.49)	NA	-8.92** (3.56)	NA	-8.80** (3.70)	NA
Controls								
Real GDP *per capita*	No		Yes		No		Yes	
Migration Flows	No		No		Yes		Yes	

Notes: Standard errors are presented in parentheses. ** and *** stand for statistically significant at 5% and 1% levels, respectively. TH stands for taxable household. NA stands for Not Applicable.

Our evidence confirms that the significance of the squared term for the P80/P20 ratio is due to the presence of an outlier. They also show that, even after winsorizing the outlier, when including only a linear term in the regressions performed, it continues to be statistically significant, as it was in our initial analysis of correlates[44]. Overall, one may conclude that the relationship between persistence and inequality is linearly negative considering the P80/P20 ratio. It is non-linear for the P90/P10 ratio and the Gini Index, being negative for lower inequality levels, while positive for higher inequality levels.[45]

Summing up, the relationship between the two variables is more complex, and deserves further research that should consider the possible mechanisms we raised.

3.4. Empirical Results and Discussion for the Informational Content of First Names

We now compute the informational content of first names for all the samples considered in the surnames part of the analysis. Table 3.9 presents the results for the ICF measure. We also analyse the regional disparities for this measure.

3.4.1. Informational Content of First Names

Table 3.9 – ICF on Samples with Different Degrees of Surnames' Rarity

Dep. Variable: Years of Education	30	25	20	15	Siblings	All individuals
Real-first name regression R^2	**4.16%*** **	4.34%***	4.66%***	4.96%***	6.57%***	1.68%***

(continues in the next page)

[44] The same conclusions are drawn if we eliminate the outlier.

[45] Table C6 presents the predicted minimum values for the corresponding nonlinear relationships found in Table 3.7 (excluding the P80/P20 ratio). The regions that would always be above the minimum found for the P90/P10 ratio are Alentejo Litoral, Alto Tâmega, Terras de Trás-os-Montes, Área Metropolitana do Porto, Região Autónoma da Madeira, Douro, Algarve, Região Autónoma dos Açores, and Área Metropolitana de Lisboa.

Table 3.9 – ICF on Samples with Different Degrees of Surnames' Rarity *(continued)*

Dep. Variable: Years of Education	30	25	20	15	Siblings	All individuals
Fake-first name regression mean R^2	**1.77%**	1.82%	1.96%	2.21%	3.94%	0.28%
ICF	**2.38%**	**2.52%**	**2.70%**	**2.74%**	**2.63%**	**1.40%**
No. of Observations	**57,698**	53,449	47,697	40,485	12,123	728,067
No. of First Names	**1,008**	984	955	905	488	2,082

Notes: *** stands for surnames being jointly statistically significant at 1% level. Unique first names, besides unique surnames, are excluded from the estimations – this is the reason why the number of observations presented here and in Table 3.4. Regressions with fake surnames are replicated 10 times and an average of their R^2 is computed, represented by the fake-surnames regression mean R^2.

Results show that when regressing educational attainment against real first names, the latter are jointly significant and explain almost 4.2% of the first variable's variance (second column of Table 3.9). Doing the same but for fake first name dummies, R^2 is reduced to around 1.8% and these regressors are jointly not significant in all the rounds for which this regression is computed. Therefore, first names appear to have a limited role in an individual's socioeconomic status, because the parents' decision is only responsible for approximately 2.4% of the differences in educational attainment across individuals. The greater the likelihood of family links existing, i.e., the rarer surnames become, the more informational content first names provide, although the differences are not that pronounced. The exception occurs for the siblings' subsample which, for the case of surnames presented the highest ICS, but for the case of first names the ICF is in the middle of the other subsamples' values. When analysing the entire sample, in which the likelihood of individuals with the same surname being part of the same family decreases, the ICF is close to the ICS.

Overall, first names appear to explain little of the variance of educational attainment, around 2%, i.e., belonging to a specific family is more crucial than parents' choice about first names in determining individuals' socioeconomic status.

3.4.1.1. ICF Comparability Across Regions

We now turn our attention to how the informational content of first names varies across the country. We estimate the ICF for the 25 NUTS 3 regions and consider the surnames with fewer than 30 individuals for the sake of comparability with the ICS. Restricting the sample also makes the first names distributions throughout the Portuguese territory comparable, as it was for the surnames case. This can be verified through the comparison of the first names distributions without imposing any restrictions on the number of surnames holders and considering rare surnames in each region.

Figure 3.10 – Comparison of First Names Distributions Across Regions (NUTS 3) Considering All Surnames

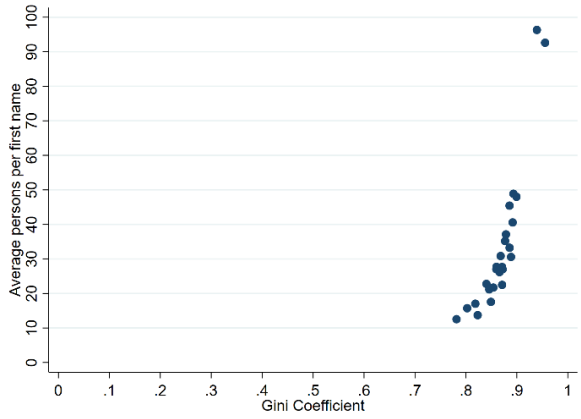
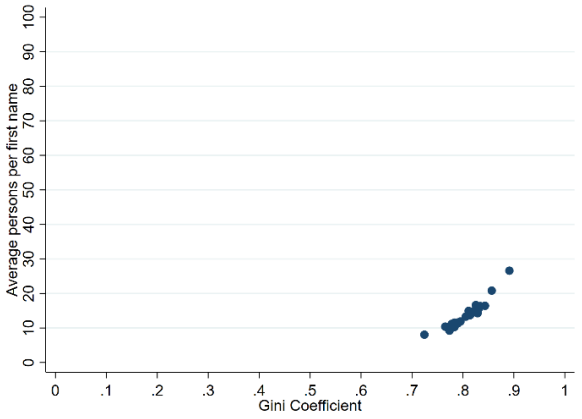


Figure 3.11 – Comparison of First Names Distributions Across Regions (NUTS 3) Considering Rare Surnames (< 30 holders)



In Figure 3.10, the dots are farther apart from each other than those in Figure 3.11, meaning that first name distributions are more comparable in the second case.

Results for the ICF considering the sample of rare surnames and the sample of individuals with the 50% of most regional surnames, as well as absolute differences are presented in Table 3.10 and graphically represented in Figures 3.12 and 3.13, respectively.

Table 3.10 – ICF for the Subsamples of Surnames Held by Fewer Than 30 People (A) and the Subsample of Most Regional Surnames (B) by NUTS 3

NUTS 3	A	B	| A – B |
Alto Minho	1.36	6.33	4.96
Cávado	3.31	3.20	0.12
Ave	0.83	1.09	0.26
Área Metropolitana do Porto	2.53	3.67	1.13
Alto Tâmega	0.81	6.78	5.96
Tâmega e Sousa	1.57	-2.12	3.69
Douro	2.91	4.63	1.71
Terras de Trás-os-Montes	0.35	5.32	4.97
Oeste	0.94	3.12	2.18
Região de Aveiro	2.46	3.40	0.94
Região de Coimbra	1.69	4.40	2.71
Região de Leiria	1.41	0.57	0.84
Viseu Dão Lafões	2.29	3.95	1.66
Beira Baixa	3.08	1.50	1.58
Médio Tejo	2.34	7.33	4.99
Beiras e Serra da Estrela	1.75	5.39	3.63
Área Metropolitana de Lisboa	2.40	2.92	0.52
Alentejo Litoral	3.64	3.78	0.14
Baixo Alentejo	3.26	4.05	0.79
Lezíria do Tejo	1.81	1.10	0.71

(continues in the next page)

Table 3.10 – ICF for the Subsamples of Surnames Held by Fewer Than 30 People (A) and the Subsample of Most Regional Surnames (B) by NUTS 3 *(continued)*

NUTS 3	A	B	| A – B |
Alto Alentejo	3.00	4.13	1.13
Alentejo Central	1.74	2.32	0.58
Algarve	2.88	5.38	2.49
Região Autónoma dos Açores	2.02	3.16	1.14
Região Autónoma da Madeira	4.33	5.04	0.72

Figure 3.12 – ICF for the Subsamples of Surnames Held by Fewer Than 30 People by Region (NUTS 3)

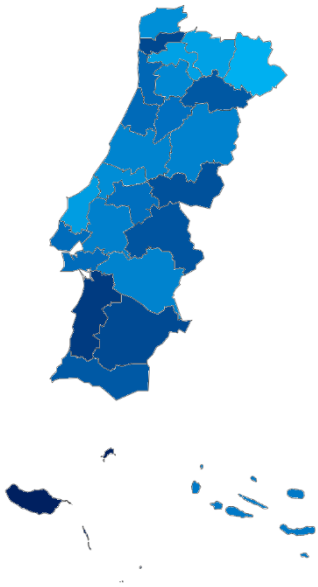
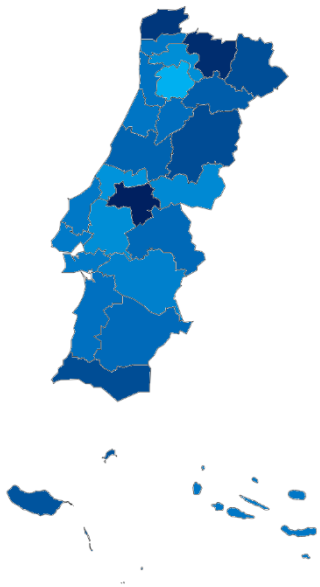


Figure 3.13 – ICF for the Subsample of Most Regional Surnames (NUTS 3)



Notes: Darker blue implies a higher ICF.

The evidence presented above shows that differences between the two subsamples exist although they are negligible (last column of Table 3.10), with only one exception exceeding a 5 p.p. difference: this occurred for most regions when comparing the ICS30 and Regional ICS30 by NUTS 3. However, the correlation between their values is 9.87%, much lower than the one verified for the surnames' analysis. Overall, the country's regional disparities are not very significant in comparison with the case for surnames, confirmed by a standard deviation of the ICF across regions of about 1% for the subsample of rare surnames and 2% in the subsample of most regional surnames: these correspond to less than half of those verified for the ICS30 and Regional ICS30 which are, respectively, around 2.4% and 4.1%. We consider that it does not make sense to analyse differences in socioeconomic outcomes that may be related with differences in the ICF, because these are too small. Nonetheless, the ICF computed for the ICS30 subsample is on average around 1.98% in the inner regions while this value increases to 2.26% when considering the coast. These values for the Regional ICS30 subsample are approximately 3.50%

and 3.68%. This means that now first names seem to matter more in the places near the ocean than in the interior.

3.5. Concluding Remarks

In this work we analysed education mobility for Portugal, applying an indicator designated the informational content of surnames (ICS) developed by Güell *et al*. (2015). This methodology has the advantage of not needing a panel data or intergenerational links between individuals. It assumes that surnames, when rare, are indicative of family links among individuals in the sample: we follow Güell *et al*. (2018) and analyse surnames belonging to fewer than 30 individuals, i.e., we compute the ICS30. Grounded on literature we also consider that first names may play a role in individuals' socioeconomic status, as they result from the choice of parents conditional on their socioeconomic status, preferences, cultural, and racial factors. Therefore, we perform an exercise analogous to the ICS measure, which we defined as the informational content of first names (ICF). We analyse it in the subsamples considered for surnames to compare if what matters is the family into which one is born or the choice parents make regarding first names. Additionally, for the first time we are able to perform a spatial analysis of education mobility across 25 Portuguese regions. We correlate the informational content of surnames with a set of socioeconomic and political variables for each region. This is done only for surnames because the variability of the information contained in surnames across regions is twice that verified for first names.

Results show that rare surnames account for 14% of the variability observed regarding educational attainment, when using the ICS30. When considering the standard ICS30 and one for the 50% most regional surnames (to account for the effect of internal migration), Regional ICS30, we conclude that regions on the country's coast present more education mobility than regions in the interior, as found for African countries. Also, first names explain only approximately 2% of the differences in educational attainment. Unlike what occurs with surnames, the informational content of first names is higher along the country's coast than in its interior. All of this together leads to the conclusion that a family into which an individual is born matters more for educational outcomes than the parents' choice about their offspring's first names.

The informational content of surnames appears to have no link with economic activity, social capital, labour market outcomes, life expectancy, and suicide rates, when analysing differences across regions. The higher the retention and dropout rates in primary and lower secondary education, the more information is provided by surnames, i.e., the greater is intergenerational persistence, as it is for Latin American countries and the USA, respectively. The weight of imports in GDP presents the opposite (and weak) relationship. The higher is inequality measured by the P80/P20 ratio, the lower is the informational content of surnames, although this is unexpected. There is a convex non-linear relationship between persistence and the P90/P10 ratio and the Gini Index.

We did not obtain strong and clear results regarding most of the regional correlates of persistence measured by the Regional ICS30, but that does not mean that they do not have a role in this context. It may be that their relationship involves more complexity than the simple exercises we perform. As in Güell *et al*. (2018) the correlations found could be the result of different political and institutional frameworks, but this might not be the case for Portugal, since it is a centralized country, so the institutional and political context is the same for all the regions.

We do not seek to find causal mechanisms for the unexpected joint behaviour between inequality and education mobility, but we address some of the possible factors that may be associated with it, grounded on the literature on the topic. First, when unemployment is high, wages accepted are suboptimal and it is difficult to monetize education. Second, if jobs are allocated based on social connections, mobility in education can occur but will not be rewarded in the labour market. Third, when education mobility and inequality are associated with merit, people will more easily accept the inequality situation. Fourth, individuals may expect high mobility to be verified in the future as well, which makes them accept their current status.

Although the institutional framework should not be the cause of the evidence we present, Narayan *et al*. (2018) consider that governments that care about social stability and growth should act to avoid individuals accepting inequality as a consequence of merit. Providing information and exposing individuals to different realities should encourage them to reject a bad *status quo*, promoting a positive feedback on mobility. Individuals should demand inequality reductions from policy makers, which may in turn enhance mobility. To strengthen the labour market when unemployment exists, the access to a social protection system and unemployment benefits, tax reductions, subsidized training, and wage subsidies may also be beneficial (Katz, 1998; Coenjaerts *et al*., 2009; Betcherman *et al*., 2010; Groh *et al*., 2016; Chen *et al*., 2018). The OECD (2012) complements the policies to tackle inequality: reducing the differences between temporary and permanent work protection, promoting active labour market policies with increased spending, strengthening the participation of women in the economic life and their work related outcomes, and, through the tax system, promoting efficient and sustainable redistribution are among the suggestions. Decreasing the school dropout rates in places where these are high, improving childhood development, investing public monies in education, facilitating the access to quality education, promoting safety, decreasing segregation, strengthening infrastructures and institutions, and easing credit to finance education may be additional mobility enhancing actions to be undertaken (Brown and Richman, 1997; Chetty *et al*., 2014a, 2016; Kline and Moretti, 2014; Chetty and Hendren, 2018a, 2018b).

A good avenue for future research would be to perform this same analysis longitudinally.

Conclusions

Our work focused on intergenerational mobility of income and education. We started with an analysis of the potential main determinants of intergenerational mobility for income and education. In the first chapter of this work we used the Rigorous Least Absolute Shrinkage and Selection Operator (RLASSO) and the Random Forest and Gradient Boosting algorithms to assess what determines intergenerational mobility in income and education, for a sample of 137 developing and developed countries considering the period from 1960 to 2018. These methods make us avoid the consequences of an *ad-hoc* model selection, given that our dataset had a large number of possible determinants, which were grounded on a comprehensive literature review. The two algorithms do not allow us to directly obtain the direction of the relationship between the variables, therefore we had to obtain them through the use of Shapley values. Finally, given that not all the countries with education mobility values in our dataset present intergenerational mobility in income values, we were able to predict missing mobility in income observations.

Results show that intergenerational income mobility presents a positive relationship with the share of married individuals. The growth rate of population density appears to negatively influence intergenerational mobility in income. The same occurs for inequality, resembling the well known Great Gatsby curve, and the share of individuals with less than primary education. Even though policies are contingent on each country's context, a better access to capital markets may mitigate the effect of inequality in mobility, because poor individuals would be less constrained and more able to finance the education of their children. Additionally, human capital investment should be promoted to make low-income individuals benefit from skills (both cognitive and non-cognitive) that have returns in the labour market.

A positive relationship was found between educational mobility and the stock of migrants. This also occurred when considering adult literacy and government expenditures on primary education. Not only this reinforces the need for human capital investment, but also highlights the positive effect that public spending on early childhood has on individuals' future. This should mitigate the difference in private investments between poor and rich parents.

We found that developing countries were penalized in both mobility dimensions, presenting larger intergenerational persistence. This result is corroborated by the lower mobility verified in countries belonging to the Latin America and Caribbean region and by the significance of the 1960s cohort, when positively contributing for income mobility, since it was composed by more developed countries. Developing economies are the ones with the highest inequality levels, highest shares of low educated individuals, and where students perform worse in literacy assessments. All together makes the need for the above policy directions more pronounced.

There is also, as expected, a positive connection between income persistence predictions and education persistence estimates. Predicted income persistence was lower for high-income economies, which is in line with the previous findings. Therefore, implementing mobility enhancing strategies should have a positive feedback effect for generations to come, if these are grounded on a sustained economic growth.

The unemployment and poverty rates appear to determine income mobility, while the growth rate of real GDP *per capita*, the degree of urbanization, the share of female population, and income mobility should influence education mobility, although the signals of the impact of these variables are not clear.

The second chapter of this thesis analyses intergenerational mobility in income and education in Portugal for men and women born in 1968-1988. With this purpose we constructed several indicators of relative and absolute mobility in income and education. For income mobility, the intergenerational income elasticity, the intergenerational correlation coefficient, the rank-rank slope, the share of individuals earning more than their parents, and the bottom to top income level probability were computed. Regarding education mobility we used the intergenerational correlation, the low to high education level probability, and the share of individuals with more education than their fathers. These variables were complemented by ordered logit transitions matrices. Besides we analysed which individual characteristics are associated with more or less mobility. With this goal, we decomposed our estimates of mobility by education level, occupation, income level, and status in employment and compare them with the estimates for the entire sample. We also estimated the relationship between income and education relative mobility measures through Mincer (1974) equations.

The results show that women generally present more income mobility than men, which may be due to a distorted labour market. The reasoning may be that richer parents, which should have a higher educational attainment when compared to poorer parents, are more able to invest in children's education and additional activities not related with education. Besides, their offspring may have access to better social opportunities due to their social status. All together should make them more rewarded in the labour market, promoting income persistence. However, if individuals are discriminated based on attributes they can't control, this mechanism may be weakened. This is the case of women whose income is penalized for the same education level of men, as shown by the gender gap that historically exists regarding income. Therefore, the likelihood that the income level of men have of being tied to their parents' level may be stronger in comparison with women.

For sons, Portugal is found to stand amongst the most relative mobile countries in income, similar to the UK, Sweden and Australia: relative mobility is higher than in France, Spain, USA, Italy, and Brazil and lower than in Canada. For daughters, it fits in existing estimates for other countries, namely the USA and the UK, and it is higher than in France.

Transition probabilities between income levels also show that intergenerational mobility is strong when examining low income fathers and that upward mobility is lower the higher the father income level, being in accordance with estimates previously done.

We also find that men present more relative mobility in education than women. This result may be the consequence of gender differences in school dropout rates. In the period under consideration, women present higher dropout rates than men in primary education, as well as in lower secondary education levels, with their reduction being more pronounced for men, meaning that opportunities are more equalized for the last ones. However, men present a large absolute educational persistence.

Women have a higher low to high education level probability than men. Furthermore, the share of individuals with more education than their fathers surpasses the 80%, with full upward mobility (100%) verified if the father is at the low education level. The probability of reaching or remaining in the high education level is larger the more educated fathers are.

Portugal presents the highest relative mobility in education for sons, when considering 20 other European countries, Canada, and India. Again for daughters it stands in the middle of known World's estimates, presenting lower mobility in comparison with France, Nordic countries, Anglo-Saxon countries, Greece, Czech Republic, and Poland and higher mobility than Canada and Austria. A larger relative increase in government expenditures in education as a share of GDP, in comparison with other countries, may be leading our evidence, by compensating the differences in private investments in education between poor and rich parents. School dropouts can also play a role here by the same reasoning as before. This country presents the highest decrease in the male school dropout rate for the primary education, while for women it stands in the middle of existing dropout statistics. Therefore, early education may be one of the keys for educational success in the Portuguese economy.

We found that children of low educated fathers are the ones presenting higher relative mobility in income. Individuals in the medium-high income level present higher mobility. This occurs regarding mobility in education for children of medium-high-income fathers and regarding mobility in income for offspring of medium-low income fathers. Lower income mobility is presented by self-employed individuals. Higher mobility is found to be higher for individuals with a high education level, which shows the importance of having more education. This is also in line with the finding that occupations requiring a higher degree of education present higher income mobility. Individuals in elementary occupations show lower mobility and offspring of fathers working as clerks present higher educational mobility. Given these results and as in the first chapter, policies targeting early childhood development, efficient public investment in education, provision of quality education and improving the access to capital markets should be implemented to improve mobility. Strengthen institutions and infrastructures and reduce segregation should also mitigate the gap between advantaged and disadvantaged individuals.

Through Mincer (1974) equations we are able to show that relative mobility in education and income present a positive relationship, although it is found to be weak. A possible reason for this to occur is again a weak labour market, reflected by the lagged harmful effects of the unemployment increase verified due to the 2008 financial crisis, which influence future labour-related outcomes. First, it decreases the likelihood of labour demand and supply matching, reducing reservation wages and harming lifetime earnings. This is particularly punishing for young individuals which tried to enter in

the labour market for the first time. Besides, their unemployment rate also increased as a result of the recession. Therefore social protection should reduce the negative impact of unemployment. Additionally, wage subsidies and taxes reductions, stimulus to hiring young individuals, and subsidized training/employment may mitigate the penalization of youth easing their integration in the labour market.

In the last chapter of this thesis we used an indicator developed by Güell *et al*. (2015) defined as the informational content of surnames (ICS). We studied the Portuguese educational attainment and avoided the need of a panel data or explicit family links between individuals. The 1956-1995 cohorts are considered. This methodology assumes that having surnames influencing socioeconomic status is the same as belonging to a specific family determining individuals' future. This means that individuals' outcomes have a certain degree of inheritance and therefore surnames may be informative about intergenerational persistence. We analyse rare surnames, which are the ones belonging to fewer than 30 individuals (ICS30) to increase the likelihood of existence of family links. We also consider that first names, chosen by parents, should influence individuals' socioeconomic status, and define an analogous to the ICS measure, designated as the informational content of first names (ICF). We performed an analysis for the country as a whole, as well as in a disaggregated way, i.e., across different regions. Since the spatial variability of the information contained in surnames is twice that verified for first names we correlate the first with socioeconomic and political variables that differ across the Portuguese territory.

Rare surnames are responsible for 14% of the differences in educational attainment. Regions in the country's interior are the ones where the informational content of surnames is higher, therefore present a higher degree of educational intergenerational persistence. When considering first names, these explain around 2% of the variability verified for educational attainment, much less than the information contained in surnames, enhancing the role of belonging to a family in comparison with the choice made by parents regarding their children's name. Furthermore, unlike the case for surnames, the information provided by first names is higher in the country's coast.

The informational content of surnames is positively related with retention and dropout rates in primary and lower secondary education, resembling one of the possible motives for some of the results presented in the second chapter. This means that decreasing the school dropout rates in places where these are high is of utmost importance. The opposite occurs regarding imports, although the relationship is weak. No connection is found with economic activity, social capital, labour market outcomes, life expectancy, and suicide rates.

The lower inequality measured by the P80/P20 ratio is, the higher is the informational content of surnames and a convex non-linear relationship is verified when considering the P90/P10 ratio and the Gini Index. The negative connection between the informational content of surnames and inequality is unexpected. As it occurred before, a weak labour market may be one of the mechanisms behind it, due to the increase in unemployment during the Covid-19 pandemic crisis, which presents a positive correlation with inequality. This increases the difficulties to monetize human capital investments with

accepted wages being suboptimal, enhancing the need for the same policy directions as before. Second, if social connections are leading the allocation of jobs, education is not rewarded in the labour market. This not only harms economic growth but also breaks the expected relationship between mobility and inequality. Besides, high education mobility and inequality, if associated with merit, may make individuals accepting the second one as an inevitable consequence. Finally, individuals expectations about future mobility may be a cause for inequality acceptance today. Individuals should be provided with clear information and exposed to different realities, encouraging them to change the unequal *status quo*, with a positive effect as well as feedback in mobility.

The main limitations of our work may be summarized as follows. In the first chapter, the existence of different methodologies for the mobility measures available for different countries may be influencing the results obtained. In the second study, unlike what occurs with Chetty and co-authors, which work with tax records for children and parents, the unavailability of parental income in Portugal made us predict it through a pseudo-parent's sample and characteristics that children recalled about their parents. The range of available characteristics is insufficient, decreasing the heterogeneity in parental income of the sample. Furthermore, when analysing the relationship between relative mobility in education and income, we do not use the same sample as in our main analysis, since this one did not present education levels with a sufficient degree of disaggregation for us to be able to convert them in years of education. This can also drive the results we have. Finally, in the last chapter, we had to impute educational attainment through the occupations classification, which may imply some bias in our findings.

Future research may consider to have a worldwide examination of mobility with higher-frequency data, i.e., smaller than 10-year averages for each cohort, to allow a panel type analysis to complement our cross-sectional framework. To the best of our knowledge, this type of data does not exist at a global level. For Portugal it would also make sense to perform a longitudinal analysis of mobility, which may be developed either when retrospective questions about parents are presented for different years or direct information about parental income is provided. Although in the last chapter of the thesis we correlate intergenerational mobility with socioeconomic outcomes, we did not aim, at this point to find any causal mechanisms. Determine the drivers of mobility in Portugal may be a good avenue for future work.

References

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

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

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

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

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

Altonji, J., Kahn, L. & Speer, J. (2016). Cashier or Consultant? Entry Labor Market Conditions, Field of Study, and Career Success. *Journal of Labor Economics*, 34(S1), S361-401. https://doi.org/10.3386/w20531

Andrews, D. & Leigh, A. (2009). More Inequality, Less Social Mobility. *Applied Economics Letters*, 16(15), 1489-1492. https://doi.org/10.1080/13504850701720197

Arulampalam, W., Gregg, P. & Gregory, M. (2001). Unemployment Scarring. *The Economic Journal*, 111(475), 577-584. https://doi.org/10.1111/1468-0297.00663

Aura, S. & Hess, G. (2010). What's in a Name? *Economic Inquiry*, 48, 214-227. https://doi.org/10.1111/j.1465-7295.2008.00171.x

Azam, M., & Bhatt, V. (2015). Like Father, Like Son? Intergenerational Educational Mobility in India. *Demography*, 52(6), 1929-1959. https://doi.org/10.1007/s13524-015-0428-8

Bago d'Uva, T. & Fernandes, M. (2017). *Mobilidade Social em Portugal*. Estudos da Fundação Francisco Manuel dos Santos.

Bank of Portugal (2022). *Economic Bulletin: The Portuguese Economy in 2021*. Banco de Portugal.

Baptista, J., Batista, F. & Mamede, N. (2006). Building a Dictionary of Anthroponyms. In R. Vieira, P. Quaresma, M. Nunes, N. Mamede & M. Dias (Eds.), *Computational Processing of The Portuguese Language* (pp. 21-30). Springer.

Barone, G., & Mocetti, S. (2021). Intergenerational Mobility in the Very Long Run: Florence 1427-2011. *The Review of Economic Studies*, 88(4), 1863-1891. https://doi.org/10.1093/restud/rdaa075

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

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

Becker, G. & Tomes, N. (1986). Human Capital and the Rise and Fall of Families. *Journal of Labour Economics*, 4(3), S1-S39.

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

Behrman, J., Garivia, A. & Szekely, M. (2001). Intergenerational Mobility in Latin America. *Economía*, 2(1), 1-44.

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

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

Bertrand, M. & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4), 991-1013. https://doi.org/10.1257/0002828042002561

Betcherman, G., Daysal, N. & Pagés, C. (2010). Do Employment Subsidies Work? Evidence From Regionally Targeted Subsidies in Turkey. *Labour Economics*, *17*(4), 710-722. https://doi.org/10.1016/j.labeco.2009.12.002

Björklund, A. & Jäntti, M. (1997). Intergenerational Income Mobility in Sweden Compared to the United States. *American Economic Review*, 87(5), 1009-1018.

Blanden, J., Goodman, P. & Machin, S. (2004). Changes in Generational Mobility in Britain. In M. Corak (Ed.), *Generational Income Mobility in North America and Europe* (pp. 122-146). Cambridge University Press.

Blanden J., Gregg P., & Machin, S. (2002). Education and Family Income. *University College London (mimeo)*.

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

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

Böhlmark, A. & Lindquist, M. (2006). Life-Cycle Variations in the Association Between Current and Lifetime Income: Replication And Extension For Sweden. *Journal of Labor Economics*, 24(4), 879-896. https://doi.org/10.1086/506489

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

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

Brenner, J. (2010). Life-Cycle Variations in the Association Between Current and Lifetime Earnings: Evidence for German Natives and Guest Workers. *Labour Economics*, 17(2), 392-406. https://doi.org/10.1016/j.labeco.2009.03.006

Brown, P. & Richman, H. (1997). Neighborhood Effects and State and Local Policy. In J. Brooks-Gunn, G. Duncan & J. Lawrence Aber (Eds.), *Neighborhood Poverty: Policy Implications in Studying Neighborhoods* (pp. 164-181). Russell Sage Foundation.

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

Bukowski, P., Clark, G., Gáspár, A. & Pető, R. (2022). Social Mobility and Political Regimes: Intergenerational Mobility in Hungary, 1949-2017. *Journal of Population Economics*, 35, 1551-1599. https://doi.org/10.1007/s00148-021-00875-w

Burtless, G. & Jencks, C. (2003). American Inequality and Its Consequences. *LIS Working Paper Series No. 339*. Luxembourg Income Study (LIS), Luxembourg.

Campos, M. & Reis, H. (2018). Returns to Schooling in the Portuguese Economy: a Reassessment. *Public Sector Economics*, 42(2), 215-242. https://doi.org/10.3326/pse.42.2.11

Cappelen, A., Konow, J., Sørensen, E. & Tungodden, B. (2013). Just Luck: An Experimental Study of Risk-Taking and Fairness. *American Economic Review*, 103(4): 1398-1413. https://doi.org/10.1257/aer.103.4.1398

Carneiro, P. (2008). Equality of Opportunity and Educational Achievement in Portugal. *Portuguese Economic Journal*, 7(1), 17-41. https://doi.org/10.1007/s10258-007-0023-z

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

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

Chadwick, L. & Solon, G. (2002). Intergenerational Income Mobility Among Daughters. *American Economic Review*, 92(1), 335-344. https://doi.org/10.1257/000282802760015766

Chen, T., Hallaert, J., Pitt, P., Qu, A., Queyranne, M., Rhee, A., Shabunina, A., Vandenbussche, J. & Yackovlev, I. (2018). Inequality and Poverty Across Generations in the European Union. *International Monetary Fund Staff Discussion Note 18/01*. https://doi.org/10.5089/9781484338445.006

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

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

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

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

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

Chetty, R., Hendren, N. & Katz, L. (2016). The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review*, 106(4), 855-902. https://doi.org/10.1257/aer.20150572

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

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

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

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

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

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

Clark, A., Georgellis, Y. & Sanfey, P. (2001). Scarring: the Psychological Impact of Past Unemployment. *Economica*, 68(270), 221-241. https://doi.org/10.1111/1468-0335.00243

Clark, G., & Cummins, N. (2012). Are There Ruling Classes? Surnames and Social Mobility in England, 1800-2011. *University of California, Davis (mimeo)*.

Clark, G., & Cummins, N. (2012). Intergenerational Wealth Mobility in England, 1858-2012: Surnames and Social Mobility. *The Economic Journal*, 125(582), 61-85.

Clark, G., & Ishii, T. (2012). Social Mobility in Japan , 1868-2012: The Surprising Persistence of the Samurai. *University of California, Davis (mimeo)*.

Clark, G., & Landes, Z. (2013). Caste Versus Class: Social Mobility in India, 1860-2012. *University of California, Davis (mimeo)*.

Clark, G., Cummins, N., Hao, Y. & Vidal, D. (2015). Surnames: a New Source for the History of Social Mobility. *Explorations in Economic History*, 55, 3-24. https://doi.org/10.1016/j.eeh.2014.12.002

Clark, G., Leigh, A. & Pottenger, M. (2017). Immobile Australia: Surnames Show Strong Status Persistence, 1870-2017. *IZA Discussion Paper No. 11021*.

Clements, B. (1999). The Efficiency of Education Expenditure in Portugal. *International Monetary Fund Staff Working Paper 99/179*. https://doi.org/10.5089/9781451859010.001

Coenjaerts, C., Ernst, C., Fortuny, M. Rei, D. & Pilgrim, M. (2009). Youth Employment. In OECD (Ed.), *Promoting Pro-poor Growth: Employment* (pp. 119-131). OECD.

Collado, M., Romeu, A. & Ortin, I. (2013). Long-run Intergenerational Social Mobility and the Distribution of Surnames. *UMUFAE Economics Working Papers 36768.*

Comi, S. (2013). Intergenerational Mobility in Europe: Evidence from ECHP. *Università degli studi di Milano (mimeo)*.
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.
Corak, M. (2013). Income Inequality, Equality of Opportunity, and Intergenerational Mobility. *Journal of Economic Perspectives*, 27(3), 79-102. https://doi.org/10.1257/jep.27.3.79
Corak, M. (2019). The Canadian Geography of Intergenerational Income Mobility. *The Economic Journal*, 130(631), 2134-2174. https://doi.org/10.1093/ej/uez019
Dahl, M., & DeLeire, T. (2008). The Association Between Children's Earnings and Fathers' Lifetime Earnings: Estimates Using Administrative Data. *Institute for Research on Poverty Discussion Paper No. 1342-08*.
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
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
Dearden, M. & Reed, H. (1997). Intergenerational Mobility in Britain. *Economic Journal*, 107(440), 47-66. https://doi.org/10.1111/1468-0297.00141
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
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
Duncan, G. & Murnane, R. (2011). Introduction: the American Dream, Then and Now. In G. Duncan & R. Murnane (Eds.), *Whither Opportunity?: Rising Inequality, Schools, and Children's Life Chances*. Russell Sage Foundation.
Dunn, C. (2007). The Intergenerational Transmission of Lifetime Earnings: Evidence from Brazil. *The B.E. Journal of Economic Analysis & Policy*, 7(2). https://doi.org/10.2202/1935-1682.1782
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
Eriksen, J. & Munk, M. (2020). The Geography of Intergenerational Mobility – Danish Evidence. *Economic Letters*, 189. https://doi.org/10.1016/j.econlet.2020.109024
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
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
Fehr, E. & Fischbacher, U. (2003). The Nature of Human Altruism. *Nature*, 425(6960), 785-791. https://doi.org/10.1038/nature02043
Ferreira, S. & Veloso, F. (2006). Intergenerational Mobility of Wages in Brazil. *Brazilian Review of Econometrics*, 26(2), 181-211. https://doi.org/10.12660/bre.v26n22006.1576
Figlio, D. (2005). Names, Expectations and The Black-white Test Score Gap. *National Bureau of Economic Research Working Paper No. 11195*. https://doi.org/10.3386/w11195
Figlio, D. (2007). Boys Named Sue: Disruptive Children and Their Peers. *Education Finance and Policy*, 2(4), 376-394. https://doi.org/10.1162/edfp.2007.2.4.376
Fleib, J. (2015). Merit Norms in the Ultimatum Game: an Experimental Study of the Effect of Merit on Individual Behavior and Aggregate Outcomes. *Central European Journal of Operations Research*, 23(2), 389-406. https://doi.org/10.1007/s10100-015-0385-8
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
Fortin, N. & Lefebvre, S. (1998). Intergenerational Income Mobility in Canada. In M. Corak (Ed.), *Labor Markets, Social Institutions and the Future of Canada´s Children*. Statistics of Canada.
Fox, W. & Lasker, G. (1983). The Distribution of Surname Frequencies. *International Statistical Review*, 51(1), 81-87. https://doi.org/10.2307/1402733

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
Friedman, M. (1957). *A Theory of the Consumption Function.* Princeton University Press.
Fryer, R. & Levitt, S. (2004). The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics*, 119(3), 767-805. https://doi.org/10.1162/0033553041502180
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
Galor, O., & Tsiddon, D. (1997). Technological Progress, Mobility, and Economic Growth. *American Economic Review*, 87(3), 363-382.
GDIM (2018). *Global Database on Intergenerational Mobility*. Development Research Group, World Bank. Washington, D.C.: World Bank Group.
Geng, Y. (2021). Intergenerational Mobility in China across Space and Time. *University of Liverpool, Department of Economics (mimeo)*.
Global Change Data Lab (2021). *Our World in Data Project*.
Groh, M., Krishnan, N., McKenzie, D. & Vishwanath, T. (2016). Do Wage Subsidies Provide a Stepping-Stone to Employment for Recent College Graduates? Evidence from a Randomized Experiment in Jordan. *Review of Economics and Statistics*, 98(3), 488-502. https://doi.org/10.1162/REST_a_00584
Güell, M., Mora, J. & Telmer, C. (2015). The Informational Content of Surnames, the Evolution of Intergenerational Mobility, and Assortative Mating. *The Review of Economic Studies*, 82(2), 693-735. https://doi.org/10.1093/restud/rdu041
Güell, M., Pellizzari, M., Pica, G. & Mora, J. (2018). Correlating Social Mobility and Economic Outcomes. *The Economic Journal*, 128(612), F353-F403. https://doi.org/10.1111/ecoj.12599
Haider, S. & Solon, G. (2006). Life-Cycle Variation in the Association Between Current and Lifetime Earnings. *American Economic Review*, 96(4): 1308-1320. https://doi.org/10.1257/aer.96.4.1308
Hao, Y. (2021). Social Mobility in China, 1945-2012: A Surname Study. *China Economic Quarterly International*, 1(3), 233-243. https://doi.org/10.1016/j.ceqi.2021.08.002
Hassler, J. & Mora, J. (2000). Intelligence, Social Mobility, and Growth. *American Economic Review*, 90(4), 888-908. https://doi.org/10.1257/aer.90.4.888
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
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
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
Hirschman, A. & Rothschild, M. (1973). The Changing Tolerance for Income Inequality in the Course of Economic Development: With a Mathematical Appendix. *Quarterly Journal of Economics*, 87 (4): 544-566. https://doi.org/10.2307/1882024
Hodrick, R. & Prescott, E. (1997). Postwar U.S. Business Cycles: an Empirical Investigation. *Journal of Money, Credit and Banking*, 29(1), 1-16.
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*. J.D. Systems Institute & WVSA Secretariat.
International Monetary Fund (2018). *Global Debt Database*. International Monetary Fund.
Jenkins S. (1987). Snapshots Versus Movies: "Lifecycle Biases" and the Estimation of Intergenerational Earnings Inheritance. *European Economic Review*, 31(5) 1149-1158. https://doi.org/10.1016/S0014-2921(87)80010-7
Joseph, V. (2022). Optimal Ratio for Data Splitting. *Statistical Analysis and Data Mining*, 531-538. https://doi.org/10.1002/sam.11583
Kalist, D. & Lee, D. (2009). First Names and Crime: Does Unpopularity Spell Trouble? *Social Science Quarterly*, 90(1), 39-49. https://doi.org/10.1111/j.1540-6237.2009.00601.x

Katz, L. (1998). Wage Subsidies for the Disadvantaged. In R. Freeman & P. Gottschalk (Eds.), *Generating Jobs: How to Create Demand for Low-Skilled Workers* (pp. 21-53). Russell Sage Foundation.
Kearney, M. & Levine, P. (2016). Income Inequality, Social Mobility, and the Decision to Drop Out of High School. *Brookings Papers on Economic Activity*, 333-380.
Kline, P. & Moretti, E. (2014). People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs. *Annual Review of Economics*, 6, 629-662. https://doi.org/10.1146/annurev-economics-080213-041024
Knudsen, E., Heckman, J., Cameron, J. & Shonkoff, J. (2006). Economic, Neurobiological, And Behavioral Perspectives on Building America's Future Workforce. *PNAS*, 103(27), 10155-10162. https://doi.org/10.1073/pnas.0600888103
Kourtellos, A. (2021). The Great Gatsby Curve in Education with a Kink. *Economic Letters*, 208. https://doi.org/10.1016/j.econlet.2021.110054
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
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
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
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
Lee, C. & Solon, G. (2009). Trends in Intergenerational Income Mobility. *The Review of Economics and Statistics*, 91(4), 766-772. https://doi.org/10.1162/rest.91.4.766
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
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
Lefranc, A. & Trannoy, A. (2005). Intergenerational Earnings Mobility in France: Is France More Mobile Than the US? *Annales d'Economie et de Statistique*, 78, 57-77. https://doi.org/10.2307/20079128
Leigh, A. (2007). Intergenerational Mobility in Australia. *The B.E. Journal of Economic Analysis & Policy*, 7(2). https://doi.org/10.2202/1935-1682.1781
Leone, T. (2022). The Geography Of Intergenerational Mobility: Evidence of Educational Persistence and the "Great Gatsby Curve" In Brazil. *Review of Development Economics*, 26(3), 1227-1251. https://doi.org/10.1111/rode.12880
Levitt, S. & Dubner, S. (2006). *Freakonomics*. Harper Trophy.
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
Lundberg, S. & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*.
Maoz, Y. & Moav, O. (1999). Intergenerational Mobility and the Process of Development. *The Economic Journal*, 109(458), 677-697. https://doi.org/10.1111/1468-0297.00468
Mazumder, B. (2005). Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data. *Review of Economics and Statistics*, 87(2), 235-255. https://doi.org/10.1162/0034653053970249
Mendolia, S. & Siminski, P. (2019). New Estimates of Intergenerational Mobility in Australia. *Economic Record*, 92(298), 361-373. https://doi.org/10.1111/1475-4932.12274
Mincer, J. (1974). *Schooling, Experience and Earnings*. National Bureau of Economic Research.
Mocetti, S. (2007). Intergenerational Earnings Mobility in Italy. *The B.E. Journal of Economic Analysis & Policy*, 7(2). https://doi.org/10.2202/1935-1682.1794

Muñoz, E. & Siravegna, M. (2021). When Measure Matters: Coresidence Bias and Intergenerational Mobility Revisited. *SSRN (mimeo)*. https://doi.org/10.2139/ssrn.3969270
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
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.
Neidhöfer, G. (2019). Intergenerational Mobility and the Rise and Fall of Inequality: Lessons From Latin America. *The Journal of Economic Inequality*, 17, 499-520. https://doi.org/10.1007/s10888-019-09415-9
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
Nicoletti, C. & Ermisch, J. (2008). Intergenerational Earnings Mobility: Changes Across Cohorts in Britain. *The B.E. Journal of Economic Analysis and Policy*, *7*(2), 1-38. https://doi.org/10.2202/1935-1682.1755
Nicoletti, C. & Francesconi, M. (2006). Intergenerational Mobility and Sample Selection in Short Panels. *Journal of Applied Econometrics*, *21*(8), 1265-1293. https://doi.org/10.1002/jae.910
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
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
Nuñez, J. & Miranda, L. (2010). Intergenerational Income Mobility in a Less-developed, High-Inequality Context: the Case of Chile. *The B.E. Journal of Economic Analysis & Policy*, 10(1). https://doi.org/10.2202/1935-1682.2339
Nybom, M. (2018). Intergenerational Mobility: A Dream Deferred? *International Labour Organization Research Paper Series.*
Nybom, M. & Stuhler, J. (2016). Heterogeneous Income Profiles and Life-Cycle Bias in Intergenerational Mobility Estimation. *Journal of Human Resources*, 51(1), 239-268. https://doi.org/10.3368/jhr.51.1.239
Nybom, M. & Stuhler, J. (2017). Biases in Standard Measured of Intergenerational Income Dependence. *Journal of Human Resources*, 52(3), 800-825. https://doi.org/10.3368/jhr.52.3.0715-7290R
OECD (2012). *Economic Policy Reforms 2012: Going for Growth*. OECD Publishing.
OECD (2018). *A Broken Social Elevator? How to Promote Social Mobility*. OECD Publishing.
OECD (2019). *Education at a Glance 2019: OECD Indicators*. OECD Publishing.
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
Owen, A. & Weil, D. (1998). Intergenerational Earnings Mobility, Inequality and Growth. *Journal of Monetary Economics*, 41(1), 71-104. https://doi.org/10.1016/S0304-3932(97)00067-6
Özalp, L. (2019). Intergenerational Income Mobility. In K. Natsuda, B. Yilmaz & K. Otsuka (Eds.), *Globalisation & Public Policy*, (pp. 25-34). IJOPEC Publication.
Pagan, A. (1984). Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, 25(1), 221-247.
Pereira, P. (2010). Higher Education Attainment: The Case of Intergenerational Transmission of Education in Portugal. *IZA Discussion Paper No. 4813*. https://doi.org/10.2139/ssrn.1570431
Pereira, P. & Martins, P. (2004). Returns to Education and Wage Equations. *Applied Economics*, 36(6), 525-531. https://doi.org/10.1080/0003684042000217571
Piraino, P. (2007). Comparable Estimates of Intergenerational Income Mobility in Italy. *The B.E. Journal of Economic Analysis & Policy*, 7(2). https://doi.org/10.2202/1935-1682.1711
Piraino, P. (2015). Intergenerational Earnings Mobility and Equality of Opportunity in South Africa. *World Development*, 67, 396-405. https://doi.org/10.1016/j.worlddev.2014.10.027

Pissarides, C. (1992). Loss of Skill During Unemployment and the Persistence of Employment Shocks. *Quarterly Journal of Economics*, 107(4), 1371-1391. https://doi.org/10.2307/2118392

Raaum, O. & Røed, K. (2006). Do Business Cycle Conditions at the Time of Labor Market Entry Affect Future Employment Prospects? *The Review of Economics and Statistics*, 88(2), 193-210. https://doi.org/10.1162/rest.88.2.193

Raaum, O., Bratsberg, B., Røed, K., Österbacka, E., Eriksson, T., Jäntti, M. & Naylor, R. (2008). Marital Sorting, Household Labor Supply, and Intergenerational Earnings Mobility Across Countries. *The B.E. Journal of Economic Analysis & Policy*, 7(2). https://doi.org/10.2202/1935-1682.1767

Reis, H. & Campos, M. (2017). Revisiting the Returns to Schooling in the Portuguese Economy. *Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies*. Banco de Portugal.

Schneebaum, A., Rumplmaier, B. & Altzinger, W. (2014). International Educational Persistence in Europe. *Vienna University of Economics and Business Department of Economics Working Papers wuwp174*.

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

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.

Solon, G. (1992). Intergenerational Income Mobility in the United States. *American Economic Review*, 82(3), 393-408.

Solon, G. (2002). Cross-country Differences in Intergenerational Earnings Mobility. *Journal of Economic Perspectives*, 16(3), 59-66. https://doi.org/10.1257/089533002760278712

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.

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

Torche, F. (2015). Analyses of Intergenerational Mobility: an Interdisciplinary Review. *Annals of the American Academy of Political and Social Science*, 657(1), 37-62. https://doi.org/10.1177/0002716214547476

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

World Bank (2018a). *World Development Indicators*. World Bank Group.

World Bank (2018b). *World Development Report 2018: Learning to Realize Education's Promise.* World Bank Group.

Appendix A

A1 – Variables

- **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 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* 1,000 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)**

This variable is given by the number of marriages (marr) as a share of mean population, *per* 1,000 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* 1,000 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 (2017) 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* 1,000 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.

A2 – Tables

Table A1 – Descriptive Statistics for Determinants of IM in Income

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
avinc	70	22,271.06	3,448.42	11,116.65	34,088.88
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
IGPI	70	0.53	0.25	0.11	1.10
IGPE	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	65,301.27	544,779.00	-837,680.70	3,939,991.00
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

(continues in the next page)

Table A1 – Descriptive Statistics for Determinants of IM in Income *(continued)*

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
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 A2 – 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	17,271.06	5,287.45	10,981.55	34,859.54
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
IGPI	338	0.54	0.18	0.11	1.10
IGPE	338	0.39	0.16	-0.21	0.98
infl	338	45.05	114.12	0.34	1,082.08
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

(continues in the next page)

Table A2 – Descriptive Statistics for IM in Education *(continued)*

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
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

Table A3 – 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.

A3 – Figures

Figure A.1 – Intergenerational Persistence of Income for the 1960 Cohort

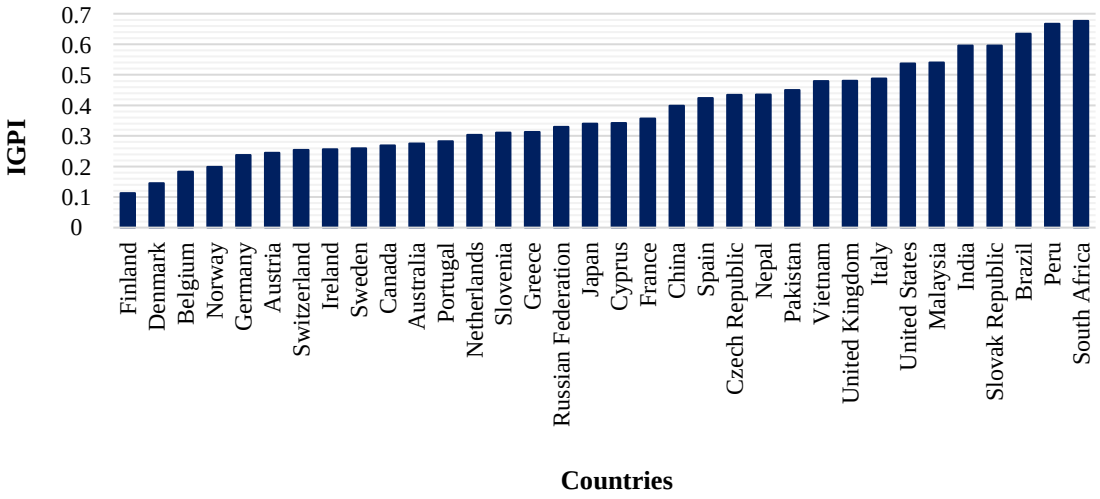
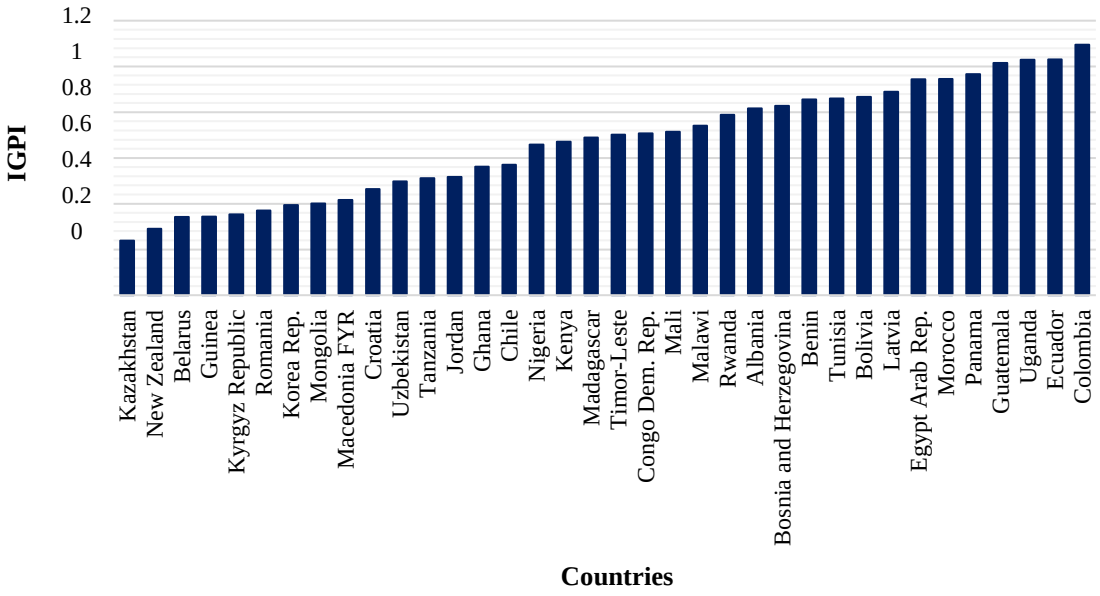


Figure A.2 – Intergenerational Persistence of Income for the 1970 Cohort



Appendix B

B1 – Tables

Table B1 – Summary (unweighted) statistics: pseudo-parents' sample

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
Age	1,025	39.23	5.85	30	50
Education: Low level	1,025	0.84	0.36	0	1
Education: Medium level	1,025	0.11	0.31	0	1
Education: High level	1,025	0.05	0.21	0	1
Main occupation: Legislators, senior officials and managers	1,025	0.02	0.13	0	1
Main occupation: Professionals	1,025	0.04	0.20	0	1
Main occupation: Technicians and associate professionals	1,025	0.09	0.29	0	1
Main occupation: Clerks	1,025	0.10	0.30	0	1
Main occupation: Service workers and shop and market sales workers	1,025	0.14	0.35	0	1
Main occupation: Skilled agricultural and fishery workers	1,025	0.04	0.20	0	1
Main occupation: Craft and related trades workers	1,025	0.29	0.45	0	1
Main occupation: Plant and machine operators and assemblers	1,025	0.16	0.37	0	1
Main occupation: Elementary occupations	1,025	0.12	0.32	0	1
Managerial position: Supervisory	1,025	0.07	0.26	0	1
Managerial position: Non-supervisory	1,025	0.93	0.26	0	1
Individual income (in logs)	1,025	9.07	0.56	6.31	11.03

Table B2 – Summary (unweighted) statistics

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
Children's characteristics					
Age	2,549	40.70	5.93	30	50
Education: low level	2,549	0	0	0	0
Education: medium level	2,549	0.51	0.50	0	1
Education: high level	2,549	0.49	0.50	0	1
Individual income (in logs)	2,549	9.10	0.73	4.54	11.64
Father's characteristics (recalled by children)					
Education: low level	2,549	0.80	0.40	0	1
Education: medium level	2,549	0.11	0.31	0	1
Education: high level	2,549	0.09	0.29	0	1
Main occupation: Legislators, senior officials and managers	2,549	0.05	0.22	0	1
Main occupation: Professionals	2,549	0.08	0.28	0	1
Main occupation: Technicians and associate professionals	2,549	0.16	0.36	0	1
Main occupation: Clerks	2,549	0.07	0.26	0	1
Main occupation: Service workers and shop and market sales workers	2,549	0.14	0.35	0	1
Main occupation: Skilled agricultural and fishery workers	2,549	0.05	0.21	0	1
Main occupation: Craft and related trades workers	2,549	0.23	0.42	0	1
Main occupation: Plant and machine operators and assemblers	2,549	0.14	0.35	0	1
Main occupation: Elementary occupations	2,549	0.07	0.26	0	1

(continues in the next page)

Table B2 – Summary (unweighted) statistics *(continued)*

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
Managerial position: Supervisory	2,549	0.28	0.45	0	1
Managerial position: Non-supervisory	2,549	0.72	0.45	0	1
Father's predicted individual income (in logs)	2,549	9.25	0.62	3.95	11.94

Table B3 – First stage results: 1995 pseudo-parents' sample

Variable	Benchmark	Educ.	Occ.	Man. Pos.	Educ. & Occ.	Educ. & Man. Pos.	Occ. & Man. Pos.
Education: Medium level	0.10* (0.06)	0.32*** (0.08)			0.12** (0.06)	0.26*** (0.07)	
Education: High level	0.46*** (0.13)	1.10*** (0.06)			0.53*** (0.15)	0.84*** (0.08)	
Main occupation: Professionals	0.33** (0.17)		0.38* (0.22)		0.18 (0.18)		0.52*** (0.20)
Main occupation: Technicians and associate professionals	0.19 (0.19)		-0.20 (0.22)		-0.02 (0.21)		0.05 (0.20)
Main occupation: Clerks	-0.001 (0.19)		-0.51** (0.22)		-0.29 (0.21)		-0.17 (0.20)
Main occupation: Service workers and shop and market sales workers	0.05 (0.20)		-0.46** (0.22)		-0.21 (0.22)		-0.15 (0.20)
Main occupation: Skilled agricultural and fishery workers	-0.54** (0.22)		-1.09*** (0.24)		-0.83*** (0.24)		-0.74*** (0.23)
Main occupation: Craft and related trades workers	-0.28 (0.19		-0.82*** (0.21)		-0.56*** (0.21)		-0.48** (0.20)
Main occupation: Plant and machine operators and assemblers	-0.20 (0.20)		-0.73*** (0.21)		-0.47** (0.22)		-0.40** (0.20)
Main occupation: Elementary occupations	-0.55*** (0.21)		-1.09*** (0.22)		-0.84*** (0.23)		-0.75*** (0.21)
Managerial Position: Supervisory	0.40*** (0.08)			0.86*** (0.07)		0.52*** (0.05)	0.53*** (0.08)
Age	-0.002	0.01 (0.06)	0.01 (0.05)	0.01 (0.06)	0.002 (0.05)	0.005 (0.05)	-0.0001 (0.05)
Age^2	0.00005 (0.001)	-0.0001 (0.001)	-0.0001 (0.001)	-0.0001 (0.001)	-6.43e-06 (0.001)	0.00001 (0.001)	0.00001 (0.001)
Intercept	9.20*** (0.95)	8.61*** (1.05)	9.60*** (0.98)	8.64*** (1.06)	9.37*** (0.98)	8.78*** (1.01)	9.38*** (0.95)
No. of Observations	**1,025**	**1,025**	**1,025**	**1,025**	**1,025**	**1,025**	**1,025**
Total Population	**774,800**	**774,800**	**774,800**	**774,800**	**774,800**	**774,800**	**774,800**

Notes: Standard errors are in parentheses. *, ** and *** stand for statistically significant at 10%, 5%, 1% levels, respectively. Parental individual income (in logs) is predicted at the age of 40 years old.

Table B4 – Predicted Probabilities for Income Mobility using an Ordered Logit

	Father Income Levels	Children's Income Levels			
		Low	Medium-low	Medium-high	High
All individuals	**Low**	35.88***	40.86***	16.11***	7.15***
	n=275 | N = 68,264	(0.02)	(0.01)	(0.01)	(0.01)
	Medium-low	27.27***	41.59***	20.83***	10.31***
	n=1262 | N = 488,814	(0.01)	(0.01)	(0.01)	(0.01)
	Medium-high	20.08***	39.62***	25.65***	14.65***
	n=451 | N = 182,071	(0.01)	(0.01)	(0.01)	(0.01)
	High	14.41***	35.41***	29.79***	20.39***
	n=561 | N = 240,934	(0.01)	(0.02)	(0.02)	(0.02)
Males	**Low**	31.65***	42.83***	17.32***	8.20***
	n=98 | N = 29,180	(0.03)	(0.02)	(0.02)	(0.01)
	Medium-low	23.47***	42.43***	22.21***	11.89***
	n=484 | N = 203,553	(0.02)	(0.02)	(0.02)	(0.01)
	Medium-high	16.88***	39.25***	26.94***	16.93***
	n=185 | N = 81,735	(0.02)	(0.02)	(0.02)	(0.02)
	High	11.85***	34.01***	30.59***	23.55***
	n=260 | N = 117,382	(0.02)	(0.02)	(0.02)	(0.03)
Females	**Low**	38.56***	39.6***	15.42***	6.42***
	n=177 | N = 39,084	(0.03)	(0.02)	(0.02)	(0.01)
	Medium-low	30.07***	40.96***	19.87***	9.10***
	n=778 | N = 285,261	(0.02)	(0.02)	(0.01)	(0.01)
	Medium-high	22.76***	39.93***	24.56***	12.75***
	n=266 | N = 100,336	(0.02)	(0.02)	(0.02)	(0.01)
	High	16.79***	36.72***	28.91***	17.58***
	n=301 | N = 123,552	(0.02)	(0.02)	(0.02)	(0.02)

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities are expressed in %. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

Table B5 – Predicted Probabilities for Education Mobility using an Ordered Logit

	Father Education Levels	Children's Education Levels	
		Medium	High
All individuals	**Low**	55.65***	44.35***
	n=2,040 | N = 745,593	(0.01)	(0.01)
	Medium	34.41***	65.59***
	n=275 | N = 124,035	(0.02)	(0.02)
	High	17.98***	82.02***
	n=234 | N = 110,455	(0.03)	(0.03)
Males	**Low**	63.07***	36.93***
	n=800 | N = 321,123	(0.02)	(0.02)

(continues in the next page)

Table B5 – Predicted Probabilities for Education Mobility using an Ordered Logit *(continued)*

	Father Education Levels	Children's Education Levels	
		Medium	High
Males	**Medium**	45.07***	54.93***
	n=131 | N = 60,726	(0.03)	(0.03)
	High	28.28***	71.72***
	n=96 | N = 50,001	(0.05)	(0.05)
Females	**Low**	50.02***	49.98***
	n=1240 | N = 424,470	(0.02)	(0.02)
	Medium	24.35***	75.65***
	n=144 | N = 63,309	(0.02)	(0.02)
	High	9.38***	90.62***
	n=138 | N = 60,454	(0.02)	(0.02)

Notes: Standard errors are in parentheses. *** stands for statistically significant at 1% level. Probabilities are expressed in %. n stands for the number of observations used and N for the total population represented by those observations using survey weights.

B2 – Proof of Theorem 2.1

Consider two Mincer (1974) equations, for children and for parents, which measure the change in logged current income (y_{it}) due to an additional year of current maximum education attained (Ed_{it}), given by ϱ, after controlling for other factors that may influence the first variable (W_{it}), which may comprehend individual, as well as employer characteristics. Moreover, consider the standard equation used to estimate intergenerational income elasticity, accounting for life-cycle effects. These are, respectively:

$$\begin{cases} y_{it}^{c} = \varrho^{c} Ed_{it}^{c} + \chi'^{c} W_{it}^{c} + u_{it}^{c} \\ y_{it}^{p} = \varrho^{p} Ed_{it}^{p} + \chi'^{p} W_{it}^{p} + u_{it}^{p} \\ y_{it}^{c} = \beta_0 + \beta_1 y_{it}^{p} + \gamma_1^{c} A_{it}^{c} + \gamma_2^{c} {A_{it}^{c}}^{2} + \alpha_{it} \\ Ed_{i}^{c} = \partial_0 + \partial_1 Ed_{i}^{p} + \pi_i \end{cases} \tag{2.17}$$

Assuming that y_{it}^{p} is orthogonal with respect to A_{it}^{c} and ${A_{it}^{c}}^{2}$, the multivariate regression coefficient corresponds to the univariate regression coefficient. Thus, we have that $\beta_1 = \frac{Cov(y_{it}^{c}, y_{it}^{p})}{Var(y_{it}^{p})}$, which, for the Mincer-type equation implies[46]:

$$\beta_1 = \frac{Cov\left(\varrho^{c} Ed_{i}^{c} + \chi'^{c} W_{i}^{c} + u_{i}^{c}, \varrho^{p} Ed_{i}^{p} + \chi'^{p} W_{i}^{p} + u_{i}^{p}\right)}{Var\left(\varrho^{p} Ed_{i}^{p} + \chi'^{p} W_{i}^{p} + u_{i}^{p}\right)} =$$

$$= \frac{\varrho^{c}\varrho^{p} Cov\left(Ed_{i}^{c}, Ed_{i}^{p}\right) + Cov\left(\chi'^{c} W_{i}^{c} + u_{i}^{c}, \varrho^{p} Ed_{i}^{p}\right) + Cov\left(\varrho^{c} Ed_{i}^{c} + \chi'^{c} W_{i}^{c} + u_{i}^{c}, \chi'^{p} W_{i}^{p} + u_{i}^{p}\right)}{Var\left(\varrho^{p} Ed_{i}^{p} + \chi'^{p} W_{i}^{p} + u_{i}^{p}\right)} =$$

[46] We will omit the subscript *t* for simplicity.

$$= \frac{\varrho^c \varrho^p \partial_1 Var\left(Ed_i^p\right) + Cov\left(\chi'^c W_i^c + u_i^c, \varrho^p Ed_i^p\right) + Cov\left(\varrho^c Ed_i^c + \chi'^c W_i^c + u_i^c, \chi'^p W_i^p + u_i^p\right)}{Var\left(\varrho^p Ed_i^p + \chi'^p W_i^p + u_i^p\right)} =$$

$$= \frac{\varrho^c \varrho^p \partial_1 Var\left(Ed_i^p\right) + Cov\left(\chi'^c W_i^c + u_i^c, \varrho^p Ed_i^p\right) + Cov\left(\varrho^c Ed_i^c + \chi'^c W_i^c + u_i^c, \chi'^p W_i^p + u_i^p\right)}{(\varrho^p)^2 Var\left(Ed_i^p\right) + Var\left(\chi'^p W_i^p\right) + Var\left(u_i^p\right) + 2\varrho^p Cov\left(Ed_i^p, \chi'^p W_i^p\right)} =$$

$$= \frac{\varrho^c \varrho^p \partial_1 + \frac{Cov\left(\chi'^c W_i^c + u_i^c, \varrho^p Ed_i^p\right) + Cov\left(\varrho^c Ed_i^c + \chi'^c W_i^c + u_i^c, \chi'^p W_i^p + u_i^p\right)}{Var\left(Ed_i^p\right)}}{(\varrho^p)^2 + \frac{Var\left(\chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{Var\left(u_i^p\right)}{Var\left(Ed_i^p\right)} + 2\varrho^p \frac{Cov\left(Ed_i^p, \chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)}}$$

We are now able to uncover how mobility in income responds to changes in mobility in education:

$$\frac{d\beta_1}{d\partial_1} = \frac{\varrho^c \varrho^p}{(\varrho^p)^2 + \frac{Var\left(\chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{Var\left(u_i^p\right)}{Var\left(Ed_i^p\right)} + 2\varrho^p \frac{Cov\left(Ed_i^p, \chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)}} =$$

$$= \frac{1}{\frac{\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(\chi'^p W_i^p + u_i^p\right)}{Var\left(Ed_i^p\right)} + 2\frac{Cov\left(Ed_i^p, \chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)}} =$$

$$= \frac{1}{\frac{\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(\chi'^p W_i^p + u_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{2}{\varrho^c}\frac{Cov\left(Ed_i^p, \chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)}} =$$

$$= \frac{1}{\frac{\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(y_i^p - \varrho^p Ed_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{2}{\varrho^c}\frac{Cov\left(Ed_i^p, \chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)}} =$$

$$= \frac{1}{\frac{\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(y_i^p\right) + (\varrho^p)^2 Var\left(Ed_i^p\right) - 2\varrho^p Cov\left(\hat{y}_{it}^p, Ed_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{2}{\varrho^c}\frac{Cov\left(Ed_i^p, \chi'^p W_i^p\right)}{Var\left(Ed_i^p\right)}} =$$

$$= \frac{1}{\frac{2\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{2}{\varrho^c}\left[\frac{Cov\left(Ed_i^p, \chi'^p W_i^p\right) - Cov\left(y_i^p, Ed_i^p\right)}{Var\left(Ed_i^p\right)}\right]} =$$

$$= \frac{1}{\frac{2\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{2}{\varrho^c}\left[\frac{Cov\left(Ed_i^p, \chi'^p W_i^p - y_i^p\right)}{Var\left(Ed_i^p\right)}\right]} =$$

$$= \frac{1}{\frac{2\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{2}{\varrho^c}\left[\frac{Cov\left(Ed_i^p, -\varrho^p Ed_i^p - u_i^p\right)}{Var\left(Ed_i^p\right)}\right]} =$$

$$= \frac{1}{\frac{2\varrho^p}{\varrho^c} + \frac{1}{\varrho^c \varrho^p} \frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)} + \frac{2}{\varrho^c}\left[\frac{Cov\left(Ed_i^p, -\varrho^p Ed_i^p\right)}{Var\left(Ed_i^p\right)}\right]} =$$

$$= \frac{1}{\frac{2\varrho^p}{\varrho^c} + \frac{1}{\varrho^c\varrho^p}\frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)} - \frac{2\varrho^p}{\varrho^c}\left[\frac{Cov\left(Ed_i^p, Ed_i^p\right)}{Var\left(Ed_i^p\right)}\right]} = \frac{1}{\frac{2\varrho^p}{\varrho^c} + \frac{1}{\varrho^c\varrho^p}\frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)} - \frac{2\varrho^p}{\varrho^c}\left[\frac{Var\left(Ed_i^p\right)}{Var\left(Ed_i^p\right)}\right]}$$

$$= \frac{1}{\frac{2\varrho^p}{\varrho^c} + \frac{1}{\varrho^c\varrho^p}\frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)} - \frac{2\varrho^p}{\varrho^c}} = \frac{1}{\frac{1}{\varrho^c\varrho^p}\frac{Var\left(y_i^p\right)}{Var\left(Ed_i^p\right)}} = \varrho^c\varrho^p\frac{Var\left(Ed_i^p\right)}{Var\left(y_i^p\right)}$$

■

Appendix C

C1 – Macroeconomic Outcomes: Variables Information

- **Average population**: arithmetic mean of the population observed at the end of two successive years.
- **Resident population**: number of individuals that lived or arrived with the purpose of living in a given residence place for a continuous 12-month period before the moment of observation, whether present in that moment or not.
- **Nominal GDP**: value of goods and services produced by productive units that reside in Portugal.
- **Real GDP *per capita***: ratio between the value of goods and services produced by productive units that reside in Portugal and the consumer price index considering a set of consumption representative goods and services, having 2012 as a base year, as a share of average population.
- **Retention and desistance rate in a given education level (%)**: number of students enrolled in a given education level that remain in the same school year due to voluntary attempt to improve qualifications or failure, as a percentage of the number of students enrolled in the same education level in that academic year.
- **P80/P20 ratio**: ratio between the 80th percentile of declared gross income net from IRS and the 20th percentile of declared gross income net from IRS.
- **P90/P10 ratio**: ratio between the 90th percentile of declared gross income net from IRS and the 10th percentile of declared gross income net from IRS.
- **Gini coefficient**: the asymmetry of the income distribution, ranging between 0 and 100, where 0 indicates that everyone has the same income and 100 indicates that one person detains all of the income.
- **Voter turnout in a given election**: number of actual voters in a given election as a share of the number of citizens meeting the legal requirements to vote for that election.
- **Registered unemployment**: number of people, given a minimum specific age, enrolled at the public employment office who have no job, are seeking a job, and are available for work.
- **Unemployment as a % of resident population given a specific age range**: monthly average number of registered unemployed persons as a percentage share of the average resident population with a given age range.
- **Employment**: total individuals working in a productive activity that fits the definition of production.
- **Imports/GDP**: value of goods entering the territory that come from another territory as a share of nominal GDP.
- **Exports/GDP**: value of goods leaving the territory and entering another territory as a share of nominal GDP.

- **Life expectancy at a specific age**: average number of years an individual of a specific age is expected to live, given the current age probabilities of dying, computed using life tables for three consecutive years.
- **Suicide rate**: number of deaths by suicide as a percentage share of total number of deaths, in a given year.
- **Crime**: action declared and described by law as liable of conviction for felony, detected by/brought to the knowledge of police authorities.
- **Municipal Balance:** difference between the town council revenues and its expenses.

C2 – Tables

Table C1 – Correspondence Between NACE Classifications

NACE two-digit codes	NACE string codes	Label
[01; 03]	A	Agriculture, forestry and fishing.
[05; 39]	B-E	Mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply; water supply.
[41; 43]	F	Construction.
[45; 47]	G	Wholesale retail.
[49; 53]	H	Transportation and storage.
[55; 56]	I	Accommodation and food service activities.
[58; 63]	J	Information and communication.
[64; 66]	K	Financial and insurance activities.
[68; 82]	L-N	Real estate activities; professional, scientific and technical activities; administrative and support service activities.
84	O	Public administration and defence, compulsory social security.
85	P	Education.
[86; 88]	Q	Human health and social work activities.
[90; 99]	R-U	Arts, entertainment, and recreation; other service activities; activities as household as employer; activities of extraterritorial organizations and bodies.

Table C2 – Summary Statistics for Key Economic Outcomes

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
Economic Activity | (1995-2020)					
Real GDP *per capita*	25	14,728.50	3,239.86	9,857.29	23,619.38
Education | (2009-2020)					
Retention and desistance rate in primary and lower secondary education (%)	25	6.89	1.68	3.91	10.76
Retention and desistance rate in upper secondary education (%)	23	15.69	2.08	12.19	20.33
Inequality | (2015-2020)					
P80/P20 ratio for TH	25	3.44	0.27	2.92	4.18

(continues in the next page)

Table C2 – Summary Statistics for Key Economic Outcomes *(continued)*

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
P80/P20 ratio for TP	25	3.05	0.26	2.53	3.58
P90/P10 ratio for TH	25	7.72	0.91	6.47	10.02
P90/P10 ratio for TP	25	6.41	0.95	5.15	8.17
Gini coefficient for TH (%)	25	44.00	1.89	41.42	48.92
Gini coefficient for TP (%)	25	40.02	2.08	37.18	44.20
Social Capital | (1987-2020)					
Voter turnout in Chamber of Deputies elections	25	0.61	0.05	0.48	0.68
Voter turnout in European Parliament elections	25	0.41	0.04	0.30	0.47

Notes: TH and TP stand for taxable household and taxable person, respectively. Time periods used to compute the averages are in parentheses.

Table C3 – Summary Statistics for Other Economic Outcomes

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
Labour Market Outcomes | (2009-2020)					
Registered unemployment – % of resident population between 15 and 64 years old	25	5.54	1.17	4.12	8.52
Registered youth unemployment – % *per* 100 inhabitants aged between 25 and 34 years old	25	8.85	1.69	6.39	12.56
Registered unemployment – % *per* 100 inhabitants 15 years old and older	25	9.11	6.84	5.43	39.80
Employment – % of resident population	25	44.19	3.33	37.87	49.78
Trade Openness | (2009-2020)					
Imports/GDP	25	0.20	0.13	0.03	0.49
Exports/GDP	25	0.25	0.15	0.02	0.59

Notes: TH and TP stand for taxable household and taxable person, respectively. Time periods used to compute the averages are in parentheses.

Table C4 – Summary Statistics for Other Socio-Political Outcomes

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
Life Expectancy | (2008-2020)					
Life expectancy at birth	25	79.84	1.04	77.01	81.18
Life expectancy at 65 years	25	19.16	0.73	16.84	19.92
Suicide rates (proportion of deaths by different causes) | (2009-2020)					
Suicide rate	25	1.52	1.74	0.50	7.63
Crime Rates (*per* 100,000 residents) | (2015-2020)					
Total crimes	25	2,849.11	571.58	2,180.11	4,759.72
Crimes against people	25	797.10	173.71	578.21	1,413.31
Crimes against property	25	1,344.51	397.16	883.13	2,695.63

(continues in the next page)

Table C4 – Summary Statistics for Other Socio-Political Outcomes *(continued)*

Variables	Obs.	Mean	Std. Dev.	Min.	Max.
Crimes against life in society	25	456.09	133.22	303.42	960.73
Crimes against the State	25	50.47	17.14	30.16	92.53
Crimes against cultural identity	20	0.58	0.66	0.00	1.94
Crimes against pets	25	18.46	6.82	8.58	32.88
Crimes against separate legislation and others	25	182.09	56.68	95.77	299.04
Domestic violence against a spouse or equivalent related crimes	25	215.12	42.78	159.58	342.95
Theft in motor vehicle	25	139.63	85.24	62.14	416.29
Burglary in residence	25	118.49	45.93	64.64	288.30
Burglary in commercial or industrial building	25	70.11	22.93	28.97	109.18
Crimes of voluntary manslaughter	25	206.67	115.57	97.23	592.12
Crimes of assault	25	505.11	105.94	386.70	869.65
Bodily harm	25	221.18	54.36	155.81	419.88
Theft/purse snatching in public places	25	43.95	40.75	15.53	202.52
Theft of motor vehicles and theft from motor vehicles	25	206.67	115.57	97.23	592.12
Driving a motor vehicle with a blood alcohol level equal to or above 1.2g/l	25	177.56	44.75	114.48	282.41
Driving without legal documentation crimes	25	80.72	24.32	43.36	126.78
Public Sector Activity | (2010-2020)					
Municipal Balance *per resident*	25	3.61	1.59	-0.22	6.23
Municipal Balance *per resident* growth rate	25	-1.47	4.22	-17.56	2.58

Notes: TH and TP stand for taxable household and taxable person, respectively. Time periods used to compute the averages are in parentheses.

Table C5 – Pearson Correlations Between Income Inequality Measures and Unemployment Rates

	P80/P20 ratio (TH)	P80/P20 ratio (TP)	P90/P10 ratio (TH)	P90/P10 ratio (TP)	Gini index (TH)	Gini index (TP)
P80/P20 ratio (TH)	100					
P80/P20 ratio (TP)	95.85*	100				
P90/P10 ratio (TH)	93.95*	93.68*	100			
P90/P10 ratio (TP)	82.84*	89.02*	88.98*	100		
Gini index (TH)	93.79*	90.77*	96.25*	80.47*	100	
Gini index (TP)	90.08*	93.34*	95.25*	93.52*	94.59*	100
Unemployment: 15-64 years old (%)	12.79	25.66	29.37	42.96*	21.25	3567*
Unemployment: 25-34 years old *per* 100 inhabitants	22.77	36.92*	29.68	36.04*	21.17	30.24
Unemployment: more than 15 years old *per* 100 inhabitants	14.99	32.88	35.15*	38.85*	24.64	35.05*

Notes: Correlations are expressed in %. * stands for statistically significant at 10% level. TH and TP stands for taxable household and taxable person, respectively.

Table C6 – Predicted Minimum Values for The Nonlinear Relationship Between Persistence and Inequality

Variables	(1)	(2)	(3)	(4)
Inequality | (2015-2020)				
P90/P10 ratio for TH	NA	NA	8.09	8.08
P90/P10 ratio for TP	7.04	7.08	7.00	7.00
Gini coefficient for TH (%)	NA	45.64	51.52	44.88
Gini coefficient for TP (%)	NA	NA	36.29	40.33
Controls				
Real GDP *per capita*	No	Yes	No	Yes
Migration Flows	No	No	Yes	Yes

Notes: TH and TP stand for taxable household and taxable person, respectively. NA stands for Not Applicable.

C3 – Figures

Figure C.1 – Box Plots for Inequality Measures

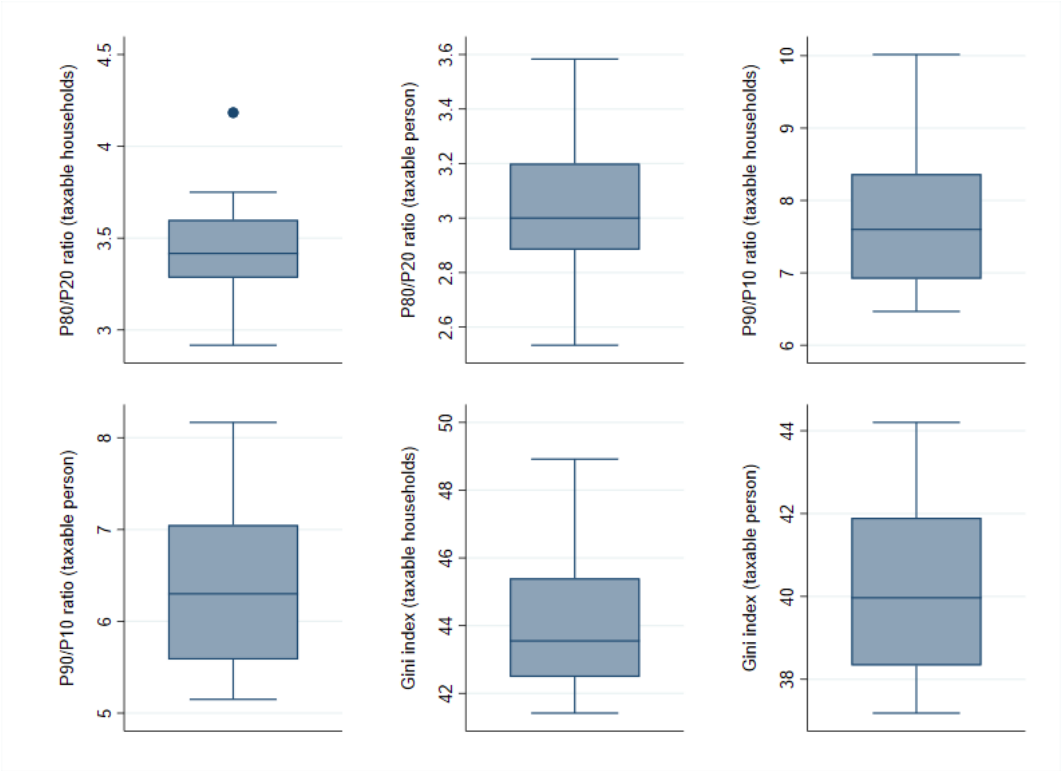


Figure C.2 – Box Plot for Regional ICS30