Universitat Autònoma de Barcelona

Facultat d'Economia i Empresa UAB

Bachelor's Degree Final Project

Faculty of Economics and Business

TITLE: Income Mobility in Spain: a look at socioeconomic position changes in the short-term

AUTHOR: *Adam Olivares Canal*

TUTOR: *Jordi Caballé Vilella*

DEGREE: Bachelor's Degree in Economics - English

DATE: *30/05/2023*

Abstract

I outline income mobility in Spain at the autonomic and provincial levels using pre-tax income collected through surveys and administrative files published by the INE. First, I build 2 repeated cross-sectional datasets, one for provinces between 2015-2020 and another one for 2007-2021, each containing the income distribution corresponding to each year. Second, I use panel-free upward mobility and relative mobility measures to characterize income mobility across different regions. The results suggest that income mobility differs across geographical units. Finally, I will consider a set of variables separated into four types of indicators: (i) Inequality, (ii) Immigration, (iii) Education, and (iv) health. These indicators will be used to model the causal dynamics that impact income mobility variation across autonomous communities (CCAA) with a principal components regression model and across provinces with a stepwise regression with both direction search.

Keywords: Income distribution, Income mobility, Upward mobility, Relative mobility, Autonomous communities, Provinces, Pre-tax income, Panel-free.

Contents

1 Introduction

Income mobility is defined as the capacity of individuals, households, or groups to change their economic position in the income distribution. It has been a hot topic in recent years, drawing the public's and researchers' attention because it provides information on the equality of opportunity and chance for the economic success of citizens in society, which can be used to identify its drivers and deal with them directly. Findings suggest income mobility differs across country regions and co-moves with economic performance, social and educational factors (Chetty et al., 2014; Güell et al., 2018).

Our objective is to replicate for Spain the analysis carried out by Chetty et al. (2014), where they demonstrated that intergenerational mobility differed across regions in the US by using the rank-rank specification put forward by Dahl et al. (2008) on the joint distribution of parents and child incomes [1]. They additionally demonstrated that their measure is correlated with factors such as segregation, race, family structure, school quality, inequality or social capital. In order to fulfill this objective, I follow an unprecedented methodology presented by Genicot and Ray (2022) to identify the upward mobility of individuals over their life and its variation across regions, which allows for pro-poor and panel-free data measurement of upward mobility by looking at the income distribution at two different points in time. An important implication is that such an approach suppresses the need for tracking the income of the same individual over time.

I outline Spain's income mobility on two territorial disaggregation levels, using repeated cross-section (semi-longitudinal) pre-tax income data from two databases elaborated by the INE to reproduce income distributions. These two databases are:

- "Encuesta de condiciones de vida" (ECV or EU-SILC) (INE, 2023b). The first belongs to the family of harmonized statistical operations for the UE countries that collect information on income distribution and social exclusion across European regions through a combination of anonymized surveys and exploitation of administrative files. It has been carried out for Spain since 2004, but because of a methodological change in income estimation, I will only create our dataset with data from 2008-2022 (14 years).
- "Atlas de Distribuición de la Renta de los Hogares" (ADRH) (INE, 2022a). The second is a project initiated in 2019 that draws solely on administrative files to construct a series of aggregated income variables and indicators, focusing on detailed disaggregation by territory. The available measures range from 2015 to 2020 (5 years), when the last update was received.

In the first part, I will introduce our methodology, carry out a data treatment to obtain our variables and implement the approach between 2007-2021 for autonomous communities using ECV microdata (CCAA are first level of political and administrative division in Spain) and 2015-2020 for provinces using ADRH data (a more minor administrative division). The objective is to detect if a meaningful variation in upward mobility exists conditional on a geographic factor.

Afterward, I explore the degree to which a few of the factors discussed in the literature may be relevant to explain upward mobility variation across Spanish regions. The first factor is inequality: I relate upward mobility with the mean log deviation by constructing a "Great Gatsby" curve. Second, Immigration: here, I will explore if there is a connection between your

[1] The joint distribution of parent and child incomes comprises the copula (the joint distribution of parents and child ranks) and the marginal distributions of both generations.

country of origin and the probability of climbing the income ladder. Third, Education: I will use the average class size of all pre-university levels as a public expense indicator. Finally, Health: a health professionals ratio will be used as a proxy to healthcare quality to determine the relationship.

2 Methodology

Examining income mobility involves analyzing how individuals' economic outcomes are influenced by their parents' income or social status. Upward mobility and relative mobility are two distinct concepts used to analyze intergenerational income or social mobility. Albeit they both relate to the movement of individuals across social classes, they center their attention on different aspects of mobility. The former refers to where a person ranks in the income distribution compared to others in the same geographical unit. In contrast, the latter characterizes the odds of an individual attaining a higher income than themselves in the past or their parents.

Various scholars have proposed different methodologies and indicators to assess and measure upward and relative income mobility. One widespread method relies on the construction of a transition matrix showing the probability of a child from a particular income quantile ending up in various income quantiles as an adult. These matrices provide a detailed picture of mobility patterns across income levels. However, if we wanted to compare income mobility results across regions in a simpler manner, we could summarize the information presented in the transition matrix in a single number at the expense of some loss of information. Other single-number measures are not reliant on transition matrices, like rank-rank slope or Intergenerational elasticity, which measures the percentage change in a child's income associated with a percentage change in parental income (a higher elasticity indicates lower mobility) (Caballe, 2016).

Although both ECV and ADRH databases are organized in such a way that allows for the formation of panel data and implement one of the aforementioned techniques, our analysis would benefit from removing the identity link of individuals between the starting and final income distributions. The reason is that some individuals' observations are not available intertemporally but in only one time period, which would force the omission of a substantial number of observations. Hence, in this research, I use the methodology recently developed by Genicot and Ray (2022), which allows for a simple panel-free implementation to measure upward mobility. To give some insights into its working mechanism, I will enunciate the 2 fundamental concepts they devised to connect growth to mobility from the already stated concepts of relative and absolute upward mobility: Growth-orientation (i), rewarding growth while punishing decay, and the Growth alignment axiom (ii), rewarding income growth rates transfers from the rich to the poor, which is equivalent to rewarding higher annualized growth rates of the poorest individuals relative to the richest. First, they build a preliminary "instantaneous" measure that precedes their discrete measure. It depends on the collection of pairs z_i of baseline incomes y_i, and instantaneous growth rates g_i of that income expressed as:

$$M_\alpha(\mathbf{z}) = \frac{\sum_{i=1}^{n} y_i^{-\alpha} g_i}{\sum_{i=1}^{n} y_i^{-\alpha}}, \text{ for some } \alpha > 0 \tag{1}$$

Equation 1 cannot be used in the current form, so they used it only as a base for a discrete upward mobility index. They state that for a measure to have an empirical application, it needs to be able to deal with the following three considerations: Data is discrete in time (observations are separated in points in time), individual income paths in time can cross, and accessibility to panel data since sometimes only cross-sectional data repeated in time exist. The combination of treatment for such concepts with their instantaneous upward mobility notion leads to a panel-free discrete-time empirical implementation that becomes both growth-oriented thanks to path independence. The fact that it is growth oriented implies that, unlike other measures found in the literature, it does not reward movement in any direction; it only quantifies upward movement by penalizing negative growth. Furthermore, the panel-free feature lets us exploit income data observed at two points without requiring reporting the income of the same individual across time; only an initial income distribution at time s and a final one at time t are required. This disassociation is possible because they found that individuals' identity links between the starting and final income distribution can be broken, enabling us to work with quantile data. This last property differentiates this measure from all the others seen in the literature. This is especially interesting to expand the existing body of literature on Spanish income mobility, which suffers from the availability of panel data particularizing the income of the same individual in more than one time period (Cervini-Plá, 2015).

The above-described expression for upward mobility is analogous to a growth rate, implying it can take positive or negative values and be expressed as a percentage. Let's denote n as the population of size, where each individual i has a baseline income $y_i > 0$ collected in a vector of incomes $\mathbf{y}$. Since we will work with percentile data from two income distributions measured at s and t, $y_i(s)$ and $y_i(t)$ will become the baseline incomes of a quantile i and n will be equal to the number of quantiles used to cut the distributions. Genicot & Ray's measure also evaluates economic mobility as the sum of individual growth rates weighted by economic characteristics, with a pro-poorness factor α that puts more weight on the growth of the poorest quantiles as it increases. The formula is presented as follows,

$$M_\alpha^\Delta(\mathbf{y}(s), \mathbf{y}(t)) = \frac{1}{t-s} \ln \left[\frac{\sum_{i=1}^n y_i^{-\alpha}(t)}{\sum_{i=1}^n y_i^{-\alpha}(s)} \right]^{-\frac{1}{\alpha}} \text{ for some } \alpha > 0. \tag{2}$$

Where $\mathbf{y}(s)$ and $\mathbf{y}(s)$ stand for individual (or quantile) income vectors in the initial and final periods s and t. Everything is divided by the normalization term, allowing us to interpret upward mobility as the overall upward movement throughout the period selected. A particularity of equation (2) is that it will become Rawlsian as $\alpha \to \infty$ and converge to the log growth of individual or quantile income growth rates $\alpha \to 0$.

Genicot and Ray (2022) recover the relative part of the index and omit the absolute aspect by netting out aggregate growth. Positive values will indicate that upward mobility exceeds average income growth, while negative values will be a symptom of the opposite:

$$K_\alpha^\Delta(\mathbf{y}(s), \mathbf{y}(t)) = M_\alpha^\Delta(\mathbf{y}(s), \mathbf{y}(t)) - \frac{1}{t-s}[\ln(\bar{y}(t)) - \ln(\bar{y}(s))] \tag{3}$$

This way, they get a relative upward mobility measure that evaluates the departure of upward mobility from the annualized average income growth, where $\bar{y}(t)$ refers to the average per capita income.

Intending to observe how the measure encapsulated in expression (2) compares to other empirical approaches in long-run contexts, they used the results and data presented in Chetty et al. (2017) for birth cohorts from 1940 to 1984 in the US, where they used a measure more commonly seen in the context of absolute intergenerational income mobility to document a decline in absolute mobility: the share of families whose absolute fortune has improved across generations. Genicot & Ray documented their upward mobility measure when $\alpha = 0.5$ for 30-year intervals only differed in magnitude but captured the same decline and co-moved closely with the panel-dependent intergenerational mobility in Chetty et al. (2014) and the Berman (2022) non-panel data approximation to Chetty $et.al$'s measure that only relies on copulas for other countries or periods. Hence, they point out that their upward mobility measure is a solid and reliable alternative in data-poor settings that may not have available income panel data.

3 Data

This section explains how income distributions are built for autonomous communities and provinces and the set of variables I use to study the differences in Upward mobility across autonomous communities and provinces. Variables for studying differences in upward mobility will be cross-sectional and not expressed as change over time, since we do not expect them to change substantially in the short-term. For all of them, I use data from the earliest year I have available income data, i.e., 2007 for CCAA and 2015 for provinces.

Table 1: List of independent variables classified by type.

Health	Healthcare workers ratio
Education	E.Infantil
	E.Especial
	Prog.GarantíaSocial
	E.Primaria
	C.F.G.M.
	C.F.G.S.
	E.S.O.
	Bachillerato
Immigration	UE28 without Spain
	Non-EU28 Europe
	Africa
	North America
	Center America and Caribbean
	South America
	Asia
	Oceania

As for income distributions, I prefer to work with pre-tax per capita income instead of post-tax income. The reason is that income redistribution offsets part of the inequality in the economy, so pre-tax income will provide a better picture of the actual inequality before any intervention is made.

3.1 Pre-tax income per capita for autonomous communities

I use de ECV database to obtain microdata for 2008 and 2022, containing data for the year before the survey was conducted, 2007 and 2021. Microdata is divided into 5 cross-sectional files for each year, but only 3 contain relevant information to pinpoint personal income. These are:

- Detailed adults' data ($\geq$ 16-year-old) (File P): It includes pre-tax and post-tax personal income data, disaggregated by sources from which it is obtained.

- Basic household data (File D): It assigns a household identification code to each unit and links it with the country, CCAA (expressed in NUTS classification), and anonymized census section where it is located.

- Detailed household data (File H): It includes pre-tax and post-tax household income data, disaggregated by sources from which it is obtained. Some sources, such as investments and rental property, are only included per household.

First, I need to identify where each individual lives. In file P, each adult receives an individual identification (variable PB030), composed of the household identification code plus two digits for each household member at the end. If I remove the last two digits, I am left with the same code as in files D and H (variables HB030 and DB030). Using file D, I will identify where that individual is living. Since census sections are anonymized and therefore unusable, autonomous communities become the smallest geographical unit available. Then I sum all the relevant sources of pre-tax income of each individual to obtain their total income before taxes. From file P, I will use the following sources of pre-tax income: monetary or quasi-monetary income of the salaried worker (PY010G), social security contributions made by the employer (PY030G), monetary benefits or losses of sole proprietorship workers (PY050G), the income coming from private pension schemes (PY080G), unemployment benefits (PY090G), retirement pension (PY100G), survivor's benefits (PY110G), sick pay (PY120G), disability payment (PY130G) and study grants (PY140G). However, pre-tax income coming from rental property (HY040G), child/family benefits (HY050G), social assistant grants (HY060G), housing benefits (HY070G) and return on investments (HY090G) is still missing and only available for households in file H. Therefore, I will impute those values by using an equal-split approach, dividing the amount of each household by the number of its members and sharing the result equally among them. Our total pre-tax income excludes any earned income not reported to the AEAT and non-monetary income of the salaried worker, like food allowances. If individuals report no income information, their per capita income will be treated as 0, not as a missing value[2].

As a result, I obtain two data frames with per capita incomes for 19 CCAA, one for 2007 (30082 observations) and another for 2021 (50147 observations). The number of observations is enough to form as many income distributions as autonomous communities each year. Before continuing, I must deal with outliers in the form of negative, zero, and extremely high incomes (see figure A.1). Our preferred approach is to truncate the lowest 0.5% to deal with extreme incomes of each distribution and then set all the remaining values equal or below €1 to an arbitrarily minimum income, say €1.1. Incomes equal to €1 translates into an insensitivity of the bottom percentiles of the income distribution in the pro-poorness coefficient.

In contrast, values below €1 make our upward mobility measure dramatically sensitive to the choice of pro-poorness. This treatment will be sufficient to include low-income percentiles in our upward mobility formula without censoring them too much. Notice that the pro-poorness coefficient, as stated in Genicot and Ray (2022), is sensitive to these imputations, especially for large values of α. Their preferred treatment is aggregating data into deciles instead of percentiles. However, this solution is insufficient here as I would still get zero-value incomes in some autonomous communities like Melilla, making expression (2) and (3) produce NaN values. Even if these CCAA were omitted, the difference in decile growth rates would be so disparate that our mobility index would become inconsistent, generating an almost completely different ranking of autonomous communities for each pro-poorness. Despite the corrections, one critical remark is that the obtained income distributions still suffer from a skewness to

[2] In fact, I found a 6.44% and 7.97% in 2007 and 2021 respectively of observations equal to 0.

the right, a long right tail, and a high concentration of low incomes, which, in turn, makes our upward mobility measure sensitive to the choice of pro-poorness, but not as sensitive as if I had chosen to work with decile data because, in this case, it benefits from a higher income distribution definition.

Once I implement the corrections mentioned above to our distribution to make it operational, I compute the percentiles and average per capita income of each autonomous community and fit them into expressions (2) and (3).

3.2 Pre-tax income per capita for provinces

The ADRH database includes income and income distribution indicators between 2015-2020 for different degrees of territorial disaggregation, being census sections the greatest level of detail. I will take advantage of this level of detail to use each census section's average pre-tax per capita income in 2015 and 2020 to form a provincial income distribution. Census sections are partitions of municipal areas that possess easily identifiable limits and include around 1000–2500 inhabitants. The main flaw of this approach is that average pre-tax income is not a substitute for individualized data. Thus, I must rely upon the assumption that there's just one representative individual per census section who earns the mean pre-tax income[3]. Such an implication precludes the possibility of accounting for individuals with zero or even negative incomes and narrows the range of incomes by censoring the poorest and richest individuals.

Besides, I will plot the proportion of missing observations of each province relative to the total number of census sections. As Figure A.2 shows, the census section's data is unavailable in the same proportion for both years. Provinces 01 (Álava), 20 (Gipuzkoa), and 31 (Navarra) must be dropped from our analysis due to this constraint; there are not enough observations in 2015 to form an income distribution. [4] After this exclusion, I will create a data frame with the percentiles for all provincial distributions and others with average per capita incomes. The number of observations (census sections without missing data) is 33565 for 2015 and 33524 for 2020.

3.3 Demographic data: percentage of immigrant population

The migration statistics (INE, 2023a) provides semiannual migration estimations broken down by different indicators, such as sex, year of birth, nationality, and country of birth. I will collect data from January 1st, 2008, and January 1st, 2016; the reason is that I want a snapshot of demographics as close as possible to 31st December 2007 and 2016 because later on, our health professionals datasets will be using that date. From these estimations, I will calculate the percentage of immigrants by country of birth (country grouping) for each Spanish geographical unit, dividing the immigrant population of each grouping by the total resident population. The variables obtained are UE28 without Spain, Non-EU28 Europe, Africa, North America, Center America and Caribbean, South America, and Oceania.

[3] The INE also offers the number of inhabitants per census section, which could be used to create a data frame with as many rows as the number of inhabitants by assuming that all census section members receive the same pre-tax income. Nevertheless, this method is computationally expensive.

[4] Provinces 19 (Guadalajara) and 42 (Soria) lack information about more than 50% of the census section, so I must not lose sight of the fact that their income distributions will have a poorer definition.

3.4 Public education expenses

Public expenses in education are defined as the average class size of mandatory and optional non-university levels of education in the Spanish system. These proxies are obtained by collecting data on the number of classes and students from the Ministry of Education database (Ministerio de Educación y Formación Profesional, 2023) for the academic years 2007-2008 and 2015-2016. Private education centers' data will be omitted because they do not receive as much public funding as semi-private (concertada) and public schools. I do not make special distinctions for adult students and online classes; these will be added to the rest of the data and treated equally. Data about schools with mixed education, including classes from different levels, will have to be excluded because no further information about the number of classes discriminated by type is provided. The variables obtained are E.Infantil (pre-school education), E.Primaria (primary education), E.Especial (special education), Prog.Garantía Social (social guarantee programs. Not available for 2015), E.S.O. (secondary education), C.F.G.M. (middle grade), C.F.G.S. (equivalent to junior college) and Bachillerato (baccalaureate).

3.5 Healthcare quality

The healthcare quality ratio is the sum of working-age collegiate health professionals per 1000 inhabitants. This proxy is obtained by collecting data on the number of collegiate professionals and total population from the INE (INE, 2022b) for 2007 and 2015. In Spain, health professionals from diverse areas must be collegiate in the province where they exercise most of the year (BOE, 1974), so it is logical to assume that most will work where they are collegiate. Collegiate health professionals' statistics include physicians, nurses, dentists, dental technicians, pharmacists, physical therapists, clinical psychologists, speech therapists, optometrists, podiatrists, occupational therapists, and nutritionists. The name given to the result of this operation is Health professionals ratio.

4 Results: Income mobility and the "Great Gatsby" curve

The results section is divided into two main parts. First, I will show the values obtained for our upward mobility measure and discuss their robustness to other pro-poorness coefficient numbers. On the other hand, I will observe the relationship between mobility and inequality in our data through a Great Gatsby curve.

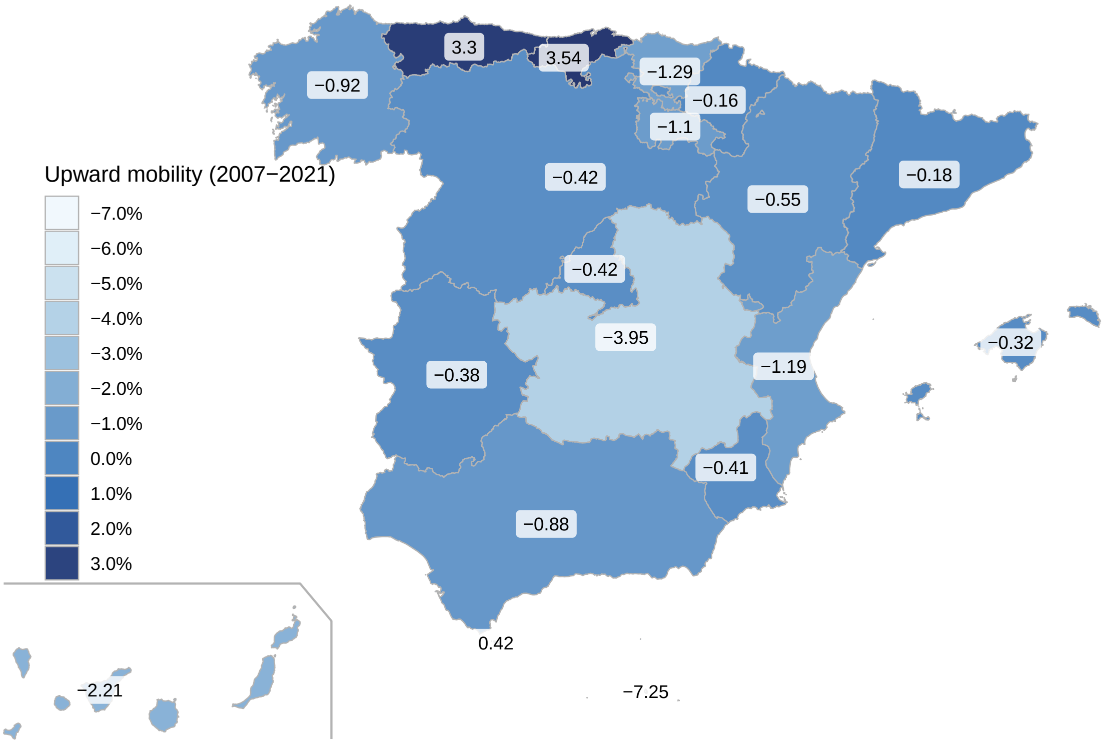
4.1 Upward Mobility and Relative Upward Mobility

I apply equations (2) and (3) for the pro-poorness value of $\alpha = 0.7$ using our percentile income data for 5-year (provinces) and 14-year (autonomous communities) periods to compute upward mobility and relative upward mobility. In Figure A.3, I test the choice of α trying its robustness for different values using the provincial data frame. The choice of α does not affect the ranking of most and least mobile regions to a great extent, but it does affect the magnitude, especially for autonomic data, which is considerably sensitive to the choice of pro-poornes, mainly due to the high number 0 income values transformed and included (as discussed earlier in subsection 3.1).

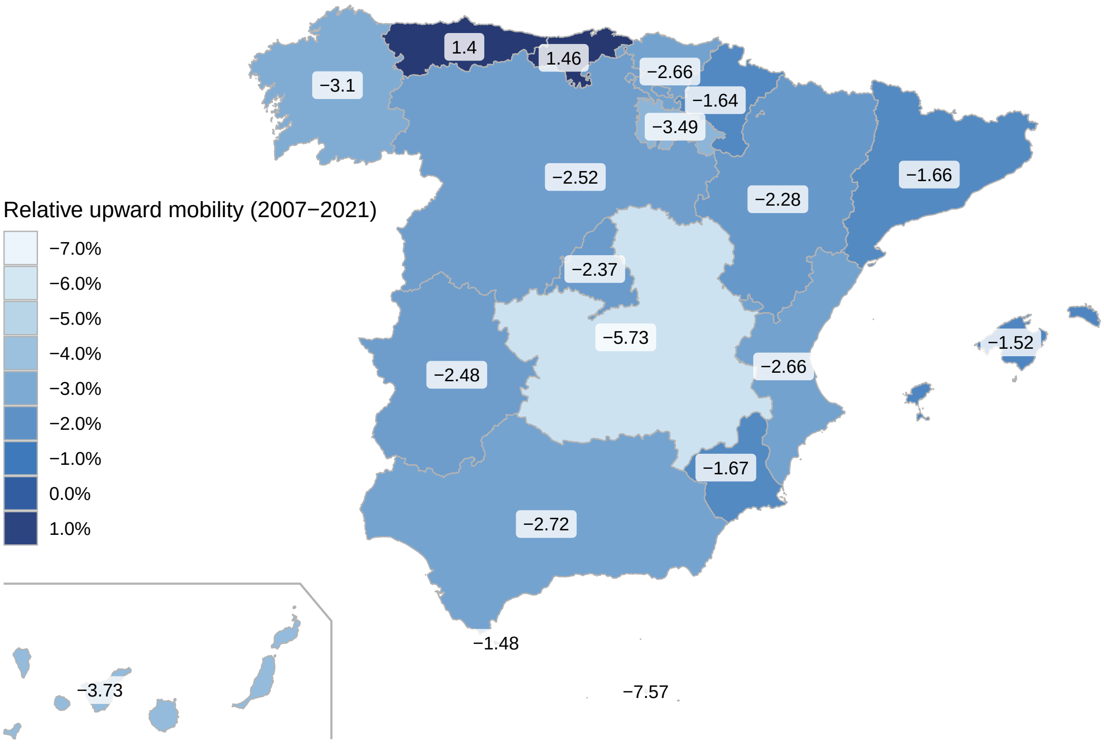
Panels (a) of Figure 1 and 2 presents upwards mobility, and panels (b) relative upward mobility. Focusing on panels (a), I can discern three notable patterns that apply to both 5-year provincial mobility and 14-year autonomic mobility: First, income mobility differs across and within autonomous communities. Whether it decreases or increases and the magnitude of the value may be conditioned by several factors throughout the periods I will discuss later on, but also by the level of geographical disaggregation and the length of the studied period. Moreover, the time availability and methodology used to collect and present income data from the ECV and ADRH databases differs considerably, which poses a problem concerning the comparability of results because it may also be driving this sign difference between provincial and autonomic results. Second, Castilla y la Mancha, Valencia and Andalucía are the least mobile regions of Spain. Third, Asturias, Cantabria are the most mobile regions, followed by Catalunya, and Illes Balears. Ceuta is also among the most mobile areas. Still, its case is particular because its upward mobility result is driven by divergent growth rates over percentiles when compared to other regions (In Figure A.4 (CCAA ES63) and A.5 (Province 51) this phenomenon is detected. All the other regions present varying relative positions in the ranking. To exemplify this, take Madrid as an example. For 2015-2020 (Figure 1), it exhibits one of the highest upward mobility increases of the period, but for 2007-2021 (Figure 2) is in a mid-position.

Maps (b) consider relative upward mobility, which measures the deviation of upward mobility from each region's annualized average per capita growth rate. The rankings according to this measure change slightly. Still, the takeaway here is average pre-tax income growth may be lower or higher than upward mobility depending on the province or autonomous community; the closer relative upward mobility to 0, the closer mobility and average income growth comove. In figure 1, only Asturias' and Cantabria's upward mobility grew more than the period's annualized average per capita income. At the same time, the rest of the autonomous communities experienced the converse phenomena, reaching values in every case. Excluding Ceuta, all the negative values in relative upward mobility are explained by the fall in upward mobility as income growth grew. In figure 2, a pattern can be distinguished; Galicia, Canarias, and all the southern provinces except Ceuta, obtain negative numbers. Nonetheless, there are more provinces in the north and the Illes Balears whose upward mobility value surpassed average per capita growth, albeit other northern regions also attain negative percentages.

If I take a closer look at income growth rates across percentiles (Figures A.4 and A.5), all the autonomous communities in Figure 1 that have seen a decrease in mobility for the period have at least one extreme negative growth rate for some percentile below 25 (except for Melilla, which also reaches extreme values, although less, but more percentiles have experiences income decreases than in any other CCAA), whereas those with positive upward mobility for the period only have positive income growth rates. Concerning the results presented in Figure 2, no percentiles experience a decrease in income, explaining why all provinces receive a positive value of upwards mobility. By observing these contrasting results and setting aside the methodological differences in the data collection between the utilized databases, I can hypothesize income mobility has declined as a whole in the period after the 2008-2014 financial crisis that struck the Spanish economy. Still, it recovered in the post-crisis period ranging from 2015 to 2020. Note that both final years for our upward mobility indexes, 2020 and 2021, are the years when the Covid-19 crisis and its economic consequences came about.

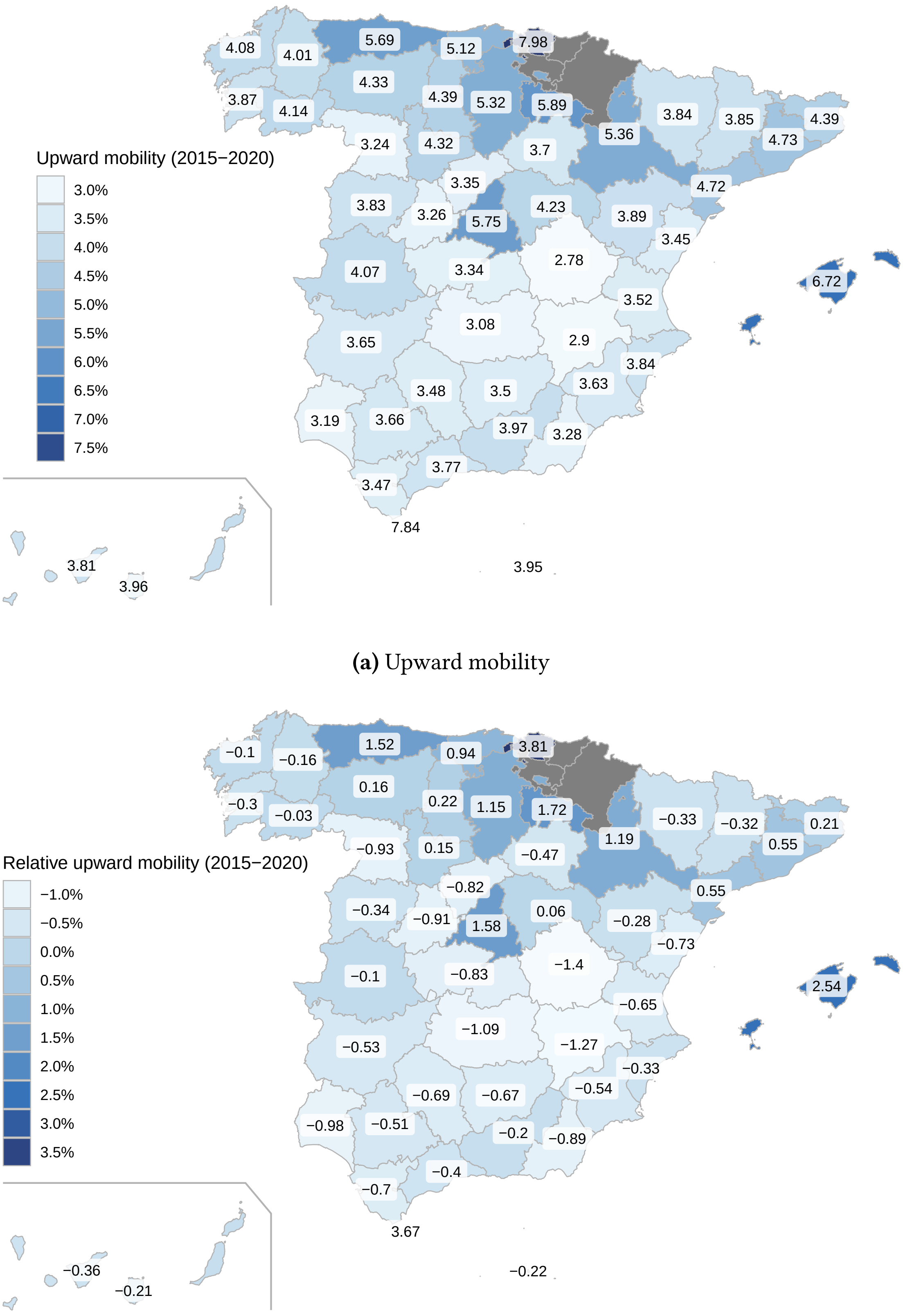


(a) Upward mobility



(b) Relative upward mobility

Figure 1: 14-year period heat maps of autonomic mobility for a 0.7 value of the pro-poorness coefficient.



(a) Upward mobility

(b) Relative upward mobility

Figure 2: 5-year period heat maps of provincial mobility for a 0.7 value of the pro-poorness coefficient.

4.2 Great Gatsby curve

A Great Gatsby curve characterizes the negative relationship between income mobility and income inequality. The concept was first introduced by Krueger (2012) and studied by Corak (2013). I explore whether this relationship holds over Spanish regions by regressing upward mobility on the mean log deviation of the earliest year. Mean log deviation is computed within each province and autonomous community using the mld.wtd command from the dineq package in R (Schulenberg, 2018). This measure reaches 0 when there is complete equality in our income data. Larger values correspond to higher inequality.

In Figure 3, I observe a weak negative relationship between inequality and mobility. Conversely, in Figure 4, a positive relationship between these variables is found. This difference may be due to the effect of working with different a different number of observations and databases. Besides, it might be that each province that constitutes Spain has a downward Great Gatsby curve, but when put in common in a single plane, they can form a curve depicting a positive relationship. Our figures should be viewed cautiously since MLD is not statistically significant to explain CCAA mobility, and R^2 indicates that only a minimal fraction of the variation in mobility is accounted for inequality in our data. Another remark is that inequality values significantly differ between provincial and CCAA income data. Recall that ADRH observations are measured as average pre-tax income per capita, which overlooks a substantial part of the differences between rich and poor by taking averages. In contrast, ECV observations contain abundant 0 or close to 0 income observations, thereby creating a much wider gap when compared to richer income values.

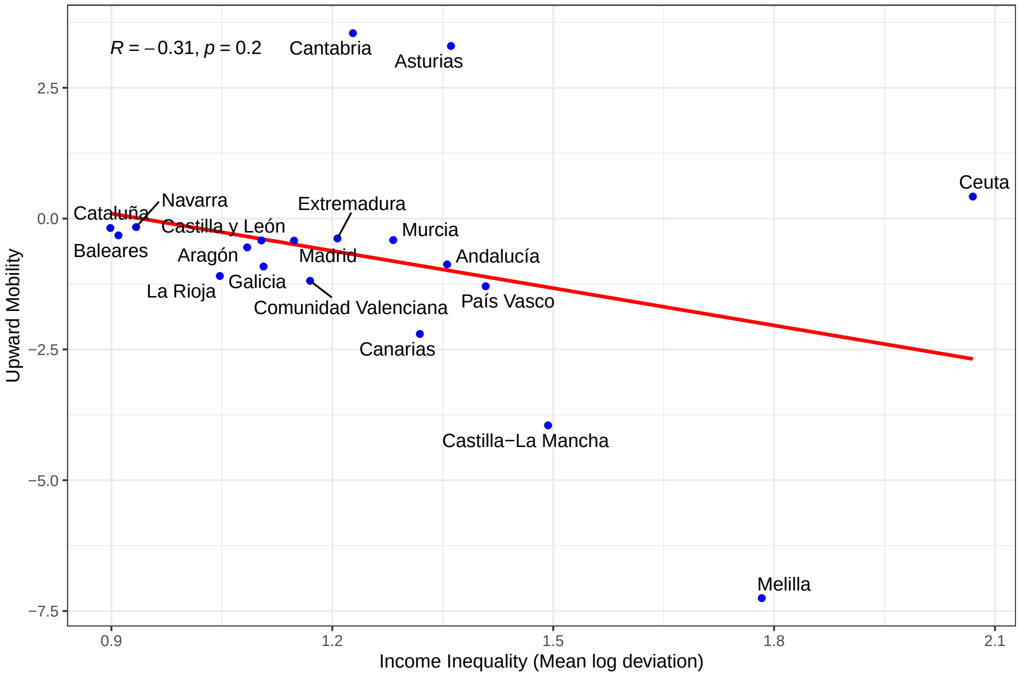


Figure 3: Linear regression performed on standardized CCAA data to obtain correlations as coefficients. UPWARD MOBILITY is plotted against inequality (MLD). This regression depicts a negative relationship between income mobility and inequality; it has a slope of -0.31 and an R-squared of 0.09337. 1 standard deviation increase in the mld index leads to a -0.31 decrease in upward mobility.

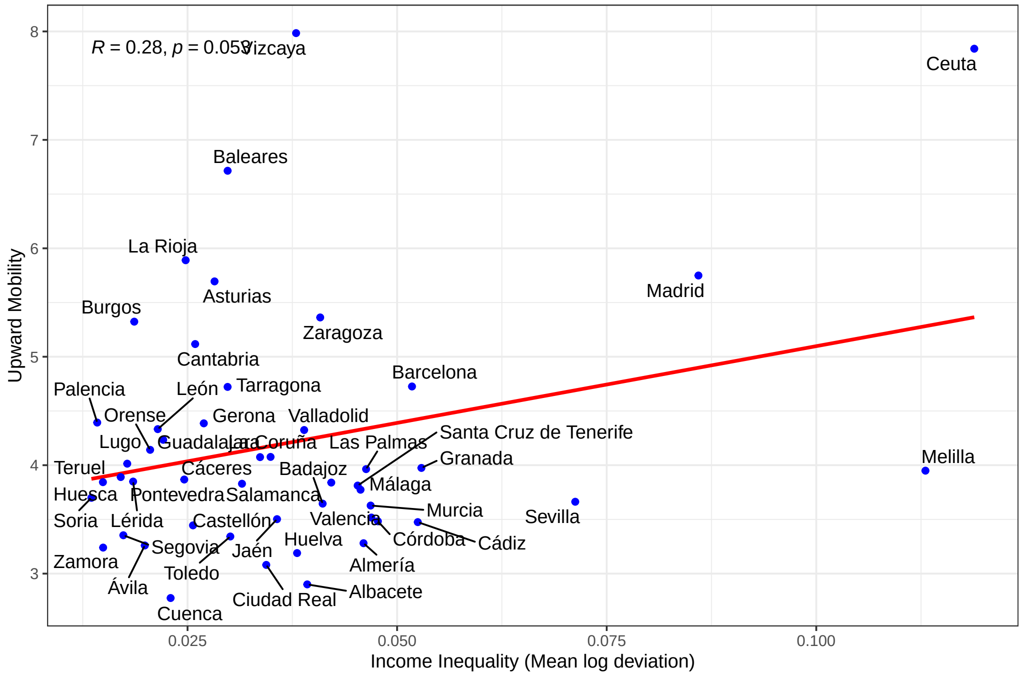


Figure 4: Linear regression performed on standardized provincial data to obtain correlations as coefficients. Upward mobility is plotted against inequality (mld). This regression depicts a positive relationship between income mobility and inequality; it has a slope of 0.28 and an R-squared of 0.07761. 1 standard deviation increase in the mld index leads to a 0.28 increase in upward mobility.

5 Factors explaining Upward Mobility

In this section, I will use two techniques to determine the geographical variation in upward mobility without making an arbitrary selection of variables. I build two multiple linear regression models for CCAA and provinces to explore these relationships. However, I am only interested in keeping variables with the most explanatory power, and, as a consequence, I must address a variable selection problem to construct parsimonious models. First, I use stepwise regression with both directions for the provincial dataset to produce a smaller set of variates with explanatory power in an OLS regression. The algorithm generates regression models through an iterative process of throwing out and adding variables based on AIC. On the other hand, CCAA data suffers from high dimensionality, as I have 17 independent variables and 19 observations. Using the stepwise algorithm only leads to an overfitted model with hard-to-interpret coefficients and high VIF (Variance inflation factor) in all variables, implying the presence of multicollinearity. One simple and reasonable solution is to reduce the number of independent variables by performing a Principal Components Regression using the 'psych' package in R (Revelle, 2023).

5.1 Factors explaining upward mobility variation across CCAA

I start from a set of 17 standardized independent variables and perform Principal components analysis to summarize the explained variance of our data set in a few variables. In this case, principal components will be obtained by conducting an eigenvalue decomposition of the covariance matrix[5] to rotate our data points such that variance is maximal on the first axis. As a result, I obtain 17 new variables called principal components. These orthogonal linear combinations of the original standardized variables can explain a portion of the data's variation.

The first components capture the most variation (see Table A.1), so only a few will be retained. A commonly used criterion combines a screeplot visualization[6] with Kaiser's rule[7].

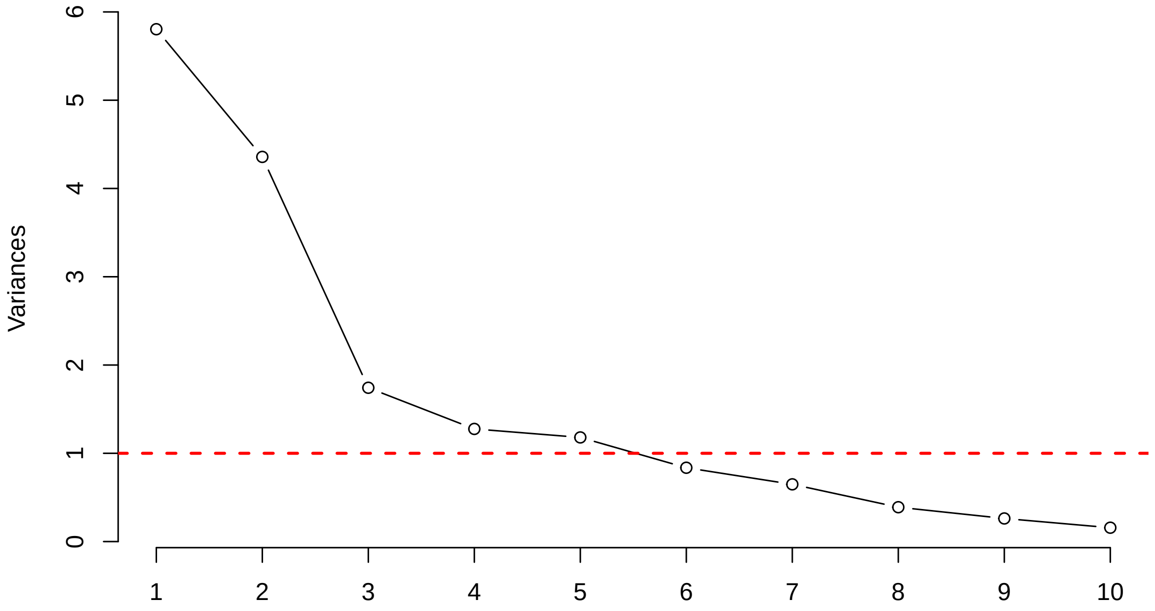


Figure 5: The screeplot shows each component's variance (eigenvalues).

Figure 5 suggests that between 2 and 5 components should be retained by applying the above-mentioned criteria. 3 components are a reasonable selection because a further number will cause a decay in model performance, and 2 would not let us effectively account for the impact of healthcare quality but would yield a better model in terms of adjusted R^2 (see table A.3 for performance comparison of regression models including 2-4 principal components). Moreover, the model interpretation becomes more complex as I introduce more components. This choice allows us to recover a 70% of the variation in the original set. Table A.2 shows the correlations between variables and components that will be used to give a logical interpretation of our components. This interpretation will change depending on the variables components are the most and least correlated with; the reason is those components assign scores to every autonomous community based on the loading matrix.

- PC1 can be read as a variable that summarizes education public expenses and the proportion of African population. CCAA that score low in PC1 will have higher average class sizes and percentage of African population. E.S.O., E.Primaria and Prog.Garantía Social are the education levels that contribute the most. CCAA with high scores will have the opposite relationship with African population and education.

[5] I work with the covariance matrix if our variables are expressed in different scales and I standardize them. Otherwise, spectral decomposition on the correlation matrix is preferable.

[6] To choose the number of components, I look at the point with the maximum curvature and retain all the components before that point. This test calls for a relative judgment of the variance accounted for by the subsequent components.

[7] Only components whose variance is greater than 1 should be retained. The reason is based on the idea that each component should account for at least as much variation as any of the original variables.

- PC2 discriminates between countries of origin. Regions with the least African population relative to other migration origins will score low, e.g., northern regions.
- As for PC3, Healthcare workers ratio is moderately correlated with this component, which tells us that areas with more investment in Health will score high in this component. Navarra and Catalunya are good examples. Even so, PC3 is also weakly correlated with other migratory and education variables that hinder a clearer explanation[8].

The next step is to regress upward mobility on the three first principal components to earn more about the causal relations between upward mobility and immigration, education, and health. A comparison of the signs obtained in the regression estimates with those of the PC scores will tell us all the necessary information to conclude. The results are as follows:

Table 2: Principal Components Regression results with standard errors in parentheses.

	Dependent variable:
	Upward mobility
PC1	0.570^{***}
	(0.183)
PC2	-0.277
	(0.211)
PC3	0.141
	(0.333)
Constant	-0.756^{*}
	(0.428)
Observations	19
R^2	0.437
Adjusted R^2	0.325
F Statistic	3.888^{**} (df = 3; 15)
Note:	*p<0.1; **p<0.05; ***p<0.01

First, our PC regression model establishes a negative relationship between the proportion of African population, average class size, and areas with higher income mobility. PC1 gets a positive coefficient; thus, I infer that areas with higher average class size and percentage of African population, namely low PC1 scores, tend to experience less income mobility. Variation across CCAA may also be explained by the immigrant population's country of origin; CCAA with a lower non-African immigrant population will achieve more mobility. Last of all, the ratio of health professionals positively impacts mobility, which is pointed to by the positive sign of PC3's coefficient, but the relationship is not statistically significant either, probably due to the small number of observations and the complex interactions of PC3 with other variables. Nevertheless, the overall regression model is valid. Its adjusted R^2 indicates our model can only explain a 32.5% of the variation in Upward Mobility, which leaves a great amount of unexplained variation by our set of variables. As discussed earlier in this section, I can gain a

[8] Regions with low E.Infantil, but high C.F.G.M. average class size and less non-Spanish European population, but more North Americans, will also score high in PC3. Although healthcare quality possesses the highest correlation, these interactions cannot be disregarded.

slight improvement in explanatory power if I exclude PC3 from the model, indicating it does not provide a meaningful amount of explanation for the variation in upward mobility.

5.2 Factors explaining upward mobility variation across provinces

To gain more understanding of income mobility across provinces, I let the stepwise algorithm select the best linear model in terms of AIC starting from our set of 16 independent variables. Stepwise regression yields the following result:

Table 3: Regression Results selected by the stepwise algorithm with standard errors in parentheses.

	Dependent variable:
	Upward mobility
Health professionals ratio	0.150^{***}
	(0.055)
E.Primaria	0.517^{***}
	(0.123)
E.S.O.	-0.221^{**}
	(0.100)
Non-EU28 Europe	-0.700^{**}
	(0.344)
Africa	-0.098^{*}
	(0.054)
South America	-0.262^{**}
	(0.128)
Asia	1.872^{***}
	(0.587)
Constant	-3.167^{*}
	(1.859)
Observations	49
R^2	0.508
Adjusted R^2	0.425
F Statistic	6.060^{***} (df = 7; 41)
Note:	*p<0.1; **p<0.05; ***p<0.01

By observing the regression model results in Figure 3, I can infer that our healthcare quality proxy is a relevant and positively related factor.

Regarding education, only two variables were considered relevant by the algorithm. However, although statistically significant, the effects of the education variables are harder to interpret. While higher average class size in secondary education (E.S.O.) exhibits a negative relationship, the same statistic but for primary education (E.Primaria) impacts positively. The problem is that higher class sizes are expected to be inversely related to per capita investment in education, thereby hampering upward income mobility. One hypothesis would be that higher class sizes in primary education translate into not allowing to segregate of students by performance, and this phenomenon facilitates upward mobility. Still, maybe it is just a misleading reaction consequence of studying mobility over a span of 5 years, which does not allow for evaluating mobility across generations. Considering that primary education is an early stage, it is hard to understand how it can affect upward mobility evaluated in the short term.

Finally, I am left with three migratory variables. The model shows a significant relationship between areas with lower upward mobility and higher concentration of immigrants from African and South American countries. The opposite interaction will be expected in areas with a higher population coming from Asia.

Altogether, the model can provide an explanation for a 42.5% of the variation in upward mobility. Other variables beyond the scope of our analysis may explain the remaining part.

6 Conclusion

Earlier, I presented an estimation of the indexes of upward mobility and relative upward mobility postulated by Genicot and Ray (2022) for two income databases, yielding a set of panel-free comparable measures across Spanish provinces and autonomous communities encompassing 5 years and 14 years, respectively. By virtue of its panel data independence, we can exploit the publicly available income data to the fullest by allowing us to work income distributions that only need to share the same level of territorial scope and removing one limiting condition: the same number of individuals observed in two different years, an otherwise determinant requirement for our analysis as a substantial number of ECV and ADRH observations fail to fulfill it.

Our analysis confirmed a contrast between upward mobility values of the studied geographical areas and that individuals do not stay in the same income percentile all their lives. The index co-move positively with healthcare quality and ambiguously with public expenses in education and the county of origin of the immigrant population, depending on the degree of territorial disaggregation and the number of years observed. More specifically, our two regression models support the idea that more mobile regions have more health professionals per 1000 inhabitants, Asian immigrants, and smaller E.S.O. classes. This opens the door for policymakers to improve economic opportunities by investing more in public health and education to increase the number of healthcare professionals and professors. However, less mobile areas will have a higher concentration of African population, denoting fewer economic opportunities to escape from poverty conditional on racial segregation toward African immigrants. Our findings additionally suggest that relative upward mobility differs across geographical units, implying a generalized deviation of average per capita income from upward mobility. On top of that, in most cases, saving Asturias and Cantabria, such deviation is driven by an underperformance of upward mobility relative to average per capita growth.

The main limitation of our results is that the modest amount of data and its presentation is a hindrance because neither 5 nor 14 years are enough to illustrate long-run mobility dynamics reliably, so what I am measuring instead is income mobility in the life of an individual in the short term. Another implication of this time problem is that it restricts the ability of the analyzed factors to explain the dynamics of income mobility in the long run, which is further worsened by the lack of more geographically disaggregated personal data (e.g., pre-tax personal incomes by source as in ECV, but linked to zip codes, census section or municipalities instead of autonomous communities). Geographical disaggregation, aside from its negative impact on upward mobility measurement accuracy and ”Great Gatsby” curves construction, becomes of utmost importance when building multifactor regression models since the absence of it outrightly translates into high dimensional data and thus less capacity to account for potentially relevant variables. Continuing this research with a longer time extension, at least 25 years, including other potentially impactful factors, such as marital status and geographically better-defined income data, could provide more insightful and accurate mobility estimates and regression models.

References

Bibliographic sources

Berman, Y. (2022). The long-run evolution of absolute intergenerational mobility. *American economic journal. Applied economics*, *14*, 61–83. https://doi.org/10.1257/app.20200631

BOE-A-1974-289 Ley 2/1974, de 13 de febrero, sobre Colegios Profesionales. (1974).

Caballe, J. (2016). Intergenerational mobility: Measurement and the role of borrowing constraints and inherited tastes. *SERIEs : journal of the Spanish Economic Association*, *7*, 393–420. https://doi.org/10.1007/s13209-016-0149-2

Cervini-Plá, M. (2015). Intergenerational earnings and income mobility in spain. *The Review of income and wealth*, *61*, 812–828. https://doi.org/10.1111/roiw.12130

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., & Narang, J. (2017). The fading american dream. *Science (American Association for the Advancement of Science)*, *356*, 398–406. https://doi.org/10.1126/science.aal4617

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly journal of economics*, *129*, 1553–1623. https://doi.org/10.1093/qje/qju022

Corak, M. (2013). Income inequality, equality of opportunity, and intergenerational mobility. *The Journal of economic perspectives*, *27*, 79–102. https://doi.org/10.1257/jep.27.3.79

Dahl, M., DeLeire, T., & of Wisconsin–Madison. Institute for Research on Poverty, U. (2008). *The association between children's earnings and fathers' lifetime earnings: Estimates using administrative data*. University of Wisconsin-Madison, Institute for Research on Poverty.

Genicot, G., & Ray, D. (2022). Measuring upward mobility. *NBER Working Paper Series*. https://doi.org/10.3386/w29796

Güell, M., Pellizzari, M., Pica, G., & Mora, J. V. R. (2018). Correlating social mobility and economic outcomes. *The Economic journal (London)*, *128*, F353–F403. https://doi.org/10.1111/ecoj.12599

Krueger, A. (2012). The rise and consequences of inequality in the united states. *The Center for American Progress.*

Databases and R packages

INE. (2022a). *Inebase / nivel y condiciones de vida (ipc) /condiciones de vida /atlas de distribución de renta de los hogares / resultados.* https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177088&menu=resultados&idp=1254735976608

INE. (2022b). *Inebase / sociedad /salud /estadística de profesionales sanitarios colegiados / resultados.* https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176781&menu=resultados&idp=1254735573175

INE. (2023a). *Inebase / demografía y población /fenómenos demográficos /estadística de migraciones / resultados.* https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177000&menu=resultados&idp=1254735573002

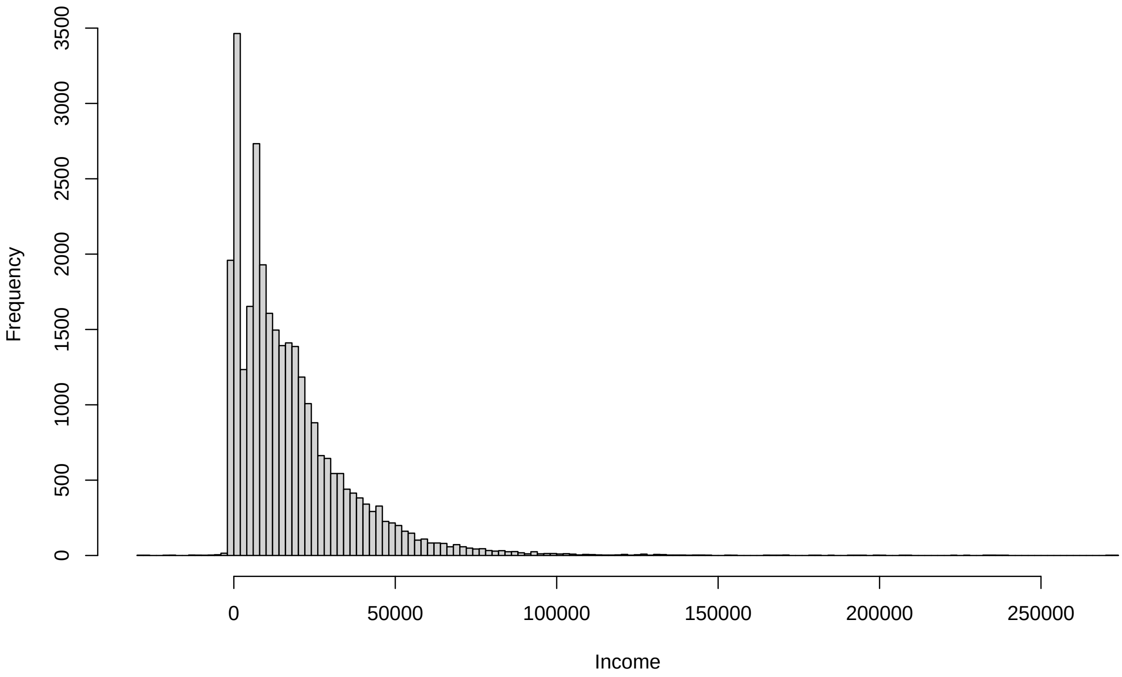
INE. (2023b). *Inebase / nivel y condiciones de vida (ipc) /condiciones de vida /encuesta de condiciones de vida / resultados.* https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=resultados&idp=1254735976608

Ministerio de Educación y Formación Profesional. (2023). *Enseñanzas no universitarias — ministerio de educación y formación profesional.* https://www.educacionyfp.gob.es/servicios-al-ciudadano/estadisticas/no-universitaria.html

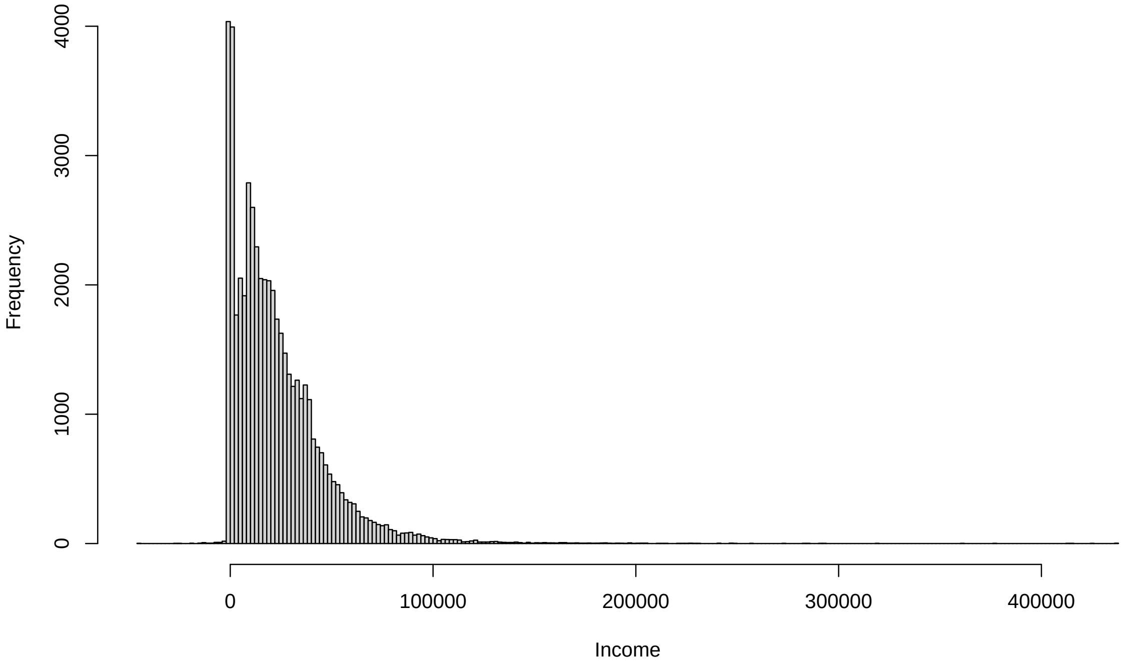
Revelle, W. (2023). *Procedures for psychological, psychometric, and personality research [r package psych version 2.3.3]*. https://CRAN.R-project.org/package=psych

Schulenberg, R. (2018). *'dineq': Decomposition of (income) inequality [r package version 0.1.0]*. https://CRAN.R-project.org/package=dineq

A Appendix



(a) Income distribution histogram of Spain in 2007.



(b) Income distribution histogram of Spain in 2021.

Figure A.1: Histograms of the Spanish pre-tax income distributions based on ECV data.

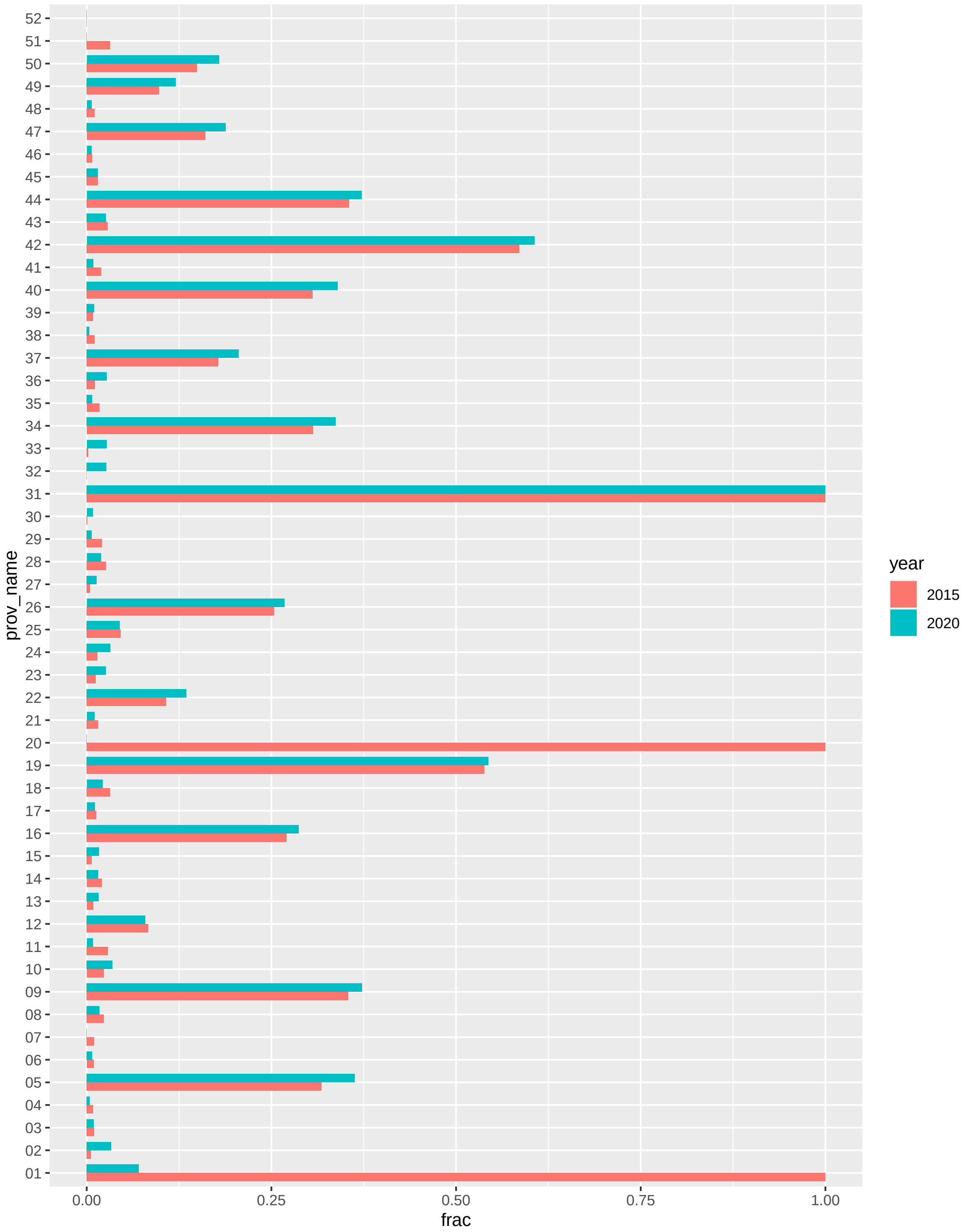
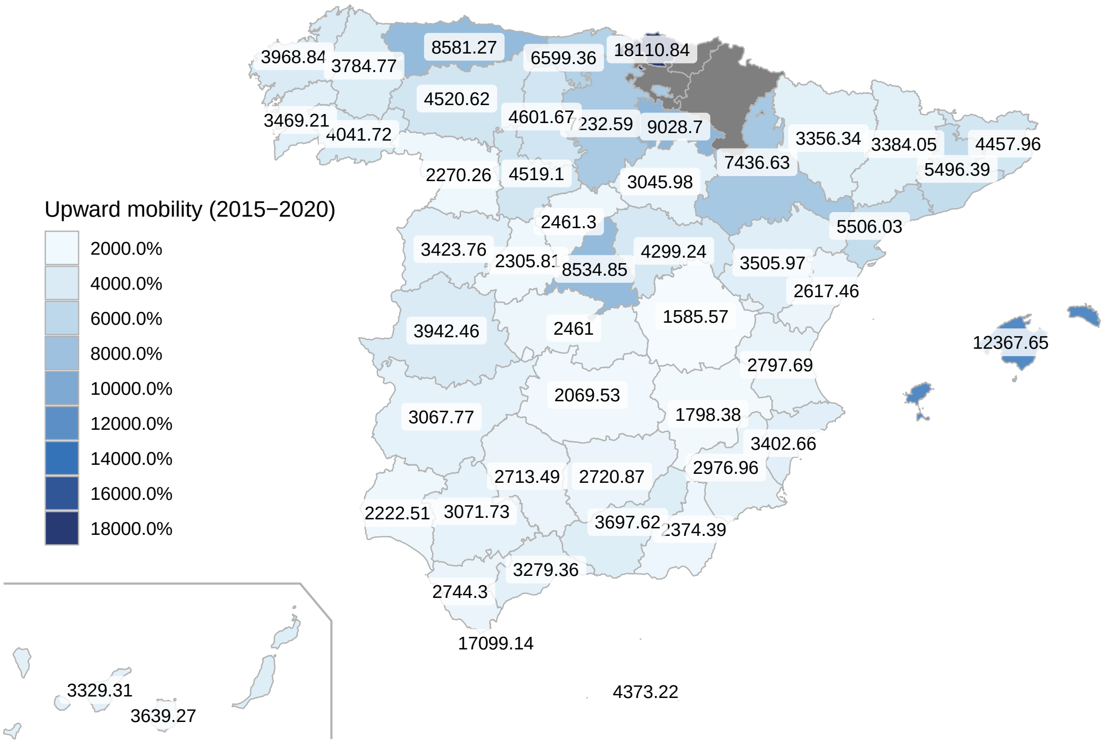
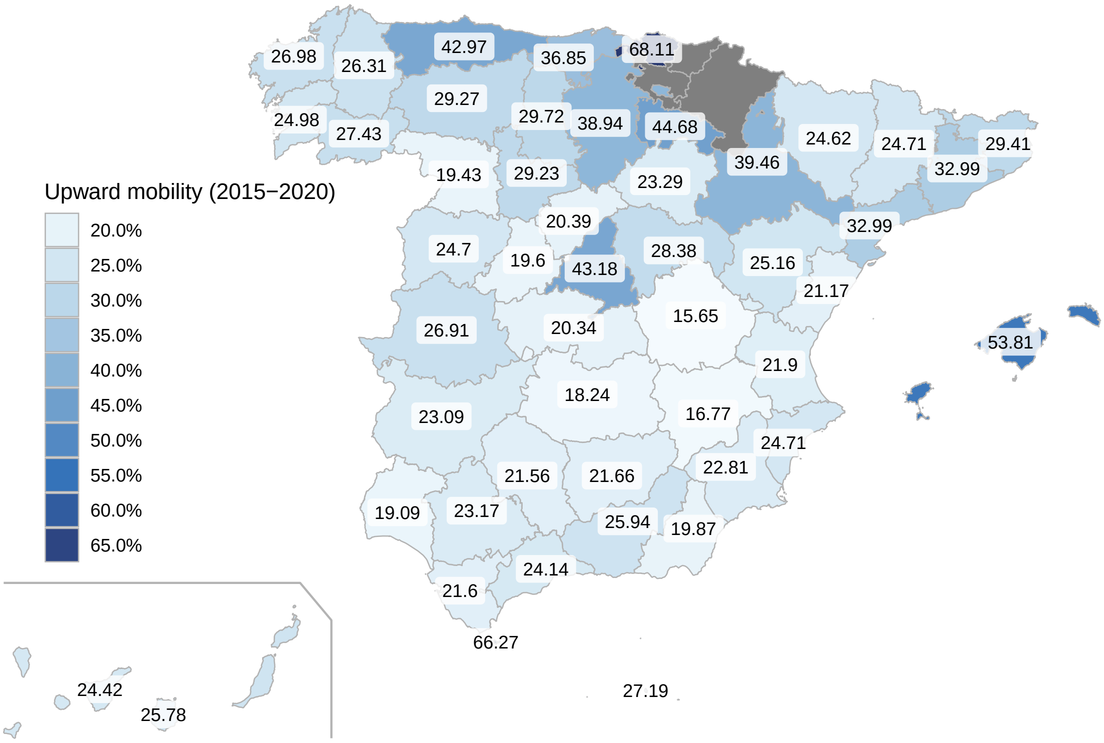


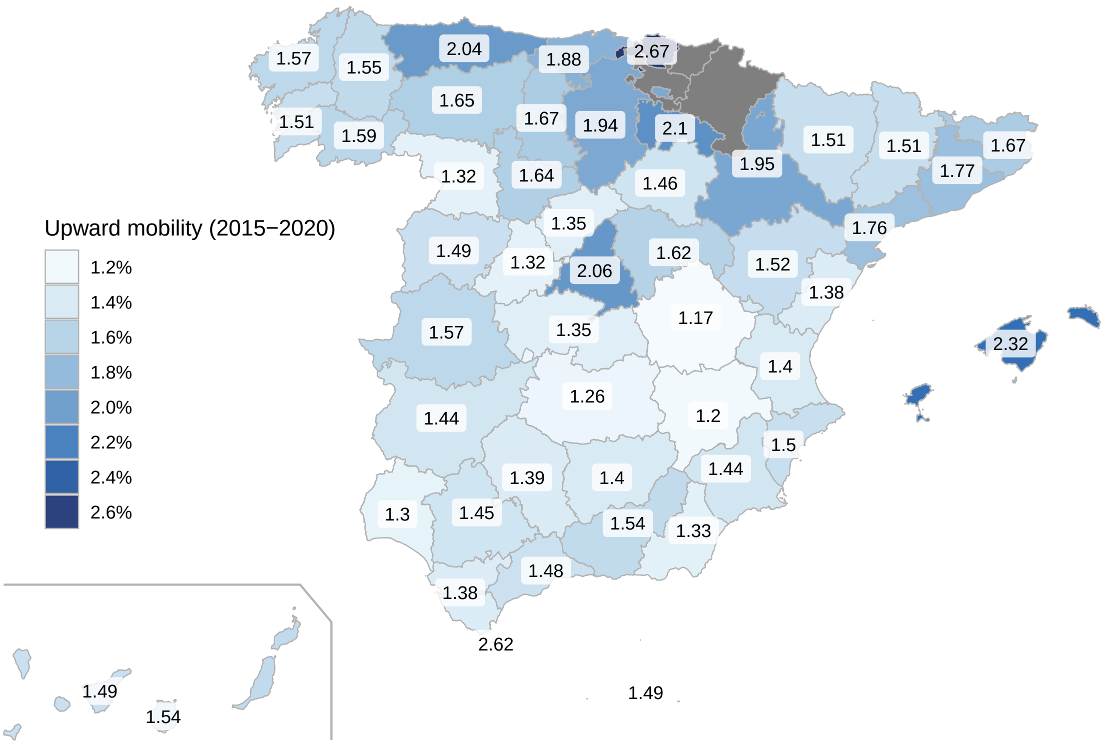
Figure A.2: Fraction of missing data in the ADRH dataset.



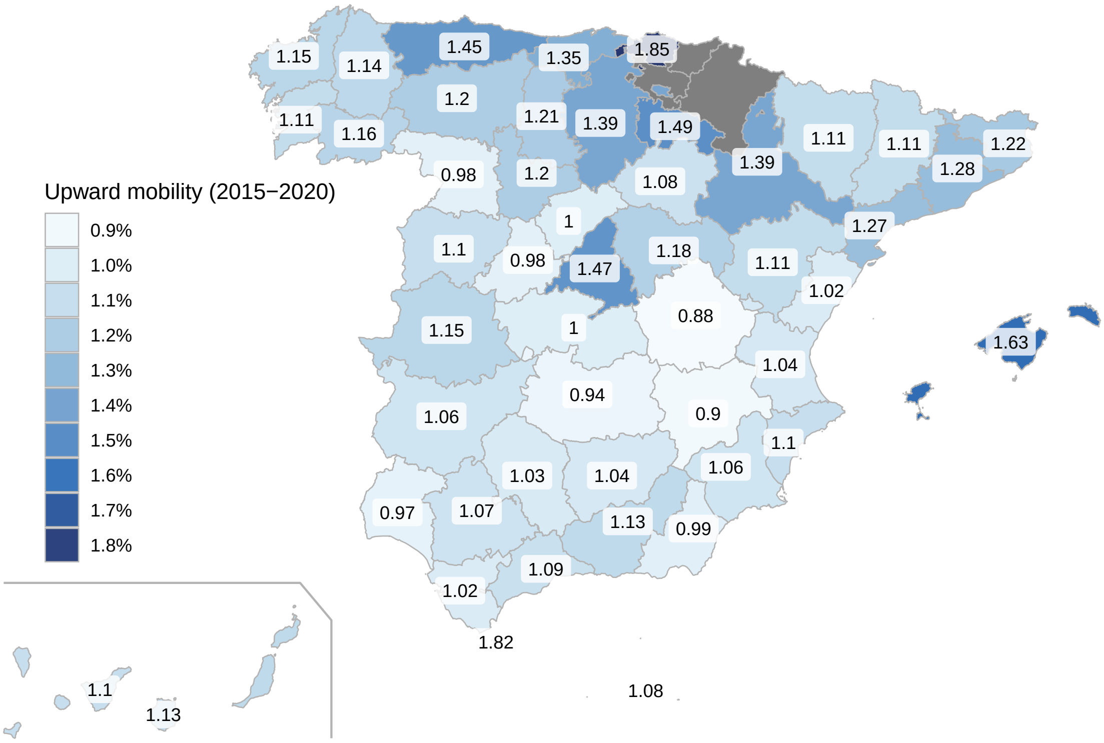
(a) Upward mobility when $\alpha = 0.3$



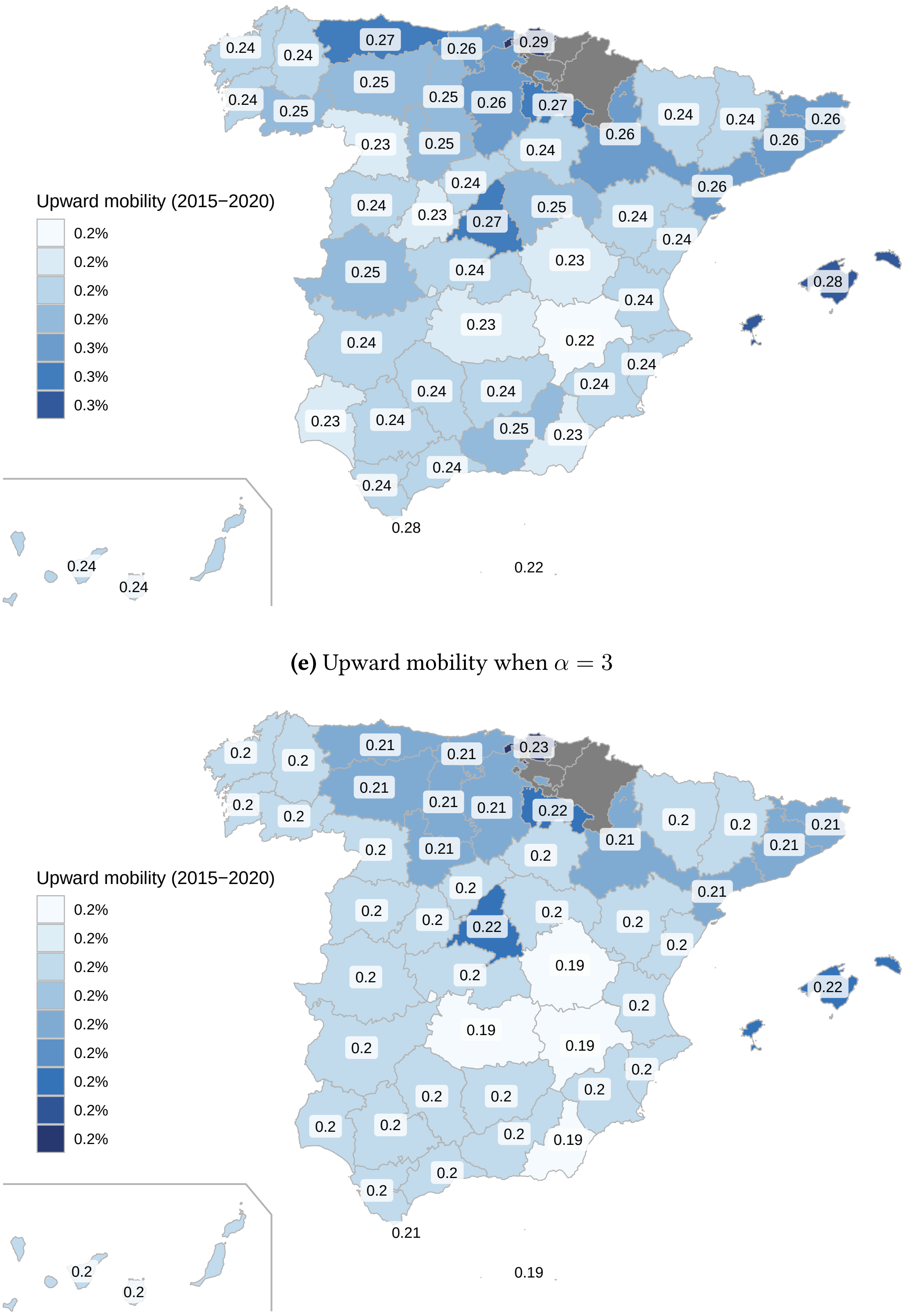
(b) Upward mobility when $\alpha = 0.5$



(c) Upward mobility when $\alpha = 0.9$



(d) Upward mobility when $\alpha = 1$



(e) Upward mobility when $\alpha = 3$

(f) Upward mobility when $\alpha = 5$

Figure A.3: 5-year period heat maps of provincial upward mobility testing different pro-poorness coefficients.

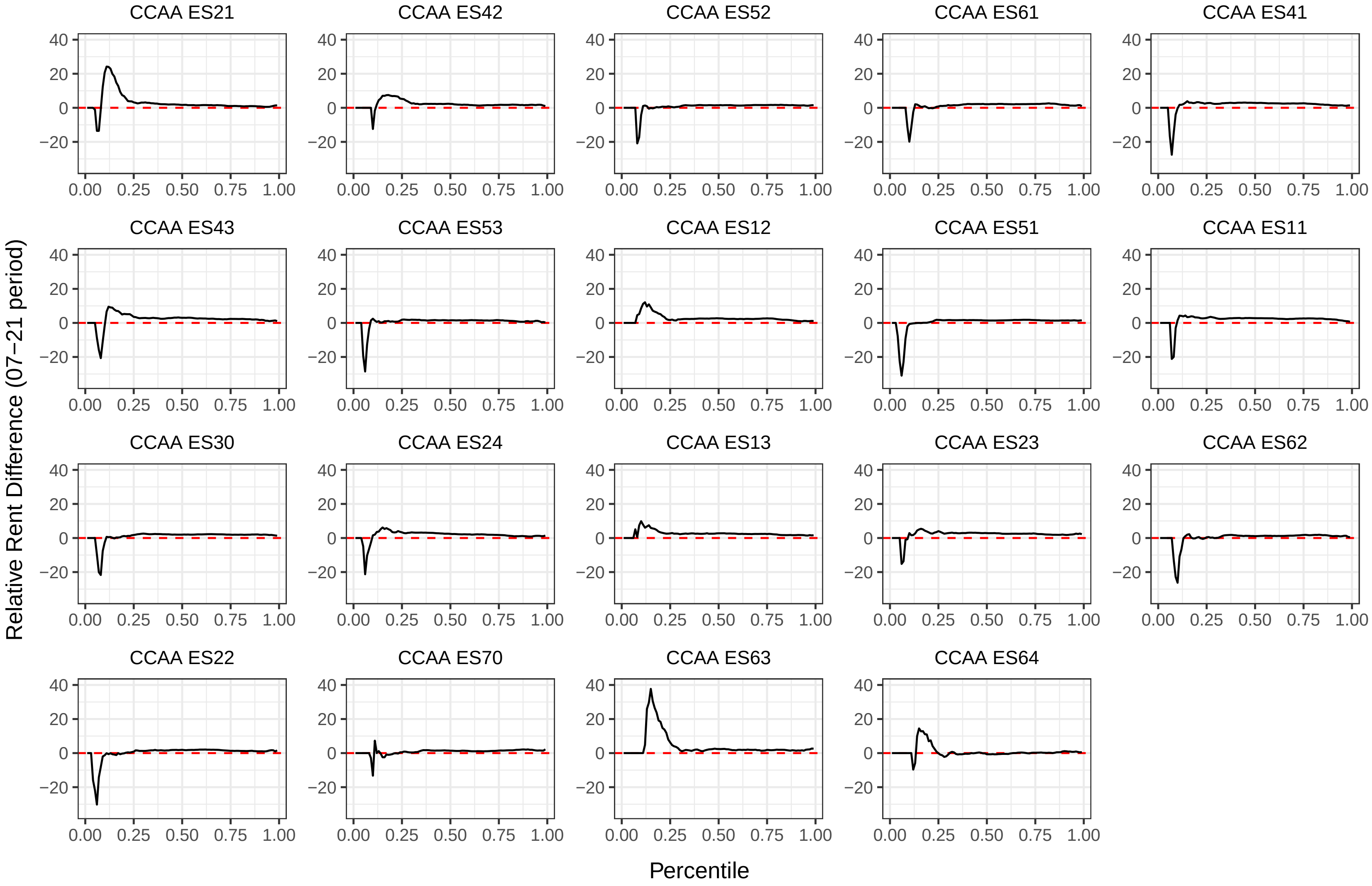


Figure A.4: Growth rates grid of all percentiles in autonomous communities using NUTS nomenclature. It uses ECV income data.

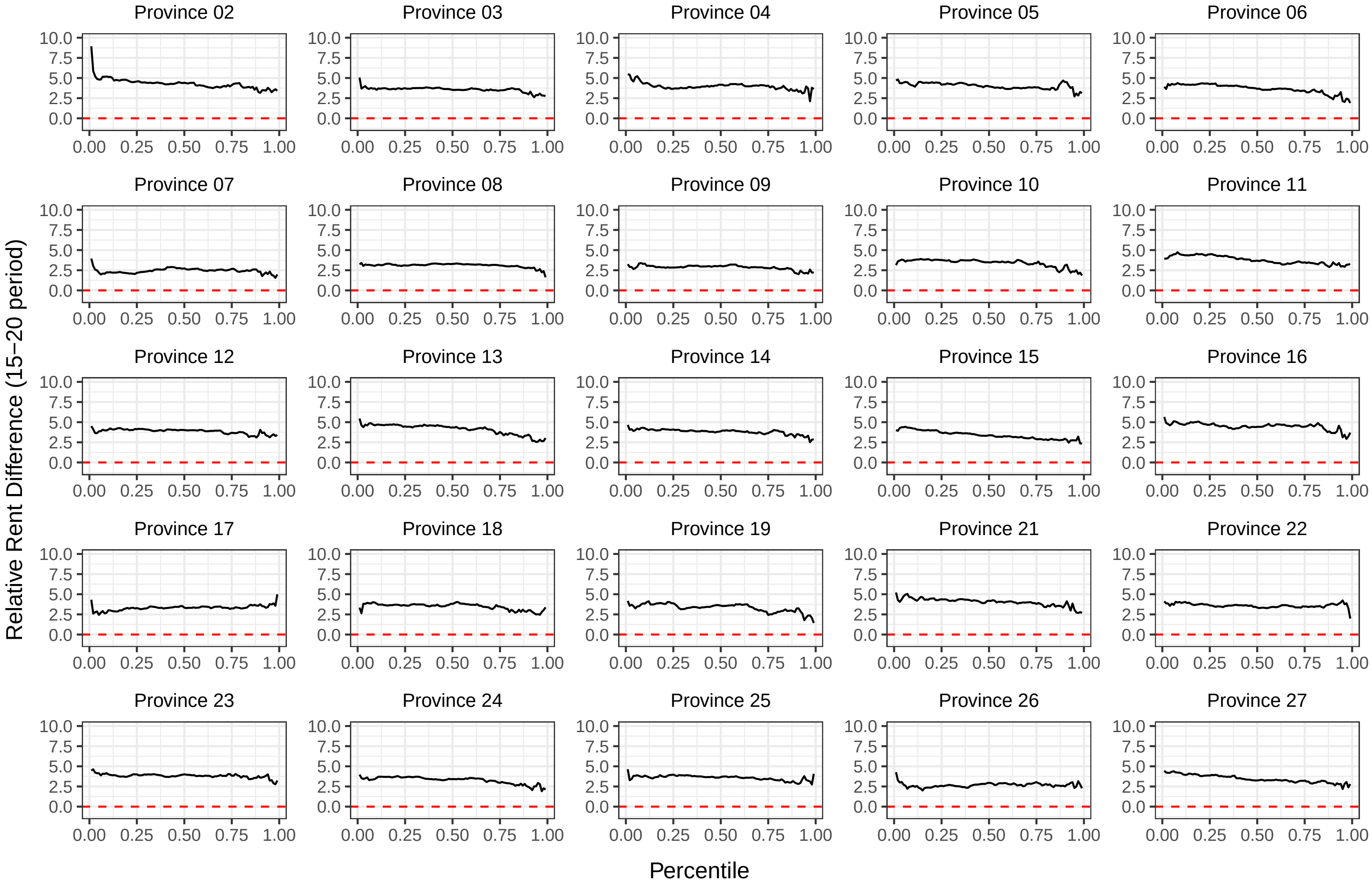


Figure A.5: Growth rates grid of all percentiles up to province 26 using cpro nomenclature. It uses ADRH data.

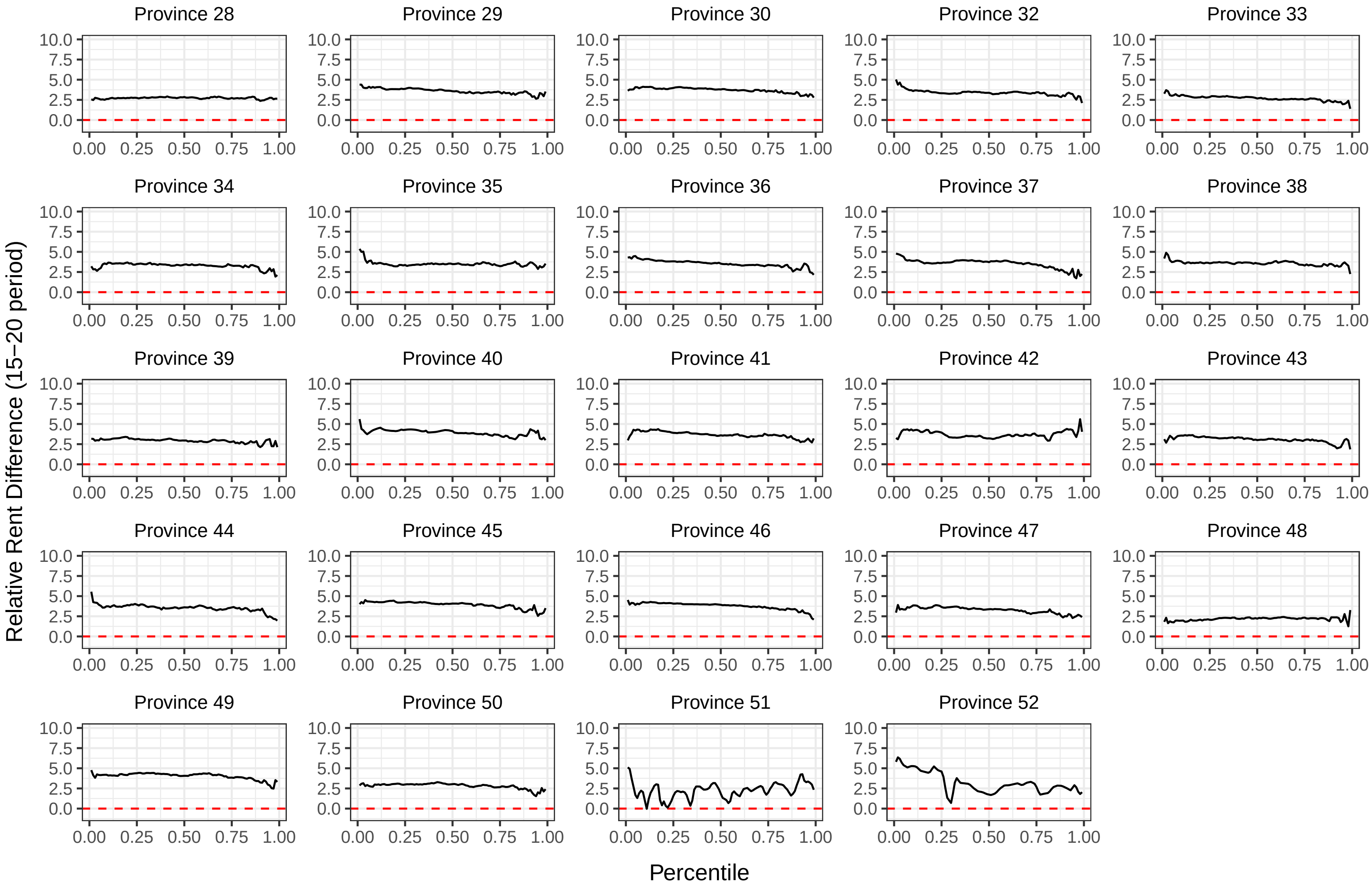


Figure A.5: Growth rates grid of all percentiles of the remaining provinces.

Table A.1: Table with the proportion of variance explained by each principal component. Notice that considering all 17 principal components will be the same as working with all your original variables but with a rotation of the axis.

Principal Component	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	24.092	0.3414	0.3414
PC2	20.875	0.2563	0.5978
PC3	13.200	0.1025	0.7003
PC4	112.959	0.07506	0.77531
PC5	10.862	0.0694	0.8447
PC6	0.9145	0.0492	0.8939
PC7	0.80524	0.03814	0.93204
PC8	0.6239	0.0229	0.9549
PC9	0.51149	0.01539	0.97033
PC10	0.39630	0.00924	0.97957
PC11	0.36983	0.00805	0.98761
PC12	0.30697	0.00554	0.99316
PC13	0.24834	0.00363	0.99679
PC14	0.16040	0.00151	0.99830
PC15	0.12830	0.00097	0.99927
PC16	0.09933	0.00058	0.99985
PC17	0.05084	0.00015	100.000

Table A.2: First 5 principal components' loadings on each variable. They explain the correlation between a component and the variables of the data frame.

	PC1	PC2	PC3	PC4	PC5
Healthcare workers ratio	0.134	-0.042	0.576	-0.395	0.050
E.Infantil	-0.328	0.102	-0.238	0.059	-0.175
E.Especial	-0.151	-0.063	-0.129	-0.186	0.752
Prog.GarantíaSocial	-0.376	0.003	0.071	-0.211	-0.122
E.Primaria	-0.373	-0.028	0.019	-0.215	-0.229
C.F.G.M.	-0.284	-0.020	0.369	0.054	0.111
C.F.G.S.	-0.261	-0.043	0.178	0.592	0.196
E.S.O.	-0.365	-0.095	0.122	-0.139	0.185
Bachillerato	-0.330	-0.158	-0.022	0.303	-0.273
UE28 without Spain	-0.055	-0.332	-0.375	-0.222	-0.023
Non-EU28 Europe	0.009	-0.334	-0.304	-0.134	0.089
Africa	-0.297	0.190	0.152	-0.281	-0.169
North America	0.155	-0.336	0.290	0.060	-0.003
Center America and Caribbean	0.018	-0.391	0.163	0.285	0.025
South America	-0.029	-0.437	-0.091	-0.087	-0.186
Asia	-0.180	-0.353	0.037	-0.052	0.172
Oceania	0.179	-0.329	0.172	-0.095	-0.270

Table A.3: Principal components regression results comparison including 2-4 PC. Model 1 is the choice with the best performance in terms of explanatory power and validity, at the expense of providing the effect of HEALTHCARE WORKERS RATIO on upward mobility (Including 2 PC accounts for 59.7% of the variability in the set of independent variables).

	Dependent variable:		
	Upward mobility		
	(1)	(2)	(3)
PC1	0.570^{***}	0.570^{***}	0.570^{***}
	(0.178)	(0.183)	(0.188)
PC2	-0.277	-0.277	-0.277
	(0.205)	(0.211)	(0.217)
PC3		0.141	0.141
		(0.333)	(0.343)
PC4			0.167
			(0.401)
Constant	-0.756^{*}	-0.756^{*}	-0.756
	(0.417)	(0.428)	(0.440)
Observations	19	19	19
R^2	0.431	0.437	0.444
Adjusted R^2	0.360	0.325	0.286
F Statistic	6.052^{**} (df = 2; 16)	3.888^{**} (df = 3; 15)	2.799^{*} (df = 4; 14)
Note:	*p<0.1; **p<0.05; ***p<0.01		