LS
7/21/08
11:15 am

Intergenerational Mobility and the Informative Content of Surnames *

Maia Güell
Universitat Pompeu Fabra,
CEP (LSE), CREA, CEPR & IZA

José V. Rodríguez Mora†
University of Southampton,
Universitat Pompeu Fabra, CREA and CEPR

Chris Telmer
Carnegie Mellon University

May 2007

Abstract

We propose an alternative method for measuring intergenerational mobility. Measurements obtained from traditional methods (based on panel data) are scarce, difficult to compare across countries and almost impossible to get across time. In particular, this means that we do not know how intergenerational mobility is correlated with growth, income or the degree of inequality.

Our proposal is to measure the informative content of surnames in one census. The more information the surname has on the income of an individual, the more important is her background in determining her outcomes; and thus, the less mobility there is.

The reason is that surnames provide information about family relationships because the distribution of surnames is necessarily very skewed. A large percentage of the population is bound to have a very unfrequent surname. For them the partition generated by surnames is very informative on family linkages.

First, we develop a model whose endogenous variable is the joint distribution of surnames and income. There, we explore the relationship between mobility and the informative content of surnames. We allow for assortative mating to be a determinant of both.

Second, we use our methodology to show that in large Spanish region the informative content of surnames is large and consistent with the model. We also show that it has increased over time, indicating a substantial drop in the degree of mobility. Finally, using the peculiarities of the Spanish surname convention we show that the degree of assortative mating has also increased over time, in such a manner that might explain the decrease in mobility observed.

Our method allows us to provide measures of mobility comparable across time. It should also allow us to study other issues related to inheritance.

Key words: inheritance; birth-death processes; cross-sectional data; population genetics.
JEL codes: C31, E24, J1.

*We thank Namkee Ann, Manuel F. Bagüés, Melvin Coles, Vicente Cuñat, John Hassler, Ramón Marimón, Laura Mayoral, John Moore, Diego Puga and Gary Solon for very useful suggestions, and Anisha Gosh, Rasa Karapanza, Ana Mosterín and Robert Zymek for superb assistance. We also thank the comments of seminar participants at Toulouse University, Universitat Pompeu Fabra, IIES at Stockholm University, Southampton University, the NBER Summer Institute, ESSLE, the CEPR Public Economics Meetings, Queen Mary, Tinbergen Institute, Universidad de Murcia, Universidad Carlos III de Madrid, University of Salerno, University of Bristol, University of Edinburgh, London School of Economics, City University London, EUI (Florence), FEDEA and CEP. JVRM thanks the financial support of the *Fundación Ramón Areces*.

†Corresponding author: Economics Division. School of Social Sciences. University of Southampton. Southampton SO17 1BJ. United Kingdom. Email: sevimora@gmail.com

1 Introduction

We do not know how important the economic status of parents is for determining the economic status of their children. At most we have very vague answers since intergenerational mobility is notoriously hard to measure. Traditional estimation methods require very long panel data linking economic outcomes of parents and children.[1] It is seldom the case that we have access to these data, but even when these are available it is well known (Solon (1992, 2002)) that they are of limited use for understanding how mobility compares across countries and time. Consequently, we have scant knowledge about the degree of intergenerational mobility, its evolution over time or its geographic distribution. Thus, we do not know its correlation with growth, income or the degree of inequality.

In this paper we attempt to overcome some of these limitations by introducing a new source of data and a new methodology that relies on cross sections, and not panels, in order to study these type of longitudinal questions. Our main claim is that by observing a cross section of surnames and income of a society (something that we argue is easy, as all governments compile censuses) it is possible to obtain information about intergenerational mobility *even if we have no explicit link between the income of parents and children.*

Thus, the new source of data is a roster of the population specifying the surnames and a measure of economic wellbeing of the individuals; a census. The main idea is simple. In societies with low intergenerational mobility, children inherit economic wellbeing (*e.g.*, income, wealth, education) from their parents. Most children also inherit their parents' surname. The joint distribution of economic wellbeing and surnames will reflect this and will allow us to identify the degree of mobility. Dividing the population by their surnames we obtain a partition of the population that is very related to family linkages, and this allow us to explore the importance of background in determining an individual's economic wellbeing.

The contributions of the paper are: (1) We present a model that determines endogenously the distribution of surnames and income. (2) Using the model we show that surnames are bound to be informative. (3) We explain the reasons why it is so, and how do they relate to intergenerational mobility and assortative mating. (4) We show that the amount of information that they contain is negatively related to the degree of intergenerational mobility. (5) Applying our methodology to Spanish data, we show that surnames are informative in the manner predicted by the model. (6) This allow us to observe that the amount of intergenerational mobility has decreased over time in a large region of Spain. (7) Finally, we use the peculiarities of the Spanish naming convention to show that this trend is explained by an increase in the degree of assortative mating.

In a nutshell, the intuition for our results lies on the fact that surnames act as a marker of the things that individuals inherit from their parents. By themselves they have no impact on the income of individuals, but they are informative because they are passed along (inherited) with other characteristics that do have effects on observable outcomes (education, income, etc.). The more inheritance of economically meaningful characteristics, the more information do surnames contain about the effects that these characteristics produce. Notice that our problem is that we can not observe inheritance directly. We do observe economic outcomes, but the only inherited

[1] A minimum of 30 years if you want only one observation of the son's income; 85 years if you want the permanent income of both (see Hertz (2007)).

characteristic that we observe is the surname. What we would like to know is how strongly economic outcomes are related to unobserved inheritance (income, genes, education, status, whatever). If the relation is strong, then the outcomes will be related to the inherited observable, the surname. Surnames leave an imprint that allow us to measure the importance of inheritance in determining outcomes. An example may help understand why surnames may have information, and why it is not obvious that they do.

Imagine a country called *Commonnamesland*. It happens to be the case that in *Commonnamesland* there are only two types of individuals, rich and poor, and *one generation ago* everybody who was rich was called *Richmanson*, while everybody who was poor was called *Poormanson*. Now we wonder how informative will the surnames be *today*. Clearly, it depends on how much upbringing matters determining your position in society.

If upbringing is very important (if intergenerational mobility is low) one generation later most males who happen to be called *Richmanson* are actually both rich and the sons of a rich person. Thus, with low intergenerational mobility, *today* it is still the case that the surname informs about the income. There are two noteworthy corollaries.

First, how informative surnames are depends on how much intergenerational mobility there is. If there was no inheritance and background would not matter at all, then to be called *Richmanson* would not inform at all about your status *today*, only about the status of your family one generation ago. This is the essence of the mechanism by which we will be able to infer the degree of mobility.

Second, the amount of information in the surnames today ought to be smaller than what it was yesterday. If the income generating process is unique and stationary, and if surnames are inherited only in this manner (surnames are never created, surnames never disappear) in the long run all surnames of *Commonnamesland* will have the same income distribution. Being called *Richmanson* in the long run would not indicate that you are the son of a rich man, and much less that you are rich yourself. Eventually a *Richmanson* and a *Poormanson* would be equally likely of being rich or poor. This second point is what makes our methodology less than obvious. If surnames provide a way of inferring intergenerational mobility there must be something else preventing the uniformity in the distribution of income per surname.

The additional ingredient is that surnames die and are born. They die when the last *male* holder of the surname dies without *male* descendants. Surnames are born when somebody (somebody *male*) changes from the one that was given to him by his father to a new surname not previously existing in the population (for whatever reason). Thus, the distribution of surnames in a population follows dynamics that are akin to the ones of the distribution of genes. A distribution of surnames like the one predicted in *Commonnamesland* is not possible under the standard western naming convention. The distribution of surnames is bound to be very skewed, with some surnames being relatively common (and from them little information can be extracted) while at the same time a very large percentage of the population has very unfrequent surnames. These unfrequent surnames are our main source of information.

Meet *Unfrequentnameland*. This is a country whose surname distribution is generated by a death-birth process as the one outlined above. In *Unfrequentnameland* the most common surname is *Smith*. There are many people called *Smith*, and any random pair of them is not likely to have a close family link. They are like two individuals

called *Richmanson* in the long run distribution of *Commonnamesland*; from them we can not learn much about the value of inheritance.

On the other hand, a consequence of the surname convention is that in *Unfrequentnameland* there are many individuals who happen to have very uncommon surnames. There are families called *C3PO* and *R2D2*. In steady state the distribution of surnames will remain constant in *Unfrequentnameland*. Which surnames happen to be more or less frequent will (of course) change, but the distribution of the frequencies of surnames will remain.

Because *C3PO* and *R2D2* are unfrequent surnames, if we take two individuals who happen to be called *C3PO*, they are very likely to have close family ties. If the degree of intergenerational mobility is high we should expect that if a *C3PO* is rich, the rest of the *C3PO*s are also likely to be rich. In the same manner, if we know that a certain *R2D2* is poor, we would infer that the other *R2D2*'s are also likely to be poor. Notice that we can make these inferences only if background is important.

Thus, in *Unfrequentnameland* we would be able to use census data in order to extract information on the degree of intergenerational mobility. And the western naming convention insures that, insofar we care, all countries following it are essentially like *Unfrequentnameland*.

In the first part of the paper we present a model of the joint determination of surnames and income. We define the informative content of surnames (ICS); and show that it is monotonously increasing in the importance of background to determine outcomes (monotonously decreasing in the degree of intergenerational mobility). The model will help us understand the different mechanisms by which surnames carry information on economic wellbeing.

In the second part of our paper we use our methodology in order to study empirically the informative content of surnames in a large Spanish region, and what light this may shed on the evolution of intergenerational mobility there.

The paper concludes by summarizing the results and indicating the next steps of this project, in particular with respect to the comparisons of intergenerational mobility across countries and regions.

Before doing that, in the next section, we review the relevant literature and discuss the need of an alternative method of measuring intergenerational mobility.

2 Literature review

Most of the empirical work on intergenerational mobility looks at the correlation between the income of parents and children using panel data and the "value" of mobility is understood to be equal to one minus this correlation. This requires very long panel data which is available rarely; but even when available there are big problems widely recognized in the literature, at least since Solon (1992): (1) current income is a noisy representation of lifetime income and this establishes an upward bias in the mobility measures. (2) children's income tends to be measured at the starting of their career, which tends to produce a bias, as the lifetime income of the educated can be very badly measured by the income of their first years.[2] (3) Samples are biased, as the attrition rate is different for

[2]See for instance Haider and Solon (2006), Hertz (2007).

different groups of the population. (4) Obviously it takes time to construct a panel data base. This hinders the possibility of looking at the dynamics of intergenerational mobility.

This traditional approach has a before and after to Solon (1992). Before his paper the estimations available on income mobility (few and only for the U.S.) indicated almost always that the correlation of parents and children's income was low (high social mobility). For example, Behrman and Taubman (1985, 1990) or the work of Becker et al. (1967, 1979, 1986) considered correlations of around 0.2. The article of Solon (1992) showed that the previous estimations were biased and misleading, and that the access to long data panel could somehow reduce this bias, diminishing the noise in the estimation of both parents and children's income and finding a correlation of around 0.4, which translates into levels of mobility much lower than previously thought. This autocorrelation has been obtained with diverse data bases and somehow it has become the consensus correlation in the U.S. for the last third of the past century.

Following the methodology of Solon estimations have been made in several European countries. In the Nordic countries there is relative facility to collect data panel of the required type. Consequently, we have good measurements for the Scandinavian countries.[3] In addition we have estimations for Great Britain, Germany and Italy. [4]. Recently Comi (2003) has provided with estimations for 12 European economies using the European Community Household Panel. For the rest of the world we hardly know anything.[5]

Unfortunately, it is very difficult to compare these estimations across countries (Solon (2002)) since the panels used are different; and thus the biases, the levels of noise and the problems of selective disappearance of the sample produced are different. Thus, we can compare the dispersion of the income between countries, but not their intergenerational mobility.

The problem has become more serious in the last decades as there has been a well documented increase in the dispersion of income. If mobility had also decreased, we should judge the problem as more severe than what it would be if mobility had increased. Unfortunately we know hardly anything about the time evolution of mobility.[6] For the US, recent papers (Lee and Solon (2006) and Hertz (2007)) show that existing widely divergent results suffer from small samples as well as the aforementioned age-related bias and sample attrition problems.[7] Taking this into account leaves the authors inconclusive about trends in intergenerational mobility. For Great Britain, (Blanden, Goodman, Gregg, and Machin (2004)) suggest a decrease in the mobility between two cohorts (born in 1958 and 1970, respectively). In any case the estimations done on the temporal evolution have serious problems of interpretation, because they suffer from many of the same problems that the comparisons between countries: they use panels that are different for the different cohorts. Not yet there is a data base panel covering fully three generations (120 years) that would allow us to look at the trend in mobility. Even if there were to be one, sample

[3] See Björklund and Jäntti (1997), Osterberg (2000), Osterbacka (2001) and Björklund et al. (2002).

[4] See Dearden, Machin, and Reed (1997); Wiegand (1997) and Couch and Dunn (1997); Checchi, Ichino, and Rustichini (1999), respectively.

[5] There are some estimations for South Africa (Hertz (2001)), Brazil (Dunn (2004) and Ferreira and Veloso (2004)), Singapore (Ng (2007)) and Malaysia (Lillard and Kilburn (1995)) and for another handful of countries in Grawe (2004). These estimations are typically done with retrospective information about the parents.

[6] See Solon (2004) for a proposed frame to study the temporary evolution.

[7] For instance, Mayer and Lopoo (2005) and Fertig (2007).

attrition would make it of doubtful utility.

Alternative approaches

Alternative approaches with a smaller dependency on panel data have been attempted in order to overcome these problems. Thus, there are studies on the behavior of siblings,[8] or neighbors.[9] These approaches have two problems that make difficult their use: (1) They require information on family bonds, which is not simple; even when available, the comparison between samples would be difficult for the same reasons as with panel data. (2) Additionally,[10] these approaches do not allow to make inferences on the direct incidence of the economic position of the parents, but only of the effect of family background.[11]

Another way of escaping from panel data is to approximate parents' income based on available information of the child (e.g. state of birth) as in Aaronson and Mazumder (2007) who exploit large samples from the US decennial Censuses. Employing a two sample estimator and, interestingly, find for the US that mobility has increased from 1950 to 1980 but has declined sharply since 1980.

Outside economics the tradition is to measure intergenerational social mobility not based on income, but on the "social prestige" associated to the professions of parents and children (Duncan, Featherman, and Duncan (1972)). Problems of this approach are that it is difficult[12] to judge the social prestige of professions, how it evolves through time and still if we knew how to do it, is difficult to interpret its meaning. The child of a very famous doctor who becomes a country doctor would be assumed to produce persistence.

Other related literature

There exists a substantial literature that deals with first (given) names. They are useful and interesting because they are *endogenous*, at least from the point of view of the parents. We want to emphasize that this is almost exactly the opposite reason of why we use surnames. We use surnames because they are completely *exogenous* to the individual, to her father and to the social position of both. Surnames are a marker, and we do not need to know nor learn its meaning. The surname "Johnson" always means that you are the son of somebody called Johnson, the first name "John" may mean that your parents were rich, but one generation later it could mean that your parents are poor (Fryer and Levitt (2004) and Levitt and Dubner (2005)). The distribution of first names follows complex social rules that depend on the specificities of group identity and social dynamic, and of the wants of imitating the better off and differentiating from the worse off. The distribution of surnames on the other hand follows the enormously simpler laws of genetics. In any case, Bertrand and Mullainathan (2004) make use of the fact that the distribution of first names differs between Afro-American and the rest of the population of the U.S. in order to test for racial discrimination in the context of a field controlled experiment.

The usage of surnames as a way of recovering information on the evolution of human populations has a long

[8] A representative study of this approach is Solon, Corcoran, Roger, and Deborah (1991).
[9] For example Page and Solon (2003), Dahan and Gaviria (2001) or Levine and Mazumder (2007).
[10] See Solon (1992).
[11] As we will see using surnames we have the potential of partially disentangling them, as we have both children and parents.
[12] At least for economists, apparently not for sociologists.

tradition in biological anthropology. George Darwin (son of Charles) published in 1875[13] an article using surnames (specifically: marital isonymy, equal surnames) in order to determine the frequency of cousin marriages in England.[14] Surnames have been since then used for determining the degree of inbreeding inside populations, for determining population movements and determining population homogeneity (see Lasker (1985)). Still, the availability of DNA mapping has induced a decrease of the usage of surnames in biological anthropology as population genetics has become more central. In the context of our paper, one remarkable feature of this literature is its predisposition at the usage of Spanish surnames. This is because they carry information on both parents and because they mutate with less frequency than in other cultures.

A relevant reference in biology on the mathematics of surname distribution is Manrubia and Zanette (2002). These authors consider a model of surname generation with exponential population growth. As in our model, newborn agents receive a new surname (a "mutation") with a fixed probability. With the complementary probability they are randomly assigned an existing surname from the existing set of names. In the latter case, the likelihood of receiving a particular, existing surname is proportional to that name's frequency in the existing population. In their baseline model the cross-sectional distribution of surnames (the frequency associated with which the n^{th} most common surname) follows Zipf's law: the frequency is inversely proportional to the surname's rank. This feature of their model is consistent with data. What's inconsistent, however, is the model's time-series behavior. The number of surnames grows exponentially, at a rate determined by the mutation rate and the population growth rate. In (some) observed data the number of surnames seems to decrease with time. Motivated by this, Manrubia and Zanette (2002) add mortality risk to their model and show that under certain parametrizations the model is consistent with a shrinking set of surnames. Our model is distinct in that – by virtue of the fact that we rely on computational simulations – we keep track over time of lineages. Doing so is critical for us since we ultimately care about the joint distribution of economic characteristics and surnames, not just the marginal distribution of the latter which is the focal point of the study by Manrubia and Zanette (2002). On the other hand, our current approach does not allow for population growth, something which we leave for future work.

There are three papers that also use Spanish surnames. Collado, Ortuño-Ortín, and Romeo (2006) is an attempt to study the degree in which consumption patterns are learned from the environment and to which degree they are inherited. They use the distance in the distribution of surnames between provinces and do not use microdata, but only aggregated distributions, finding that food consumption patterns are less likely to be a consequence of the environment. Perhaps closer in spirit to us it is Angelucci, De Giorgi, Rangel, and Rasul (2007) who use surnames in microdata in order to identify family links in Mexico. Nevertheless our objectives and methodology are also very different, as they use surnames as family links explicitly and intensively (this is, determining who is linked with whom in a small sample) while we use them implicitly and extensively for the whole population. Finally Bagüés (2005) uses very long and unusual surnames in order to determine family relationships (and the possibility of corruption) in the grades obtained in public examinations in Spain.

[13] Darwin (1875), cited in Lasker (1985).

[14] He was worried about the possible nocive effects of consanguinity between parents, as his father and mother were first cousins.

To our knowledge ours is the first paper that uses surnames in order to learn about intergenerational mobility, that uses extensively the surname information of census data, and that builds a theoretical framework to understand the information that can be obtained from surnames.

3 Surnames are Informative about Socioeconomic Status

In this section we present a model of joint determination of surnames and income. To do so, we start by defining the concepts that we will use in the rest of the paper.

3.1 Definitions

3.1.1 Census

We define a census as a list of all the individuals of a certain population. For each individual we have a minimum of two variables: (1) her surname, and (2) a measure of her economic wellbeing (income, education, consumption, etc.). The census does not need to specify the family linkages between the individuals. It may contain information on other individual characteristics (gender, age, place of birth, ethnic origin, etc.) but these are not necessary and for most of our analysis we will do without them.

3.1.2 The Informative Content of Surnames (ICS)

We define the informative content of surnames (ICS) as the difference between the (adjusted) R^2 of two regressions.[15]

The first regression has on the left hand side an index of the economic wellbeing of an individual. On the right hand side it has all the individual controls deemed necessary [16] and also a dummy of the surname that the individual holds. This last variable is the focus of the paper. Notice that it refers to the *specific* surname that the individual holds. It does not refer to general characteristics of the surname, like ethnic origin or the relative frequency of the surname. What we measure in this regression is how much can we get to know about an individual if we know his specific surname. Obviously, the larger the R^2, the more information that surnames, in general, have. Notice that even if we look at specific surnames we measure the informative content of *all* surnames by looking at the R^2.

The second regression is identical to the first one, but on the right hand side instead of placing a dummy for surname we place a dummy per "fake–surname". These "fake–surnames" are generated by us and assigned to individuals in a random manner. Our restriction is that *the distribution of "fake–surnames" is identical to the distribution of real surnames*. We are measuring how much information can be obtained by grouping individuals by surnames independently of their family linkages.

[15] To ease the reading, hereafter, we will simply refer to the R^2.

[16] In the simulations they will be unnecessary, in the empirical regressions of section 4 they refer to the background of the individual, and are exogenous to her: place and date of birth, gender, etc.

We are interested in surnames because they are a partition of the population that orders them in a way that is informative about their family links, and not because is just a partition. This might be a concern, particularly taking into account that the distribution of sizes of the groups is very skewed with some surnames holding a large percentage of the population (thus, not inducing any order) while others are very small, and much more likely to induce order (popping up the R^2). The very skewness of the distribution could make the adjusted R^2 unsuitable to correct for this. Thus, we take a conservative approach to insure that we measure the effect of family links.

We define "fake–surname" as a partition of the population that has the same distribution that the one of "real" surnames, but that is orthogonal (by construction) to the family linkages between the individual members of the census. Then, our definition of the informative content of surnames is:

$$ICS = R^2_{\text{regression with surname dummies}} - R^2_{\text{regression with "fake–surname" dummies}} \tag{1}$$

There are notorious advantages of this definition. Imagine a country where for some reason each individual had a different surname. The regression with the true surname dummies would have a R^2 of one, but their information would be all fallacious. According to our definition the informative content would be zero.

The use of "fake–surnames" and our definition of ICS acts as an insurance policy. It reassures that we measure only the family related information, and not other issues consequence of the particularities of the type of data that we use.

3.1.3 Lineages

It is very useful to define a *lineage* as the set of individuals from all generations who have a common male ancestry *and share a surname*. Two individuals may have a common ancestry, but they might not share the same surname if one of them (or one of their male ancestors) changed his surname. They would not be part of the same lineage. Death and birth of lineages is at the root of the process that we study.

3.2 A simple model of surname and income co-evolution.

In this section we present what we believe is the simplest model that generates a joint distribution of surnames and income. The procedure is the following. The model takes as exogenous (1) the income process and the degree of inheritance and (2) the rules that determine the distribution of the population and the birth and death of surnames.

The endogenous variable is the joint distribution of surnames and incomes. We will determine it in steady state, *as a function of the exogenous parameters, including the degree of inheritance.* From this distribution we will center our attention in one moment: the informative content of surnames (ICS), as defined in equation (1). Our main focus will be the correlation between the ICS and the degree of intergenerational mobility. This is, we want to know if a larger ICS means the the degree of inheritance is larger, and the reasons for their relationship.

The final objective of the project is to develop a method for extracting information on the degree of mobility from a census, and that is precisely what we do in section 4 for Spanish data. Nevertheless, in order to achieve

that objective, we do here exactly the opposite exercise. We *assume* values of the parameters that determine intergenerational mobility, and from them we generate a census. Then we look at how the characteristics of the census depend on the parameters that we have assumed. Doing this we hope to learn how to interpret a census.

Surnames are usually inherited from parents to children across the male line. This male-centered inheritance of surnames makes convenient to disregard the role of women, and we will assume that this is a male only society where procreation happens by some kind of asexual process.[17] In sections 3.5 and 3.6 we will reintroduce and give a role to women.

For simplicity the model has successive non-overlapping generations and the exogenous ingredients are constant over time. The income process is the same for all agents of the economy at all times, and the same rules for transmission of surnames apply at all times, even though they can be different for different individuals depending on their income (see below).

At each moment of time the economy is characterized by a census. For each individual the census has only two entries. These are the surname of the individual and in the income that he may have. The two components of the census are very different in their character and in the way that they are inherited. We define the vector of incomes in census t as Y_t and the vector of surnames as A_t. Each of them being a vector with as many entrances as individuals are alive at period t. The surname of the i-th individual is $a_t(i) \in \mathbb{R}$, and his income is $y_t(i) \in \mathbb{R}$.

Individuals live for one period only. Thus, they appear only in one census. A father and his son can not appear in the same census. The closest family relation that an individual may have in the census where he appears, is his brother. This is an important simplification, as it makes the model more tractable. It will be clear from the argument that our results would only be reinforced if we were to include more generations coexisting in the same census. We next turn to the different exogenous ingredients of the model.

3.2.1 Ingredient number one, get an income transmission process

We define inheritance as the coefficient of correlation between the income of a father and that of his child. This is in line with the bulk of the literature on the topic. For the shake of simplicity we assume that the income process follows a mean reverting AR(1) process.

$$y_t = \rho y_{t-1} + \varepsilon_t \tag{2}$$

where y_t is the income of a person of generation t, ε_t is a white noise shock and y_{t-1} is the income of the father of the individual. The conditional variance of the process is the variance of ε_t, that we denominate V_ε.

Conditional on the father all his children have the same ex-ante stochastic distribution of income. The eventualities of life accounting for the differences between them. We assume further that ε_t is identically distributed across all individuals.[18]

[17] As a consequence, we find convenient for the rest of the paper to hold to the convention of referring to everybody male with the male pronoun (he-him-his, rather than the usual, neutral, she-her-hers). In this case clarity overcomes the advantages of political correctness.

[18] In this paper we follow the main line of the literature and set our focus in ρ. Thus, we do not allow the variance of the shock to be different among siblings than among non family related individuals. We could do so in a overlapping generations model (thus including both parents and their children in the same census), but that would complicate the analysis giving relatively little in return. In any

A high value of ρ implies that the incomes of father and son need to be more similar than they would otherwise be. A second consequence of having a large degree of inheritance is that the more you know about a person, the more you know about their siblings *relative* to everybody else. The *unconditional* variance of the distribution generated by the income process equals the variance of the income in the population:[19] $V_{unconditional} = \frac{V_\varepsilon}{(1-\rho^2)}$.

If you know the income of the father, the conditional variance of the income of his children is simply V_ε, thus the ratio of the variance of the income of siblings to the variance of the income of the population at large is $(1-\rho^2)$. The larger the inheritance, the more homogeneous siblings are *relative* to how homogeneous is the total population of the economy (the smaller the ratio of the variances). The reason for this is that if the degree of inheritance were larger, the conditional variance would not change, but the unconditional variance would increase. The larger ρ, the more similar should be the income of two siblings relative to how similar are the incomes of two persons taken randomly.

The larger the degree of inheritance, the more homogeneous the income of a group of people who have relatively close family links (that can trace a joint descendant to a common ancestor) *relative* to how much we know about the total population.

If the model were of overlapping generations (thus including father and son in the same census), our results would only be strengthened. The reason is that there would be yet another mechanism by which more inheritance would produce more information. It would be not only that the variance of siblings is small *relative* to the variance in the population, but in addition to it the income of people with the same surname would be linked by the direct fact of inheritance.

The main exercise of this section will be to do comparative statics on ρ and V_ε, and see how they affect the ICS (and in general the distribution of income and surnames) in a census.

Notice finally that the surname of the individuals is assumed not to have any effect on their income. The income process (2) is completely independent of the surnames of the individual or his father (if they were going to be different). The only remarkable characteristic of the surname is that most often is going to be passed along form the father to the son. Nevertheless, that is enough to allow us to use them as markers. We turn now to the process of surname determination.

3.2.2 Ingredient Number Two: Get a Process of Surname Determination. Lineages

If we had an economy where each father had only one son to whom it would pass the surname with certainty, then the distribution of surnames would be constant over the whole economy. In such an economy surnames would contain no information at all. The reason is that the surname partition would bind together people with no family relationship between them because (1) there are no siblings, nor cousins, as each father has only one son, and (2) there are no father, nor uncles, in the same census than the individual, as we have only one generation per census.

In effect the surname and the *"fake–surname"* would be the same variable: a random distribution of people in

case, we intend to do so in an extension of the present paper.

[19] This is not true in the case of differences in fertility across income groups (to be seen later).

a random exogenous partition. There would be nothing real binding together people with the same surname.

We want to generate a distribution of surnames and income that can match the data from relatively few exogenous parameters. Thus, one of the conditions is that the distribution of surnames needs to be sufficiently skewed. Fortunately, birth-death processes allow to generate the power distributions akin to the ones observed in surname data. They are also the most natural approach to determine the evolution of the surname distribution.[20]

Lineages die and are born. A lineage is born whenever a new surname appears among the male population. A lineage dies whenever one of two possible circumstances occur: (1) Either the last male holding the surname dies without male descendants, or (2) the set of his male descendants leaves the lineage by changing their surname.

Mutations

We model the surname of an individual (a_t) as a number that is equal to the the surname of his father (a_{t-1}) with probability $1-\mu$. With probability μ, the individual adopts a new name that we assume that is a random extraction from a very large set of possible surnames Σ. This set of possible surnames is much larger than the number of individuals living in the economy. Thus, the new surname is almost certainly unique. We denote this probability μ as the *"mutation rate"* of surnames.[21] Thus, mutations are at the origin of lineage birth.

We are going to assume for most of the paper that the probability of a mutation is equal for all individuals. There is a trivial extension where this mutation rate depends on the status of the individuals. We discuss it briefly below, but it is irrelevant for our results.

In any case the mutation rate can not be too large, otherwise there would be only a small role for inheritance, and surnames could not work well as a marker connecting people with family linkages. Obviously, it is in the nature of surnames to be inherited. That is what differentiates them from first (given) names. Just to get an idea of the magnitudes of surname mutation, in Spain in the year 2001 there were 1570 applications to change the surname.[22] Out of this number 1426 were granted. Assuming that 2001 is a representative year, and a population of around forty million with a live expectancy of around 70 years, this amounts to a *mutation rate* of around 0.0025. Thus, in Spain around 99.8% of the individuals dies carriying the same name than his father. This mutation rate is probably as low as it can get, as relative to other western countries Spanish law is notoriously restrictive in allowing surname changes (see the discussion on Spanish names in section 4.1.1). We will use 0.2% as our baseline number for mutation rate, stablishing the frequency of lineage births, in our numerical analysis. In any case, we check that our results are robust to changes in this number.

Another source of new surnames in a population is migration. In the model we abstract from this possibility because in the empirical part we clean our data from migration and ethnicity issues. In any case, it is easy to extend the model to allow for migration to be a source of new surnames.

[20] See for instance Manrubia and Zanette (2002).

[21] We do so because it is akin to genetic mutations, and it might be useful to draw comparisons with between our point and population genetics.

[22] Data from the Office of Public Records (*Registro Civil*). We thank Manuel F. Bagüés for providing us with these data.

The Last of the Mohicans

Thus, the mutation rate while being positive is a very low number. This implies that the death of the last holder of the surname across the male line is overwhelmingly the most important of the two possible causes of the death of lineages. Thus, we need to specify a fertility process.

We will assume that an agent may have male descendants with probability $Q \in (0,1)$. Conditional on having sons, he has a fix number $M \in \mathbb{N}$ of them. The expected number of children being $E = Q \times M$.

In principle, we allow these numbers to depend on the economic status of an individual (so rich agents may have more, or less, fertility than poor ones), but for the moment we consider the fertility parameters to be exogenous and identical across all individuals. For simplicity we also assume no systematic population growth, so $E = 1$; in the aggregate the number of agents follows a random walk.

If you are the last male member of your lineage, the probability that your lineage disappears in one generation is $1 - Q$. The probability that it disappears in two generations is $1 - Q + Q \times (1 - Q)^M$, and so on.

Irrespectively of the number of members of the lineage who are contemporaneously alive, there is a positive probability of lineage death. Of course, this probability is larger the less people have a surname. If the distribution of surnames is such that there are many (unfrequent) surnames, it is easy to see that many lineages will have to die. If in addition there were no birth of new surnames, ($\mu = 0$), then the number of surnames would become very small in the medium run. The surname sizes would be very large, and the people holding a same surname would not be particularly likely to share family links. The surnames could not be informative. Thus, in a world with were lineages may die, the mutation rate insures an inflow of new surnames in the distribution, and the subsistence in the population of a large number of surnames with relatively few holders.

The new lineages enter the population with a minimum size, so they are the most likely ones to disappear. Still, some of them will survive, even expand. Many will remain for a long time in the population maintaining relatively small sizes. People holding these unfrequent surnames are overwhelmingly family related. From them we extract information.

The economy is in steady state if the joint distribution of surnames and income is constant. By this we do not mean that the proportion of people with any particular surname is constant, but that the proportion of surnames with any particular frequency is constant. This is, in steady state some surnames that were relatively large will became smaller, but other surnames will take their place by becoming relatively large.

3.3 The Information Contained in Surnames

The distribution of surnames in the steady state depends on fertility parameters (Q and M), and the mutation rate μ, but it does not depend on any variable of the income process. This is because in this model surnames not only not affect income, but *are not affected by it neither*. The process of birth and death of lineages (and surnames) is completely independent from both ρ and V_e.[23] The income process does not affect the probabilities of death/birth

[23] This will be different in section 3.4.

of lineages, and consequently the surname and income distributions are determined by different set of parameters ($\{Q,\ M,\ \mu\}$ and $\{\rho,\ V_e\}$ respectively). Still, the ICS is not an obvious object, as it is determined by *both* sets of parameters.

The model is perfectly akin to a genetic model where the reproduction is asexual and the genetic material has no survival value. Specifically, the model would be identical if we were talking about the junk DNA in the mitochondria or the Y chromosome. Junk DNA does not code for proteins, so it has no effect in the differential survival and reproductive chances of the individual. It is like surnames, having no incidence on the chances of the agents becoming rich or poor. Mitochondrial or Y-chromosome DNA does not reproduce sexually, so there is no mixing. Except for the possibility of a mutation, it is inherited in exactly the same manner that your mother (the mitochondria) or your father (if you are a male, the Y chromosome) had it.

Surnames are informative because they act as markers. Once a mutation appears, it will remain in the population unless the number of individuals who have it all die without descendants. Among the lineages that appear at a certain time a fair number will remain alive for quite a while. A few will increase in size substantially and perhaps once they have grow its mere size will insure that they will not die (remember that expected growth of the lineage is zero, $E = 1$). Some will disappear quite fast.

For us the more interesting ones are the surnames that remain a long while in the population, but without growing much. Those surnames are of people who has high chance of being related among each other. Their income is going to be more homogeneous than the income of the population as a whole. Thus, knowing the surname of the individual informs about his income.

Of course, for frequent surnames there is little information to extract from the surname. It is in the seldom heard surnames where you get the information. It is irrelevant that for many agents the surname says little, because *there is a large number of agents whose surname is unfrequent, and thus informative.* The distribution of surnames is going to follow very closely a power law, so that the huge majority of the surnames are going to have very small frequencies; the sum of all of them is bound to be a considerable percentage of the population.

The more inheritance there is, the more that you know about the individual *relative to how much you would know about him if you did not know his surname.* Simply because the family link (more specifically the link of belonging to the same lineage) is partially captured by the surname.

We proceed by assuming a set of parameters $\{\rho,\ V_e,\ \mu,\ Q,\ M\}$, and assume an initial population of one million individuals. We run a simulation of the workings of this economy during 120 periods. As initial conditions we give a random set of surnames and incomes.[24] In order to be sure that we arrive to steady state we let the economy run during the initial 100 periods. Each period the model generates a census, that itself is used as the base of the census of the following period. From the period 101th and on, we collect the census thus generated and on each of

[24] We have done the same exercise with very different initial conditions, like only one surname, or each individual a different surname, etc. The results are robust to these manipulations.

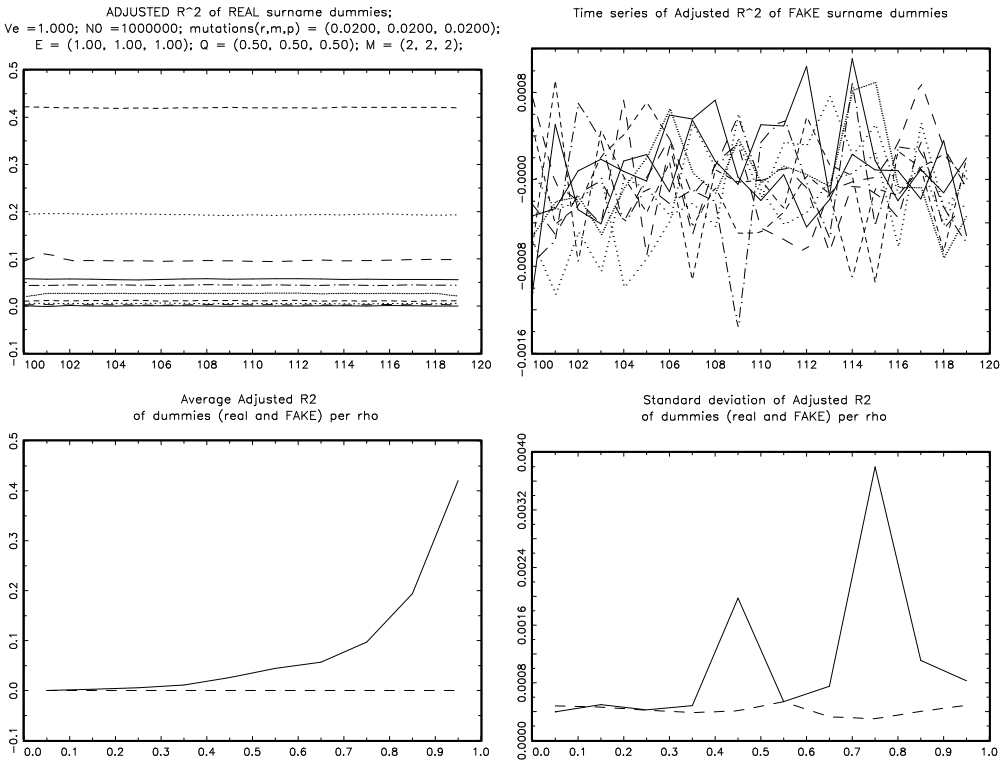


Figure 1: Surnames are informative, and their informational content increases with the degree of inheritance.

them we run two regressions; one with the real surnames, and another with fake–surnames. They are respectively:

$$Y_t = b' S_t + u_t \tag{3}$$

$$Y_t = b' F_t + u_t \tag{4}$$

where Y_t is a vector with incomes. S_t is a matrix of surname dummies. It has as many columns as different surnames exist in the surname vector of the simulated economy, A_t. F_t is a matrix of dummies of fake–surnames. Fake–surname dummies are a random extraction of a distribution defined by S_t.

We use the R^2 of both equations to analyze and interpret the information that surnames have.

Figure 1, captures the essence of the message of this paper. It is the graphical representation of the model. It shows that surnames are informative, and the more inheritance there is, the more informative the are. The parameters that we use for this simulation are $V_e = 1, \quad Q = 1, \quad M = 1, \quad \mu = 0.2\%$, and values of ρ ranging from 0.05 to 0.95. The initial number of agents was one million.[25]

Next, we explain the four panels of the figure. Notice that the scale of the coordinates differs in different graphs.

- **NW corner**, each line is the time series of the adjusted R^2 of the *true* surnames for a certain value of ρ (regression (3)). In the vertical axis we plot the R^2, and in the horizontal axis we plot time.

  What is remarkable is that

[25]This is a representative simulation, we have run many more. All of them provide the same results.

1. The time series are all flat: the adjusted R^2 is constant for a certain value of ρ.
2. The adjusted R^2 is higher the larger is ρ.

- **NE corner**, each line is the time series of the adjusted R^2 of the *fake–surnames* for a certain value of the inheritance (regression (4)). In the vertical axis we plot the R^2, and in the horizontal axis we plot time.

  What is remarkable is that

  1. The time series are not flat. The adjusted R^2 are not constant given the value of ρ.
  2. The adjusted R^2 are independent of the value of ρ.
  3. The value of the R^2 is very low for all values of ρ.

- **SW corner** plots the average adjusted R^2 of each value of ρ both for the real and the fake–surname distribution. In the vertical axis we plot the average R^2 of the time series, and in the horizontal axis we plot the value of ρ.

  What is remarkable is that the real surnames have more information the larger is the value of the inheritance, while the fake ones do not. It shows that *surnames are informative, and they are more informative the more inheritance there is.* This is the main result of this section. It will allow us to interpret a larger value of the informative content of Surnames as more inheritance, and consequently, as less mobility.

- **SE corner** plots the variance of the time series of the R^2 for each value of ρ and both real and fake surnames. In the vertical axis we plot the standard deviation of the time series of the R^2, and in the horizontal axis we plot the value of ρ.

  The remarkable thing is that it is so low in both cases. This is particularly informative in the case of real surnames: given parameters (and in particular, given ρ), the R^2 does not fluctuate over time. The variance is perhaps slightly higher for large values of ρ (when the income process approximates a random walk), but still in that case it is a very low number.

In appendix A we show that this result is invariant irrespectively of the conditional variance of the income shocks, the variance of family size or the mutation rate.

Thus, the key result: the informative content of surnames increases with inheritance. Ceteris paribus, if we measure a larger ICS, we should infer that there is more inheritance, less intergenerational mobility. This fact is independent from the degree of conditional variance, the mutation rate and the fertility parameters.

The reason is that the birth and death of surnames insures that in steady state there are many surnames with low frequency. The individuals sharing one of this unfrequent surnames are very likely to be close family relatives, sharing a recent common ancestor.

Notice that the critical thing is to determine that the distribution of surnames is skewed, and that for a large percentage of the population to share surname approximates to share recent blood lines. Once this is provided, the rest follows.

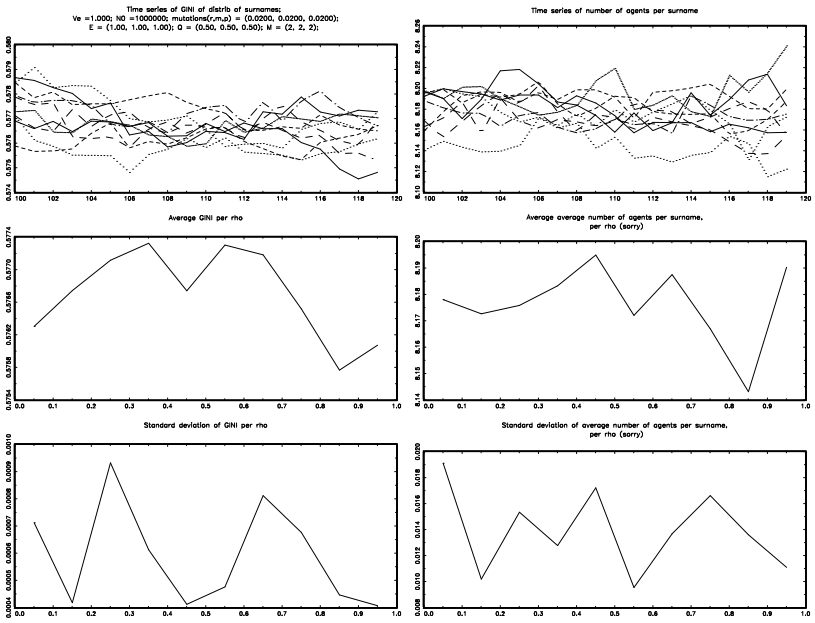


Figure 2: Surname Distribution in Baseline. The distribution of surnames is independent from the income process

An important consequence is that if there were an increase in the value of ρ, there would be an immediate increase in the value of the ICS. The reason is that a larger ρ would make more similar the outcomes of individuals sharing blood lines, making their surnames (if unfrequent) more informative; increasing the ICS. This feature will allow us to study the evolution of intergenerational mobility across time.

3.4 The Surname Distribution as a Function of the Income Process. Information in the Frequency of Surnames

In section 3.3 we have considered that rates of death and birth of lineages were independent from the economic status of the members of the lineage. This has two immediate consequences.

(1) That the surname and income distributions are independent of each other. A larger ρ affects the ICS, but it does not affect the distribution of surnames in the population. We see this in figure 2, which shows that in the aforementioned case the distribution of surnames is not affected by the degree of inheritance.

The three panels of the left show (from top to bottom) the time series of the Gini index of the distribution of surnames for different values of ρ, the average value of the Gini index of the distribution of surnames and its standard deviation. The three panels on the right show the same (time series, averages and standard deviations) for the number of agents per surname. We use the Gini index and the number of agents per surname, because these two moments completely characterize a geometric distribution, which itself is a very good approximation to a surname distribution.

It is immediate to observe, and not surprising given our reasoning above, that with the parameters from our

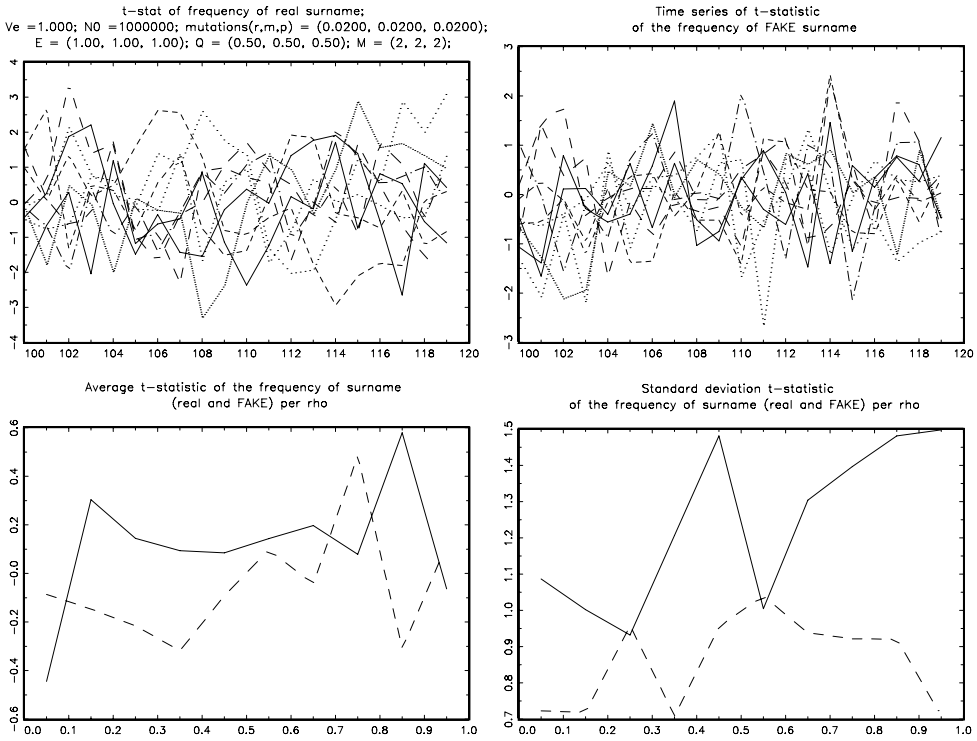


Figure 3: No frequency effect without fertility differences

previous simulation the distribution of surnames is independent from the value of ρ, as the birth and death of the lineages have nothing to do with the status of the members of the lineage. It is also reassuring to observe that the standard deviations of each of the time series is so small: the economy is at a steady state.

(2) That it is the specific surname, not its characteristic, that has information. In particular the frequency of the surname contains no information on its holder.

In order to capture the information in the frequency of the surname we run another set of regressions on the census data. As before both regressions have on the left hand side a measure of economic wellbeing of the individual.

The first of the regressions has in the right hand side the *frequency* of the surname of the individual. If the t–statistic of this variable is sufficiently high, we can reject the hypothesis that the frequency of the surname has no information on the economic wellbeing of the individual. Notice the difference with what we were doing before. Here we do not say if the specific surname has information, but if the *frequency* of the surname has it. If the surnames $C3PO$ and $R2D2$ have the same frequency, they have the same value in the right hand side, even if the $C3PO$'s and the $R2D2$'s have absolutely no family links.

In order to make sure that skewness and the usage of censored variables do not bias our results, we follow an identical strategy that we followed when dealing with the ICS. We run a second regression that places in the right hand side the frequency of fake–surnames. Thus, the frequency of surnames is informative only if the parameter of frequency is significant in the first regression (with real surnames) while in the second (with fake–surnames) it is not.[26]

[26]The specific regressions are $Y_t = c + b \times s_t + u_t$ and $Y_t = c + b \times f_t + u_t$ respectively. With s_t and f_t being vectors assigning to each individual the frequency of his real and fake surname respectively.

Figure 3 shows that, with the parameters that generate figure 1, the frequency of the surname has no informative content. It is very similar to the figure 1, but in the SW corner (where before we plotted the R^2 of equation (3)) we now plot the t-statistic of frequency in the regression with real surnames. In this case the value of the t-statistic is in general lower than two in absolute value: the frequency of the surname is not informative.

We will now show that if the probabilities of death or birth of a lineage depend upon the relative income of their holders the distribution of surnames will in general not be independent of the distribution of income. In particular not only the surname, but also its frequency, will carry information. The reasoning of how this information is generated is somehow more involved than in the previous section. We first extend the model to allow for this possibility, and then interpret the results.

We extend the model so that there are 3 income groups *rich, poor and middle class*, the first two representing the 20% richer and poorer respectively.

We assume that the probability of having children and the number of children that you would have in such an event are different across the different groups. $\{Q_r, Q_m, Q_p\}$ represents the probability that rich middle class and poor people will have sons, and $\{M_r, M_m, M_p\}$ the number of children that each would have in such a case.

In order to insure that the total population does not have systematic growth we impose $\frac{1}{5} \times Q_r \times M_r + \frac{3}{5} \times Q_m \times M_m + \frac{1}{5} \times Q_p \times M_p = 1$, but otherwise the expected number of children ($E_j = Q_j \times M_j$) can be different for different groups.

We also assume that the mutation rate of the surnames might be different depending on if the individual belongs to the rich, medium or poor classes. $\{\mu_r, \mu_m, \mu_p\}$ being the respective probabilities of mutation.

There are three dimensions in which the differences between the groups may lead to differences in the distribution of surnames across income groups. (1) Differences in the probability of having at least a male descendant (Q_j); (2) differences in average fertility between income groups (E_j), and (3) differences in the probability of starting a new lineage (μ_j). We look at their effects separately.

3.4.1 Differences in the probability of having male descendants.

There are differences in Q_j, and we refer to them as the *"hereu effect"*. They have direct incidence in the probability of survival of the surname, but have no effect on the probability that the size of the surname grows or decreases.[27]

Imagine that we have a society where rich individuals and poor individuals have the same expected number of children, but the rich have them with certainty while the poor have them stochastically ($Q_r = 1$, $M_r = 1$; $Q_p = \frac{1}{2}$, $M_p = 2$). Then the probability of survival of your lineage if you are rich is one, while it is only $\frac{1}{2}$ if you are poor. Say that there are 100 mutations among the rich and another 100 among the poor. After one period the

[27] In traditional Catalan society the property of the family farm was inherited by the oldest son (not daughter) who was called "hereu" (inheritor). The other children would typically be compensated by other forms of education (like becoming a priest), or by dowry, or with cash. This institution had important consequences relating to average size of farms (and avoiding that they became too small), but it had the drawback that you needed of a son if you wanted your farm to remain in your lineage. Somehow it seems that old time Catalan farmers did want their farms to remain on their lineages, so they wanted sons, only daughters would not suffice. The way to insure this is to keep having children at least until you hit a boy. This means that the probability of your lineage dying was very low if you had a farm, because at least you would have a male child who would continue the lineage alive. If you had no farm you would be less obsessed with the male child thing, and thus the probability of disappearance of the lineage would be larger.

mutations of the rich will still be there, while only 50 of the new surnames of the poor will be there (and each with two people).

Notice that the key mechanism here is that surname death rate is different for different income groups, while the inflow is the same in all of them. The groups with a larger survival rate are bound to accumulate a larger number of unfrequent surnames.

In figures 4(a), 4(b) and 4(c) we show the results of a simulation where individuals have all the same expected number of children ($E_j = 1 \quad \forall j$) but where the rich *always* have a male child ($Q_r = 1; \quad M_r = 1$), while for the middle class is ($Q_m = \frac{1}{2}; \quad M_m = 2$), and the poor have an even higher variance of family size ($Q_r = \frac{1}{4}; \quad M_r = 4$).

We learn three things from this simulation:

1. It is clear from figure 4(b) that the frequency of the surname is informative (more frequent surname, less income), and *that the more inheritance there is, the larger the absolute value of the t-statistic of the frequency.* This second feature is perhaps the more important.

   Imagine two mutations. One taken place among the rich, where the lineage *Richmanson* is born. The other among the poor, the lineage *Poormanson*.

   If the degree of inheritance is large, the lineages that started in a mutation among the rich will stay alive for a very long time, *and they will have a small frequency all that time.* This is a consequence of the income persistence being large: they children of Mr *Richmanson* remain rich, which insures that they will have children, and the survival of the lineage. In spite of being very certain that the surname will not disappear, it is also unlikely that the surname will grow. This is because the rich have for sure a son, but they are unlikely to have many.

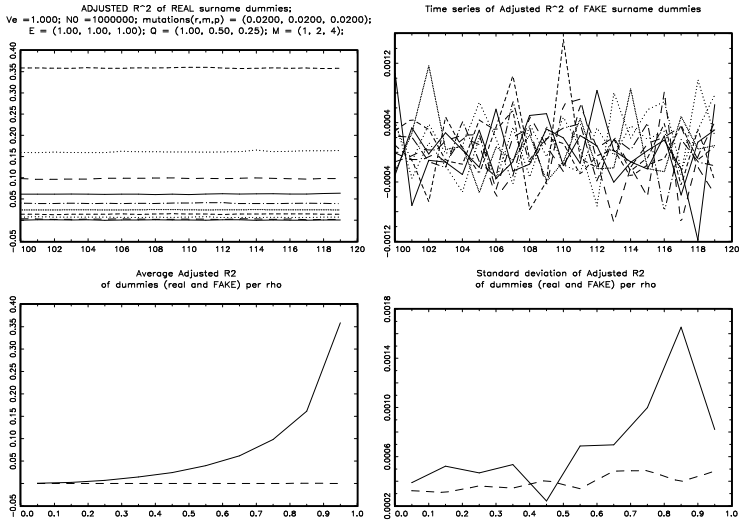
   It is on the other hand very difficult that the lineage *Poormanson* will survive *and remain unfrequent.* Mr *Poormanson* and Mr *Richmanson* have the same expected number of sons, but Mr *Poormanson* has a larger variance. He is more likely to have no sons (with which the lineage would disappear), but if this does not happen he will have more sons than the *average* rich guy. As a consequence the unfrequent surnames will tend to belong to rich people and only seldom you will find a poor person with an uncommon surname.

   If the degree of inheritance were smaller you would not get such a large effect of frequency, as lineages that started rich have a large probability of becoming poor (and then disappearing). There will be less concentration of rich people among the unfrequent surnames.

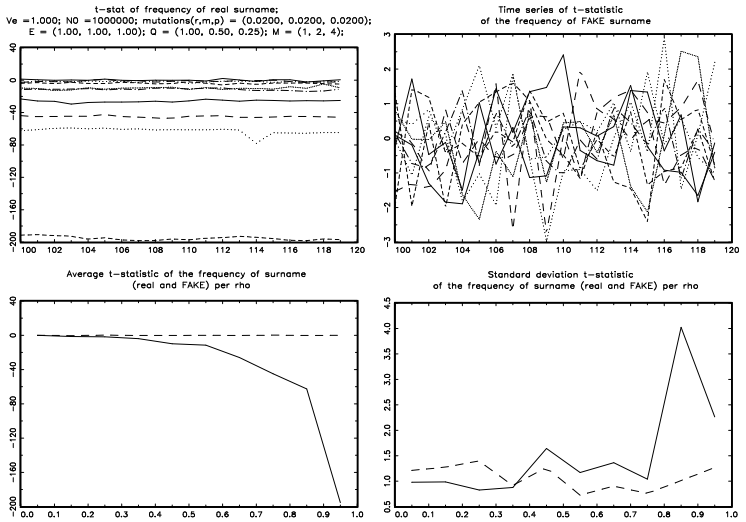
2. In figure 4(c) we observe that the distribution of surnames actually does depend on the income process. This is very different of what happened when income did not affect the lineages birth/death rates (figure 2)

   The distribution of surnames, being very well approximated by a geometric distribution) is characterized by the number of people per surname and the Gini index of the surname distribution.

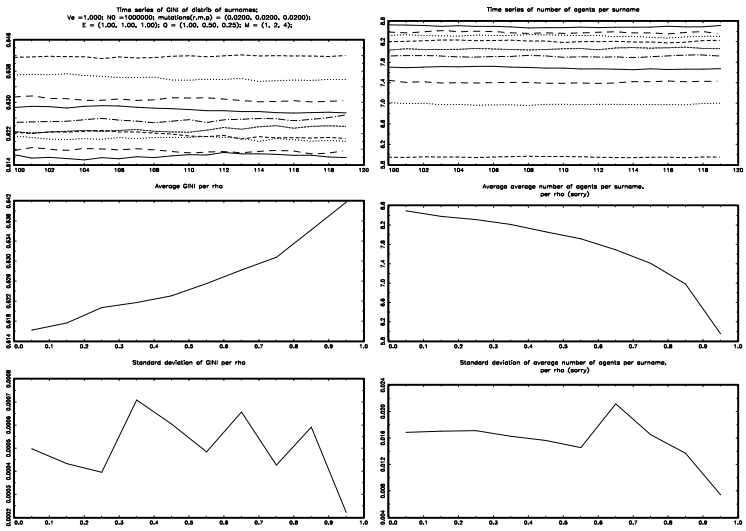
   The number of agents per surname decreases with the degree of inheritance, while the Gini index increases.



(a) Adjusted R2



(b) Hereu frequency



(c) Hereu distribution

Figure 4: "Hereu Effect": Differences across socieconomic groups in the probability of survival of surnames.

The reason for the first is that if inheritance is very important (high ρ) rich agents tend to have one of a kind surname. Once they get the surname it only changes if there are mutations, but it does not grow either.

The Gini index is very high because a few surnames (the ones of the poor) hold a very large percentage of the population (those non rich). The distribution becomes very skewed.

3. Finally, notice in figure 4(a) that all what we argued in the previous section is still true. When conditioning on the specific surname, and thus approximating family relationships, the ICS increases with ρ in the same manner that it did before. The mechanism of grouping siblings together (surnames being an informative partition of the population, as it relates to family) is still working.

   This will be important for our empirical approach: irrespectively of the t-statistic of frequency, we can infer the degree of mobility by looking at the ICS alone.

3.4.2 Differences in average fertility.

Differences in average fertility[28] between income groups (differences in E_j) are substantially more involved to study. The reason is that, in addition to affecting the probability that the lineage disappears, they alter the probability that the surname will either grow or decrease.

Differences in E change also the relative population holding the surnames. Imagine that the rich individuals have a larger E, then not only they have a lower probability of their lineage disappear (which makes the rich having unfrequent surnames) but also induces the surnames of the rich to become large relatively fast (with which the rich would tend to have frequent surnames). The key to determine if an odd surname is going to indicate wealth or its absence is the play of M's with Q's.

Notice finally that by inducing differences in reproductive patterns between rich and poor agents, the unconditional distribution of income in the population will *not* be the same that the unconditional distribution of (2). For instance, if the average fertility of the rich is relatively large, then a positive income shock in one generation will transmit to more individuals (on average) than a negative one of the same magnitude. The income distribution in that case would switch toward higher levels of average income.

In appendix B we present the result of simulations with differences in E. What we can learn from it is that the ICS, as defined in previous sections, maintains its monotonous relationship with inheritance, as surnames are still approximating recent common ancestry. The relationship between frequency and inheritance is very complex (sometimes positive, sometimes negative). The relationship between ICS and inheritance is stable, clear, always increasing and positive. This is the reason why we are going to focus our empirical approach on the ICS.

[28] In any case we want to remark that we refer to *males*. We refer to the average number of (reproductively capable) *male* offsprings that a *male* adult has. The correlation between "male fertility" and income can go in exactly the opposite direction than female fertility. Educated females are known to have less children than uneducated ones, but that is not necessarily the case for males. It is not uncommon for successful males to have children with more than one female; either by re-marriage, polygamy or out-of-wedlock relationships.

3.4.3 Differences in the mutation rate.

It is straight forward to see that frequency of the surname has information on the income of its holder if there are differences in the rates of birth of lineages associated to income differentials, and for essentially the same reasons than above.

Say that the mutation rate were larger among rich than among poor people. In such a case the inflow of new lineages would be larger among the rich than among the poor. Unfrequent surnames would tend to belong to the rich, because the new surnames are by definition unfrequent, and they belong to the rich.

Of course, if the mutations were to take place mostly among the poor, the opposite would happen, and surname frequency would be correlated with higher income. It complicates matters the fact that it is difficult to know if mutations happen more often among the rich or among the poor.

On one hand there are reasons to believe that surname mutations are more likely to occur among the rich. The number of hyphenations, and even the sheer length of the surname are probably associated to higher income, as rich people may like to signal their status through their surnames. This could well work in a form akin to first (given) name allocation. It is well known that the better-off choose names for their offspring that are new, and different from the most common ones in their society.[29]

On the other hand, migration is probably the most common form of introducing new surnames into a given population, and in our context it could be interpreted as mutations. Emigrants tend to be poor. They also tend to have surnames that from the point of view of the recipient population are unusual. Most often they are simply unique because the possibility of mutation is very likely to increase a lot as a direct consequence of migration. Transliteration of foreign scripts and alphabets, orthographic and phonetic differences between countries all this adds up to generate new surnames that are new not only from the point of view of the recipient populations, but also in the original population of the migrant.

An additional complication is that the relationship between migration and mutation depends on the difference between the surname distribution of the origin and recipient populations. A migrant from Morocco to Spain is more likely to introduce a new surname in Spain than a migrant from Ecuador. In the same manner, if migration happens between regions that are "close" from a surname distribution point of view the number of observed mutations will be lower than if the regions are far apart.

Thus, we finish this subsection remarking that: (1) there are reasons to expect that the surname distribution should be a function of the income process, and (2) that characteristics of the surname (like its frequency) and not only the specific surname should be informative on the individual's economic wellbeing. (3) In spite of it, there are many possible causes of it, and they are difficult to interpret, as they move in different directions. (4) Consequently, it is better to restrict ourselves to the study of the ICS when studying matters related to inheritance.

In the empirical section we do so. Nevertheless, for the sake of completeness, we report that in Catalonia, and among non emigrants, unfrequent surnames are associated with more economic wellbeing.

[29] See Fryer and Levitt (2004) and Levitt and Dubner (2005).

3.5 Assortative Mating and Surnames

Assortative mating refers to the fact that people are likely to marry individuals with similar characteristics to themselves.[30] We will consider two dimensions where there can be assortative mating, and look at the consequences that an increase in the degree of assortative mating in each of these might have on the informative content of surnames. First we will look at assortative mating with respect to income and secondly (in section 3.6) to ethnicity.

It may seem intuitive that assortative mating according to income may generate by itself informative content of surnames, because it may give "organization" to the distribution of surnames. One could think that if the rich marry the rich and the poor marry the poor, then the surnames of the rich and of the poor will be kept different from each other. This intuition is wrong, and deeply misleading. Assortative mating in ethnic characteristics may induce surnames to be informative about the ethnicity of the holder (we will see this in more detail below), but the process of mating *by itself*, in absence of the birth/death process that generates the surname distribution, would not generate any informative content of surnames. This is because the degree of assortative mating does not affect in any direct way the distribution of surnames in the population.

The reason[31] is simple: surnames are inherited through the male line, thus your surname is completely independent from the surname of your mother. You are not in the lineage of your maternal grandfather. It does not matter how your mother was called, and it does not matter why she mated with your father, and it does not matter if she was rich or poor: her surname will not be passed *by you* to anybody. It could as well be that females had no surname.

Females do not pass their surname, but they pass along other economically meaningful characteristics that affect how do you fare in life. So, the degree of assortative mating partly determines the composition of households, and the amount of inheritance that individuals have. We will show next that in our context a larger degree of assortative mating is isomorphic to a larger correlation between the incomes of fathers and sons, which in the previous sections we have associated to the value of inheritance (ρ). Thus, there exists an effect from the degree of assortative mating into the amount of information that surnames contain: assortative mating generates high values of ρ, which in its turn generates information in surnames.

We develop here an extension of the model that includes assortative mating as one of its determinants. Our main result is that (more) assortative mating increases the correlation of the income of sons and their fathers *even if the correlation between the income of the sons and the average income of their family is unchanged.* Consequently, the informative content of surnames increases because when we look at the male line an increase of the degree of assortative mating is isomorphic to an increase in the correlation of the income between fathers (not mothers, not families) and children (not the household set up by the child).[32]

[30] Some of the existing literature on mobility (see Lam and Schoeni (1993), Chadwick and Solon (2002), Ermisch, Francesconi, and Siedler (2006) and Holmlund (2007))) has considered the contribution assortative mating among *the child and his (her) spouse* to the intergenerational mobility of the *child's household*. Our approach is quite distinct, as we consider the mating process of the *parents* of the child under consideration and the *child's (own)* mobility. Even if some of the existing papers sometimes consider the correlation between the income of the child and both his parents, to our knowledge we are the first ones to consider the assortative mating of parents.

[31] We are thankful to Melvin Coles for this insight.

[32] In what follows we refer to the characteristic that determines the appeal of individuals as mates, as "income". We do this with the same caveats than in the previous sections.

3.5.1 Set up of an extended model.

First we explain which are the exogenous parameters and how they interact. Then, we relate the model to the one that we have developed in the previous sections.

Inheritance. Agents are either female or male, they form households of two individuals (one from each gender). They have children, and their children's income is partially determined by the characteristics of the household. We denote as x the deviation of the income of a female from the mean, as y its equivalent for males , and as z the average income of the household. With r we denote the degree of "inheritance", the correlation between the income of a child (once he/she grows up) with the (average) income of the household where he/she grew up. Finally, e is normally distributed noise, uncorrelated across time and across agents. Its variance (denoted σ_e) is the variance of the income of an agent conditional on the average income of the household where he (or she) grew up.

$$x' = r\ z + e\,;\quad y' = r\ z + e \tag{5}$$

where a prime signifies that the individual belongs to the generation of the children. We assume the income process to be identical for both genders. So, the unconditional distribution is also identical for both.

Mating process. Males and females mate. They mate, but they are discriminating. They want to mate only with rich people (as that would give their children larger inheritance). Unfortunately, they are restricted to mating with one person, so it is not that Bill Gates (and only him) mates all the girls; nor that only the Queen of England (and only her) mates with all the boys.

The mating process is notoriously difficult, and we do not model it; suffice to say that the members of a couple are expected to have no difference in incomes. The difference in their incomes is determined by a white noise process u:

$$x = y + u\,;\quad u \sim N\left(0, \sigma_u\right) \tag{6}$$

It is useful to interpret this as if the income of the husband were a signal on the income of the wife. This signal has a certain precision, which is larger the smaller is the variance of the mating noise (σ_u). We will say that "there is more assortative mating" whenever the variance of the difference between the incomes of a couple is smaller; whenever the income of the husband is a more precise signal on the income of the wife.

Steady State. Abstracting from lineage and surname determination, the only structural parameters of the economy are r, σ_e and σ_u. They determine the distribution of households incomes and, through it, the distribution of income for all the individuals. Insofar $r < 1$ all the processes are stationary, and this insures the existence of an steady state with constant distributions.

In steady state the *unconditional* distribution of income for the individuals (which is identical for both men and women) has a zero mean and a certain variance V. This distribution is endogenous, the value of V depends

(in a non obvious manner) on r, σ_e and σ_u.

$$x \sim N(0, V) \; ; \quad y \sim N(0, V) \tag{7}$$

3.5.2 Solution of the extended model.

The model of section 3.2 refers only to the relationship between fathers and sons. It does *not* refer to the household or the mother. This is because surnames are inherited *only* through the male line. Thus, they are informative only about the relationship between individuals and their fathers.

To understand the relationship between the informative content of surnames and assortative mating we need to be able to determine how assortative mating affects the relationship between the incomes of father and son, as different from the relationship between the incomes of parents (which includes the mother) and their children.

We can rewrite the income process generated from r, σ_e and σ_u in a manner that is analogous to the one that we have developed in section (3.2):

$$y' = \rho \; y + \varepsilon \tag{8}$$

where y' is the income of a son (not a daughter), y is the income of a father (not a mother, not the household), ρ is the correlation between the income of father and son, and ε is stochastic noise whose variance (σ_ε) is the variance of the income of an individual conditional on the income of his father. Both ρ and σ_ε are endogenous, and a function of r, σ_e and σ_u.

We want to remark that when addressing issues on intergenerational mobility the proper measure of the value of inheritance is ρ; not r. The reason is that ρ establishes a relationship between two comparable elements (the incomes of child and father), while in equation (5) the two elements are just not comparable. It explains the income of one individual with the the consolidated income of the household where he grew up. The determination of the second a one has gone through the noisy lottery of mating, while the LHS has been exempted of such a noise.

It would be coherent to use an equation relating the consolidated income of the household when one individual grew up with the consolidated income of the household that the individual eventually establishes; as itwould include the noise due to mating in both sides. This is the path that has been taken by some of the literature relating assortative mating and mobility. We have instead opted for ρ and equation 8 because (1) it is perfectly coherent, (2) it is the one that is estimated in the vast majority of the literature, and (3) it happens to be intimately related to the process of surname diffusion.

Thus, we want to know how the degree of assortative mating affects ρ and σ_ε, because they determine the degree of intergenerational mobility.

In appendix C we prove the following result:

Result 1 *In steady state ρ and σ_ε are the unique solution to the following system of equations:*

$$\left[\rho - \frac{r}{2}\right] \times \left[\frac{1-\rho^2}{r-\rho} - \frac{r}{2}\right] = \frac{\sigma_e}{\sigma_u} \tag{9}$$

$$\left(\frac{\sigma_\varepsilon}{\sigma_e} - 1\right) \times \left[\frac{1-\rho^2}{r-\rho} - \frac{r}{2}\right] = \frac{r}{2} \tag{10}$$

At the solution: $\rho \in \left(\frac{r}{2}\ ,\ r\right)$, $\frac{d\rho}{d\sigma_u} < 0$, $\sigma_\varepsilon > \sigma_e$ *and* $\frac{d\sigma_\varepsilon}{d\sigma_u} > 0$.

A large degree of assortative mating (a large precision of the mating process, a small value of σ_u) translates into a large value of ρ.

The more assortative mating, the larger the correlation between the income of father and son; even if the correlation between the income of sons and the joint income of their parents is kept constant. The intuition is straight forward: the more assortative mating the more information has the income of the father on the income of the mother. Both father and mother contribute to the characteristics of the son. Thus, the more the income of the father can explain the income of the mother, the more it also explains of the income of his son. The more assortative mating, the less intergenerational mobility.

Surnames are passed exclusively along the male line. They do not provide with any information about the mother directly. They may provide information about what is inherited from the mother only insofar the characteristics of the father contain information about the mother. Thus, the informative content of surnames depends only on the correlation and conditional variance of the incomes of fathers and sons (on ρ and the variance of ε), and it does so in a manner that we already understand. Nevertheless, it depends on the level of assortative mating because ρ does depend on it. A mating process that is more assortative translates into larger values of ρ, which itself translate into a larger ICS. [33] We will use this to interpret our empirical findings.

3.6 Ethnicity and Assortative Mating

We consider now the possibility of the existence of another inheritable dimension in which individuals differ in addition to their surname and their income. Lacking a better word, and in order to stress that this characteristic is inheritable, we call it "ethnicity".

We assume that ethnicity affects directly the income of the individual. This can be because of discrimination, or because of different ability, or for any other reason. We do not study why ethnicity has effects on income, we only try to measure these effects. Our definition of "ethnicity" is wide and encompassing. For instance, castes in India would certainly be "ethnic" in our treatment. We define as "ethnic" as anything that is inheritable, that is not *directly* related to the income of the parents and that has effects on the income of the children.

Our next point is very intuitive: if there is assortative mating in the ethnic dimension (or if the "ethnic" characteristics are inherited only through the male line), then surnames will be informative *not only because they*

[33] As we have seen the ICS depends also on the conditional variance, even if it is not very elastic to it. In any case the ICS would increase as a consequence of a decrease of σ_ε, and thus an increase of the degree of assortative mating would have the same qualitative effect through this channel.

are informative about family linkages, but also because they inform about the ethnic status of the individual. The surname partition of the population is informative not only about family links, but also about the ethnic partition of the population.

In general, the more that the "ethnicity" of an individual can be predicted by knowing her father's ethnic group, the more that surnames capture ethnicity. Again, this could be either because (1) "ethnicity" is passed along through the male line or (2) because there is a large amount of assortative mating along the ethnic dimension.

In order to see the first case, consider again (like in subsections 3.2 and 3.3) a world with only males. The difference now being that in addition to surname and income, people has a characteristic called "ethnicity" that is perfectly inherited through the male line. The income of an individual is now:

$$y_t = \alpha_{eth} + \rho y_{t-1} + \varepsilon \tag{11}$$

where α_{eth} is the effect of the ethnic group on income. For simplicity let's assume that there are only two ethnic groups (w and b) with $\alpha_w > \alpha_b$. Ethnicity and surname are passed along together. If there is a difference in the surname distribution of both groups, the surname is going to be informative on the ethnicity of the individuals; and, via this mechanism, on their income. A surname that indicates that you are likely to be b would also indicate that you are likely to be poorer, as you are likely to have a low α.

Upon reflection it is easy to see that eventually the distribution of surnames is bound to be different across ethnic groups. The reason is that the process of birth and death of surnames is independent across the ethnic groups. Thus, their surname distributions will drift apart as time passes by, even if the the process starts with identical surname distributions across the groups. The pool of surnames in the two groups will have different mutations, and the surnames that will die will be different.

Take any two groups of males, say *red* and *green* and let's look at the lineages that emanate from them. As time passes by the distribution of surnames between the descendants of the two groups becomes more and more different. The reason is that the new surnames that arise in both groups are bound to be different (as a consequence of the randomness of the mutations), and the surnames that die are also going to be different in the two groups. Thus, after a number of generations have passed the surname of an individual will be informative on whether he belongs to group *red* or group *green*. This is independent from the possible characteristics differentiating both groups. These characteristics may go in one direction, or the other, or be inexistent. The lineages from individuals randomly assigned today to groups *red* and *green* will eventually have different distribution of surnames. Of course to have a different distribution to start with will make things go faster, but the death/birth process of surnames insures that the surname distributions will drift apart. This is true for *any* group of males. Thus, it is also true for the ethnic groups w and b, for which there is an ethnic difference in the distribution of income.

Surnames capture ethnicity *only insofar* the ethnic character is transmitted across the male line. Thus, it is easy to see that large degrees of assortative mating are necessary for surnames to contain information about ethnicity. Normally ethnicity is not inherited only through the line of the father. Insofar the "ethnic" character

comes from the mother if there were no assortative mating the father's surname would not correlate with the mother "ethnicity".

Judaism is a particularly clear example, as the attribute "to be Jewish" is inherited only through the maternal line. Abstracting from the case of conversion, you are Jewish only if your mother was Jewish. If people were to marry completely at random, neither income nor ethnicity being a concern, then after a short length of time surname would not be informative on if its holder is or it is not Jewish, because the surname informs about the father, and not about the mother.

For a surname to be informative on if the holder is Jewish (or of any other ethnic group) the ethnicity of the father, *and not the one of the mother*, needs to have predictive power on the ethnicity of the child. This, of course, is what happens when there is a large degree of assortative mating across ethnic lines. If a Jewish person always marries another Jewish person, then the ethnicity of the child and the ethnicity of the father are always perfectly correlated, and thus the surname distribution of Jewish people would be different from the surname distribution of non-Jewish people.

The story can be easily extrapolated to any other ethnicity trait. Given that in general ethnic traits are not going to be inherited only (nor probably mainly, and certainly not exclusively) across the male line, then large degrees of effective assortative mating across ethnic lines are necessary for surnames to carry information on ethnicity. Now, normally ethnicity and assortative mating walk hand in hand. Ethnic groups are characterized, one might even say that they are *defined*, by the prevalence of intra-group marriage.

A last remark is that for the surname to be indicative of ethnicity it is needed that ethnicity is predictable across the male line. Notice that this does not mean necessarily that assortative mating needs to be across the ethnic dimension, even if that is overwhelmingly the case in the real world. If the differences in income associated with ethnicity ($\alpha_w - \alpha_b$) were large enough, then even if assortative mating appears exclusively across income, but if it is very accurate (low σ_u in the model of section 3.5) then surname would capture ethnicity: the reason why people is rich being mostly that they have a large α_w, not only that they have parents with large income.

Surnames may also have information on an individual as a consequence of her being either an emigrant or a descendant of emigrants. Upon arrival to the receiving country emigrants have an initial distribution of surnames that is different from the distribution of surnames in the country at large. Very often this differences are notorious even from a phonetic and linguistic point of view, which makes the surname a salient feature in order to recognize a US citizen as "Irish", "Polish" or "Italian".

Upon arrival the immigrants have not only a different surname pool, but also they typically are substantially poorer than the recipient population. With time the distribution of income across the descendants of the emigrants would become more like the one of the descendants of the original population; the distributions becoming identical in the fullness of time. In the mean time it is possible to measure the speed of integration of the migrants in the population by observing how much does their surname explain about their income.

This is an approach that we do not exploit in this paper, but that we plan to use in future work.

4 Surnames in Spain are not named in vain

In this section we use census data from a large Spanish region to show that our methodology allows to make inferences and comparisons on intergenerational mobility. Specifically we show: (1) that the informative content of surnames is large, even after controlling for ethnicity/migration; (2) that the manner in which surnames are informative is coherent with the model; and (3) that the informative content of surnames has grown over time, suggesting an increase of the importance of background as determinant of outcomes.

We also use the peculiarities of the Spanish naming convention in order to show: (4) that the correlation of income of siblings is large and it has also grown over time; and (5) that assortative mating has also increased over time. All this is consistent with the observed fall of intergenerational mobility, and suggestive that our methodology is robust.

We proceed as follows. In subsection 4.1 we explain the peculiarities of Spanish naming convention, why they make Spanish data particularly useful for our purposes and we present the data used in the paper. We also show the characteristics of the distribution of surnames in Catalonia and Spain. In subsection 4.2 we explain our controls for ethnicity/migrant origin of individuals. In the following subsections we show our empirical findings on the degree and evolution of the informative content of surnames, the frequency of surnames and the degree of assortative mating. Additionally, in subsection 4.4 we make a calibration fitting our findings to the model and suggesting values for the degree of inheritance in the economy that are comparable with the existing literature.

4.1 The Distribution of Surnames in Spain and Catalonia.

4.1.1 Spanish Surnames

The Spanish naming convention differs from the standard western naming convention in two manners. (1) Spaniards have two surnames, the first one being inherited from the father (it is the first surname of the father), and the second from the mother (it is her first surname). (2) Females never change their surname, irrespectively of marital status.

For survival of lineages purposes this is identical to the better known Anglo-Saxon convention, in which only the father's surname is inherited and the female adopts her husbands' surname upon marriage. The reason is that in Spain after two generations all trace of the mother's surname has disappeared and the only thing that effectively has been passed along is the father's surname. If (when) we use only the first surname and restrict our data to males, Spanish census data would be identical in kind to US (or any other western) census data. The Spanish data are the different only by (1) including the equivalent to the mother's maiden name for all individuals, and (2) including the maiden name (but not the husband's surname) for all females.

As a consequence Spanish data are particularly useful in order to illustrate the possibilities of our methodology. The main[34] reasons are:

[34]There are other reasons. We can use women in our regressions, as we are certain that their surnames contain information about their blood line and not about their husband's blood line. The surname of an Spanish female tells things about her father, and her siblings; not about the father and siblings of her husband.

(1) The second surname (the mother's first surname) allows us to use one of the surnames to control for ethnic issues, while we use the other to measure the informative content of surnames due to family linkages. (2) The fact that we have both surnames available allows us to approximate the surname partition to family linkages with a very large degree of accuracy. The reason is that the probability that two individuals who are not siblings share the same two surnames in the same order is very small; it is vanishingly small if the surnames are quite unfrequent. (3) Having two surnames we can also look at the manner in which the parents of the individual married. We have a direct view to the degree of assortative mating. (4) As already stated, restricting to the first surname (and to males) Spanish data that are perfectly comparable (actually, identical) to any other western surname data. (5) Previous to 2001 there was no significant migration toward Spain. Migrations were substantial, but were either internal or toward outside Spain. There was no significant influx of people (and foreign surnames) in Spain practically since medieval times. This proves to be very useful, as allows us to use only Spanish data in order to control for ethnicity (understood as regional origin) without the need of addressing foreign sources of data. We use the Spanish Phone Directory as a comprehensive source of information on ethnicity.

4.1.2 Data Sources

Our data comes from two sources. (1) From the 2001 Spanish Census, we have very detailed microdata for a large region of Spain (Catalonia). (2) We also use the Spanish Phone Directory; these allow us to analyze the distribution of surnames as well as to obtain ethnic information.

Census Data

The Census Data for Catalonia covers the whole population (6343110 individuals). To our knowledge this is the first paper that makes use of the surnames in extensive census microdata. We have the two surnames of each individual, demographic characteristics (age, education, gender, marital status, place of birth, place of residence) as well as employment status, level of proficiency in Catalan language and housing characteristics (tenancy, size, inheritance, availability of a second house). For privacy reasons, we do not have some of the data hypothetically available. For instance, we do not have any information that links individuals (i.e., parents and children). We have some household data (e.g., the number of members and the size of the house), but we can not determine who are the members of the household nor family linkages across households. This is not a restriction in our case, as the point of our exercise is to show that we can study intergenerational mobility just using cross sectional data on surnames and measures of economic wellbeing.

In Spain, census data do not include information on income. We use years of education as a proxy for income

Additionally, the Spanish surnames are quite constant over time. Their historic mutation rates being in all likelihood much lower than in other western societies. This is partly a consequence of the peculiarities of Spanish orthographic rules and phonetics make that reproduction mistakes are much less likely in Spanish than in other countries.Thus, there is only one Spanish spelling of "Rodriguez" (meaning "son of Rodrigo"; "Rodrigues" is a Portugese surname and spelling) while there are multiple spellings of Johnson (meaning, son of John). This is a well known fact and the reason why the studies on genetic inbreeding done with surnames were very often done in Spanish speaking countries (see Lasker (1985)). It is also notorious that in Spain changing surname involves a complicated bureaucratic procedure. A consequence is that we have data on the number of surname changes in Spain and a direct measure on the mutation rate, as we have already seen. In 2001 there was a simplification of the procedure, particularly in what relates to surname order, but we use data previous to the change.

Table 1: Surnames Distribution: Gini Index and People per Surname in Catalonia and Spain

	Spain	Catalonia	
	(PhoneBook)	(Census)	(PhoneBook)
Number of People	11,397,116	6,123,909	2,073,219
Number of Surnames	155,782	91,568	61,396
People per Surname	73.161	66.878	33.768
Gini Index	0.9485	0.9304	0.9028

Source: Spanish Phone Book & Catalan Census. Sample: All phones with first & second surnames not missing.

and economic advantage. We have done some of the analysis also with other proxies of income (in particular availability of second residence and information on if the residence has been inherited) and obtained the same qualitative results. We do not think that this is a particularly damaging issue for proving our methodological point. Of course, it would be good to also have income data in the LHS, but the combination of education, gender and place of birth is probably good measure of lifetime income. In any case we believe that our analysis (using only education in the LHS) is robust enough to allow us to do comparisons of intergenerational mobility across time and regions.

For the purposes of our study we consider only individuals aged 25 and above, who have potentially finished full time education. We also consider only individuals with Spanish nationally in order to focus on the informative content of surnames for family reasons and not due to migration (more on this in section 4.2).[35] Finally, we exclude individuals whose first surname is unique to them, as the information that their surname may carry does not refer to family links. We refer to this as "Complete Population".

Telephone Directory

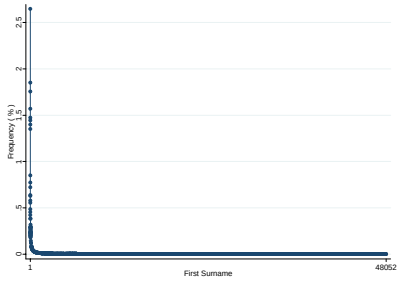
We also use the telephone listings in order to (1) calculate the distribution of surnames in the whole of Spain, and (2) in order to determine how "Catalan" (vs. from the rest of Spain) a surname is. This will allow us to use ethnic/migration controls out of the second surname of the individual.

The Phone Directory is available from commercial sources. It contains information only on fixed telephone lines, not mobile phones, which would be a substantially larger number. We use only information on private lines. For Spain, there are around 11.4 million of them, while there are around 14 million households in Spain. Certainly, some households have more than one line, while others have none (precisely because of the prevalence of mobile phones). Nevertheless there are no reasons to suspect that the distribution of people doing this differs between Catalonia and the rest of the regions of Spain, so this database is well suited for our purposes.

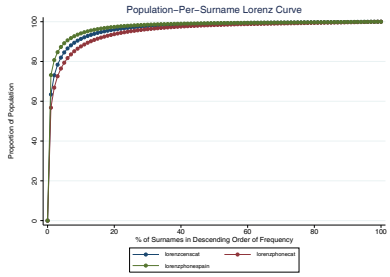
4.1.3 The Surname Distribution.

Figure 5(a) plots the distribution of the first surname. It places first the most common surname and assigns to it its frequency, then with the second most frequent, and so on. The distribution is very skewed. There exists a

[35] Individuals living in "collective households" were also excluded from the population as there is no information on education for them. These represent less than 0.6 percent of the population.



(a) Distribution of First Surname in Catalonia



(b) Lorenz Curve of the Surname Distribution in Catalonia (Census & Phonebook) and Spain

Figure 5: Distribution of the first surname in Catalonia and lorentz curves for Spain and Catalonia.

For 5(a): Source: Catalan Census. Sample: Spanish citizens in Catalonia aged ≥ 25.

For 5(b): Source: Spanish Phone Book & Catalan Census. Sample: All phones with first & second surnames not missing.

large number of low frequency surnames and the few most frequent surnames include a large percentage of the population. The 10 most popular last names cover around 11 percent of the sample population.

Figure 5(b) offers a graphical representation of these skewness. It draws the Lorenz curves of the distribution of surnames for all Spain and Catalonia as obtained from the phone–book, and for Catalonia as obtained from the census. Table 1 offers the relevant statistics for the three distributions. Notice that the number of people per surname is larger in the whole of Spain than in Catalonia (probably because Catalonia is a net receiver of immigration and Catalan language has had historically less orthographic rigidity than Spanish[36]) and census than in phone–book data (because tipically there is only one phone per family).

We can get an idea of the magnitude of the flows in and out of the surname pool with available data. Regarding the flow out of the surname pool, the probability that a woman aged 50 or above have had no male kids is 0.33. This allow us to calculate that it seems reasonable to assume that the number of dying surnames is in the range 100 to 110. [37]

It is more difficult to measure the inflow into the surname pool. It is composed of the inflow emigrants with surnames non existent in the Catalan surname pool and the mutations in surnames that there might be in Catalonia. We do not refer to the emigrants because we will take them out of our population and because it is difficult to know how likely are their surnames to be "new" in the Catalan pool.[38] We have already seen that the registry of applications for change of surnames suggests a mutation rate of around 0,26%. Using the sample of all Spanish males living in Catalonia this amounts to around 107 surnames new yearly. Thus, our calculations suggest that

[36] We did not consider the fact that the same surname could have different spellings in Catalonia. One in Catalan and another in Spanish. e.g. "Suñol" and "Sunyol". We counted them as two different surnames.

[37] If males and female had equal to the probabilities of having no male children, given the distribution of surnames this would imply that each year 96 surnames would die. This probability is more likely an underestimation because of two reasons. (1) It does not take into account the women age younger than 50 that die without having male children. (2) On the top of it, it is likely that the number of males having children is smaller than the number of females having children (some male mating with more than one children–bearing women and others with none). We have no way of looking at this, but it is reasonable to assume that this caveats not account to much: On one hand the life expectancy of women is very high in Spain, substantially more than 71 years at birth. Additionally the uneven distribution of the number children–bearing mating partners for males is likely to be less pronounced when one refers to surnames than when one refers to actual births. The reason is that children are likely to carry the surname of the husband of their mother, not necessarily of their father. We thank Namkee Ann for providing the data based on the 1991 Spanish Socio-Demographic Survey.

[38] In particular if we take into account that a large percentage of the emigration came from South-America, their names having first "migrated" from Spain, and now coming back to their origin.

except for the process of immigration the flows in and out of the surname pool are of similar magnitudes.

4.2 Immigration and Ethnicity: How Catalan a surname is?

As we have emphasized in previous sections surnames carry information both on the ethnicity of the individuals and on their family links. Our interest in this paper is mostly focused on the information that the surname partition may contain on family linkages, as our model shows that this is directly related to intergenerational mobility. For this reason we concentrate our empirical analysis on Spanish citizens[39] excluding foreign immigrants.

Even after these restrictions we would like to filter away the "ethnic" component of surnames. There are two reasons: (1) There could exist differences in education acquisition that can be traced to cultural and linguistic differences among individuals who have forefathers of different geographical origins within Spain. These cultural differences are what we deemed in this paper as "ethnic": differences in *regional origin* within Spain. (2) Second generation migrants are likely to be poor due to specific reasons that do not affect the native population (racism or other forms of bigotry and discrimination). There is also the effect due to the different starting points. This effect should decrease over time, but nevertheless be prominent during large periods after a large migration.

These reasons are particularly relevant in the Catalan context, because: (a) There were huge migration flows between 1955 and 1975 from the rest of Spain toward Catalonia, [40] and (b) The obvious existence of a linguistic divide. People that have Catalan as their mother tongue are very likely to be descendants of the original population, while people that has Spanish as their mother tongue most often are descendants of emigrants from the rest of Spain. Each of these linguistic groups represents around half of the population. It is notorious and well known that income per capita, education attainment and any other measure of economic wellbeing are substantially larger among those who have Catalan as their mother tongue.

In order to control for these ethnic/migration effects we use the fact that the distribution of surnames differs between different Spanish regions.[41] We calculate a variable (*CatalanDegree*) that determines for each surname the probability that a Spanish individual holding is a Catalan resident. We then use this variable as a proxy of the "ethnic" origin of the individual.

CatalanDegree is calculated for each surname as the probability that an individual with that last name and has a phone registered in Catalonia. Specifically, this is

$$CatalanDegree(j) = \frac{\text{Number of Phones under surname j in Catalonia}}{\text{Number of Phones under surname j in the Spain}}$$

It captures the probability that a Spaniard holding name j is living in Catalonia. We use this variable as a proxy of the "catalonianess" of the holder ("how much" Catalan she is), and through it we will characterize the information that surnames have on ethnicity (and its effects). Table 2 shows its summary statistics and figure

[39] Born in Spain and with Spanish nationality.

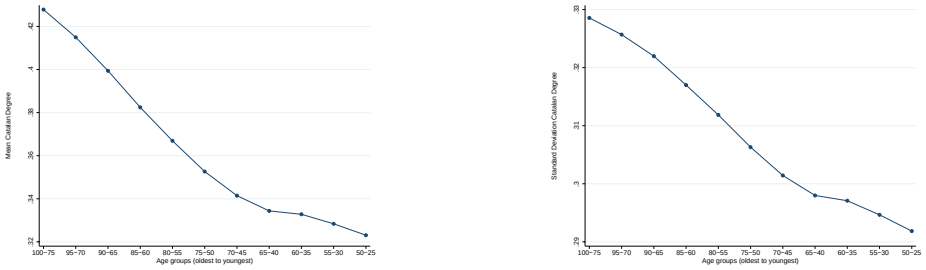
[40] Without this migration the population in Catalonia would be of around $2,7$ instead from above 6 millions the year 2000, as reported by Cabré (2004).

[41] Indeed it is different between different provinces, counties, etc. By these we do not mean the general skewness but the actual surnames, except for the most frequent ones which tend to be the same in all regions, provinces and counties.

Table 2: *CatalanDegree* Summary Statistics. Complete Population

	All residents in Catalonia	Born in Catalonia before 1950	Born anywhere in Spain before 1950	Born anywhere in Spain after 1950
Mean CatalanDegreeSurname2	0.344	0.5672	0.367	0.322
Standard deviation	(0.302)	(0.3241)	(0.312)	(0.292)
Share with CatalanDegreeSurname2>0.16	0.568	0.8365	0.596	0.542

Source: Catalan Census. Sample: Spanish citizens in Catalonia aged $\geq$ 25 with frequency of first surname$>$ 1.



(a) Average *Catdegree* (b) Standard Deviation of *Catdegree*

Figure 6: Evolution of *CatalanDegree* over (a moving window of) cohorts. Complete Population.

Source: Catalan Census. Sample: As in table 2.

6 plots its evolution over time. We calculate the average and standard deviation of *CatalanDegree* for different cohorts in a moving window. In the left we place the individuals who in 2001 were between 75 and 100 years of age. Next we place the individuals whose age was in the range 95-70, and so on.

We notice two things. First, the surnames of Catalans have become dramatically more like the average distribution of surnames in Spain as time has gone on, at least in the sense that the Catalan surnames are much less specifically Catalan that they used to be. This is a consequence of the large migrations in the 60's. Immigration took place mostly after 1950, so among the individuals born in Catalonia before that date there is little issue about migration (as they are an ethnically homogeneous population). Catalonia represents around a 16% of the Spanish population. Thus, if your *CatalanDegree* is larger than 0.16 you are more likely to be Catalan than from the rest of Spain. In table 2 we can see that the number of individuals who are ex-ante likely to be Catalan (as judged by their surnames) has decreased dramatically.

Second, *CatalanDegree* is a good measure of the migrant status of the individual. The average *CatalanDegree* and the probability of being Catalan are much larger for those agents born in Catalonia before 1950 than for those born anywhere in Spain in the same period and living in Catalonia in 2001. Our index measures well if a resident born before 1950 was an emigrant or if it was born in Catalonia. Thus, it is also a good measure on the "ethnic" background of their children.

The fact that *CatalanDegree* is a good proxy for the ethnicity of the individual is important for our results. We run two regressions to reassure that it is correct.

In table 3(a) we run a probit that has as LHS a variable that takes value 1 only if the individual has full

Table 3: *Catalandegree* & Probabilities of knowledge of Catalan language and being an immigrant.

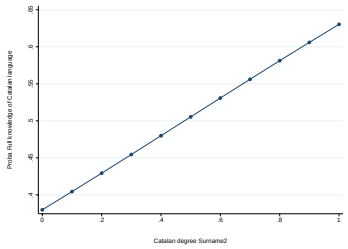
(a) Probability of knowledge of Catalan language

LHS: *Knowledge of Catalan language*	(1)	(2)
CatalanDegreeSurname2		0.639 $_{(0.003)}$
Log likelihood	-2248320.8	-2219774.9
Pseudo R^2	0.2387	0.2483

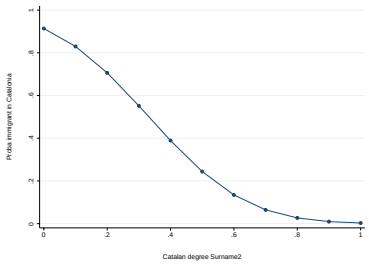
(b) Probability of being an immigrant

LHS: *Immigrant*	(1)	(2)
CatalanDegreeSurname2		-4.121$_{(0.006)}$
Log likelihood	-1418949.8	-903746.31
Pseudo R^2	0.0052	0.3664

Probit Estimates. All regressions include individual controls and regression in table 3(a) also includes place of birth controls. Source: Catalan Census.
For 3(a). Sample: Spanish citizens in Catalonia aged $\geq$ 25 with frequency of first surname$>$ 1. The variable *Knowledge of Catalan language* takes value 1 for individuals who can understand, speak, read and write the Catalan language and zero otherwise. Number of observations: 4,293,173.
For 3(b). Sample: Individuals born before 1950 of sample in table 3a. The variable *Immigrant* takes value 1 for individuals who were not born in Catalonia and zero otherwise. Number of observations: 2,057,831.



(a) Probability of knowledge of Catalan language



(b) Probability of being an immigrant

Figure 7: *CatalanDegree* & Probabilties of Catalan language knowledge and being an immigrant.

Notes as in table 3. For figure 7(a), reference individual is a male, aged 50-55, born in the county of Barcelona. For figure 7(b), reference individual is a male, aged 60-65.

knowledge of the Catalan language.[42] In the RHS we include individual controls (gender, place of birth, age). In column (2) we add *CatalanDegree*. It affects positively the probability. Figure 7(a) shows that the probability of knowledge of Catalan is increasing in *CatalanDegree*.[43]

Table 3(b) calculates the probability that an individual older than 50 did migrate into Catalonia from the rest of Spain. Clearly *CatalanDegree* affects this negatively and is enormously explicative. This can be seen graphically in figure 7(b).

Thus, we use *CatalanDegree* as a measure of ethnicity and as a mean of controlling for it. We do it two different ways. In most of the paper we will use the information that the second surname (the one inherited from the mother) has on ethnicity in order to control for migration/ethnicity issues. We will place on the RHS of the regressions the *CatalanDegree* of the second surname of the individual. We will use a dummy of the first surname of the individual in order to measure the ICS, as in our model of section 3.2. Thus, constructing *CatalanDegree* on the second surname we minimize multicollinearity issues.[44] A second manner of controlling for ethnicity is to restrict the population to ethnically homogeneous groups. Thus, we calculate the geometric mean of the *CatalanDegree* of both surnames and order the population according to this ranking. We then select the 40% of the individuals with the highest geometric average as an ethnically homogenous population of non-immigrants.

[42] In the census she answers that she understand, speaks, reads and write in Catalan. This is 45% of the population aged above 25
[43] See Ortega (2007) for a further study on Catalan knowledge.
[44] We look at the degree of assortative mating in education and "catalonianess" below.

Table 4: ICS. Complete Population.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.692(0.007)	1.017(0.008)	1.692(0.007)		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.3363	0.3440	0.3653	0.3440	0.3629	0.3363
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.384	0.000	0.332

All regressions include individual and place of birth controls. Notes: Standard errors in parenthesis. Fake-surnames have the same distribution as Surnames and are allocated randomly. (*) F-test if Surname dummies are jointly significant. Source: Catalan Census. Sample: Spanish citizens in Catalonia aged ≥ 25 with frequency of first surname> 1. Number of observations: 4,293,173. Number of surnames: 38,024.

4.3 Surnames contain information beyond ethnicity

Table 4 shows that surnames do contain information about an individual socio-economic status for the Complete Population.

The first column shows the results of a regression of years of education on individual controls (dummies for gender, age and place of birth[45]) for each individual in the complete population, the adjusted R^2 is 0.3363.

The second column adds *CatalanDegree* of the mother's surname (the second surname of the individual). It is no surprise that its coefficient is positive. The parameter value of 1.692 means that an increase of one standard deviation in the degree of catalonianess translates into 0.5 additional years of education, which is around a 10% of the standard deviation of education.[46] The standard errors of the estimates are always very low due to the very large samples. Large population size implies that the variables are bound to be statistically significant unless they have really no possible logical connection to the LHS. However in our case the important thing is not statistical significance, but that it is economically meaningful: it has a sizable effect.

The third column adds to it a dummy per surname. This is, all individuals who share the same *first* surname share a dummy, and we test for the joint significance of all these variables (there are $38,024$ surnames, around 113 people per surname). There are three noteworthy results of this equation: (1) The surname dummies are jointly statistically significant (given the large number of RHS variables involved this is not obvious in spite of the large population size). (2) The coefficient of *CatalanDegree* falls substantially, but is still economically meaningful (an increase in one standard deviation translating to 4 extra months of education). (3) *The adjusted R^2 of the regression increases to* 0.3653. By knowing the surname of a person[47] you know more of the person than if you do not know the surname.

In the fourth column we substitute the surname dummies by "fake–surname" dummies as in sections 3.1.2 and 3.2. "Fake–surnames" are not jointly significant, and the R^2 increases only marginally with their presence. The value of the Catalan degree index being not modified from the one that it had in the absence of surname controls.

For further reference we can calculate the informative content of surnames as 2.13%, this is, the difference of

[45] Our control of the place of birth of the individual differs if she was born in Catalonia or anywhere else within Spain. If Catalan-born, we use county dummies, otherwise we use dummies for the province. County is an administrative unit somewhat smaller than a typical US county. Province is a larger administrative unit, somewhat larger than a French "departament". We have done the same exercise with town of birth dummies instead, and the same results arise.

[46] The mean and standard deviation of education are respectively 8.004 and 4.751.

[47] We mean to know the surname *and its meaning*. This is, to know if to be called $C3PO$ has a larger value of the "dummy" than being called $R2D2$.

Table 5: ICS in subpopulations.

(a) Born in Catalonia.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.703_{(0.008)}$	$0.994_{(0.009)}$	$1.703_{(0.008)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Ajusted R^2	0.2256	0.2379	0.2697	0.2378	0.2661	0.2255
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.862	0.000	0.855

(b) Born in Catalonia before 1950.

CatalanDegreeSurname2		$0.952_{(0.013)}$	$0.575_{(0.014)}$	$0.956_{(0.013)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.1578	0.1622	0.1947	0.1621	0.1932	0.1577
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.709	0.000	0.772

(c) 45% Most Catalan Surnames.

CatalanDegreeSurname2		$0.772_{(0.010)}$	$0.712_{(0.010)}$	$0.773_{(0.010)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.3186	0.3207	0.3435	0.3207	0.3419	0.3187
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.500	0.000	0.165

All regressions include individual and place of birth controls. Notes as in table 4. Source: Catalan Census.
For 5(a) Sample: Individuals born in Catalonia of sample in table 4. Number of observations: 2,721,917. Number of surnames: 31,987.
For 5(b) Sample: Individuals born in Catalonia before 1950 of sample in table 4. Number of observations: 1,017,123. Number of surnames: 21,602.
For 5(c) Sample: Individuals with 45% Most Catalan (first) Surnames of sample in table 4. Number of observations: 1,931,187. Number of surnames: 25,980.

the R^2 of the regressions with with real and fake surnames.

Columns (5) and (6) repeat the exercise of columns (3) and (4) respectively, but without controlling for the degree of "Catalonianess". The results are qualitatively alike. Nevertheless, surnames dummies now capture not only close common descent, but also ethnicity. Consequently, along with what the model would predict, the ICS is in this case larger (2.66%).

Tables 5(a) and 5(b) repeat the exercise of table 4, but restricting the population to those born in Catalonia (table 5(a), emigrants are not included, even if their children are) and to those individuals born in Catalonia before 1950 (table 5(b)). The results are parallel to those we have already seen, ICS being somehow larger because these are more ethnically homogeneous populations.

It could be a concern that in spite of controlling for *CatalanDegree* we are capturing ethnicity, and not family. To show that it is not the case table 5(c) restricts the sample to the 45% of the population who has the most catalan surnames in the manner expresed above. This population is much more ethnically homogeneous, nevertheless the results do not change qualitatively. If anything the ICS is larger: 2.28%. More importantly, unlike in the previous tables the ICS is the same whether controlling for the *CatalanDegree* or not.

To conclude, in this section we have shown that surnames do contain information, even after controlling for ethnicity/migration in different ways. We read our results as implying not only that (1) ethnic/migrant status matter (and can be captured by surnames), but also that (2) surnames have informative content once the effects of

migration/ethnicity are taken into account. As argued in the previous sections this information is a consequence of the fact that the surname partition of society informs about close common descent of the agents. This is one of the main empirical results of our paper. It is robust to everything that we have try to control for. It is more than a funny empirical singularity; it allows us to use the informative content of surnames as a proxy for inheritance. In the next pages we give economic content to this result.

4.4 Calibration

In this subsection we calibrate the model of Section 3.2 to the characteristics of our surname data from Catalonia. The key feature of our procedure is that we choose the inheritance parameter, ρ, to match the informative content of surnames as measured by the incremental R^2 from the regression in Table 5(a).

The interesting result is that the value of ρ that matches the data are in line with the estimations of ρ obtained in the US and Europe (it is somewhat larger than 0.4). By all means this is a rough calibration, and a noisy estimation of ρ. It is also very important to notice that we are calibrating data on education, not income, which makes comparisons difficult. Still, we present this calibration because it is additional indication that the alternative method that we propose (to measure the informative content of surnames) is a quantitatively coherent approach.

Given the purpose of our exercise, we calibrate the simplest of our models to the data. Thus, we assume the existence of no differences in the rates of birth/death of lineages across income groups. In the simulations of section 3 we modeled economic wellbeing as "income". Since our Catalan data are on education, we now adopt this interpretation and change the model slightly in order to adapt to it. Denote the educational attainment of the j^{th} male in the population at date t as $e_{j,t}$. Education is specified to be a positive, continuous variable which follows an AR(1) from one generation to the next: $\log e_{j,t+1} = (1-\rho)\theta + \rho \log e_{jt} + \sigma\varepsilon_{j,t+1}$, where $e_{j,t+1}$ refers to all descendants of household j at date t.

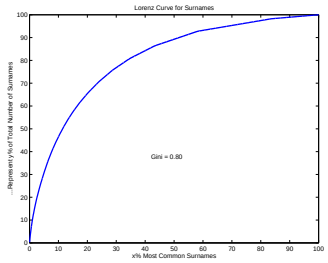
We calibrate the parameter ρ to match the informative content of the surnames from Table 5(a). What this means is that we choose all other parameters to match other aspects of data, and choose ρ so that the R^2 from the following regression matches that of the data: $e_j = \alpha + \beta D_j + \text{residuals}_j$, where D_j is an S-vector of dummy variables such that $D_{ij} = 1$ if $h_j = s_i$, $i = 1, 2, \ldots, S$, and β is an S-vector of coefficients.[48]

- *Education.* The mean and variance of the level of education, e_{jt}, depend on θ, σ and ρ.[49] But, for any given value of ρ, there are unique values of θ and σ implied by the (unconditional) mean and variance of e_{it}. Therefore, for each value of ρ that we consider (to match the informativeness of surnames) we'll set θ and σ as to match $E(e_{jt}) = 9.18$ and $Stdev(e_{jt}) = 4.56$ respectively.

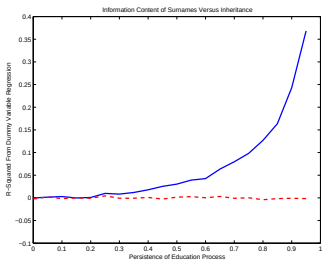
- *Fertility.* We choose M and Q such that the population size is stationary: $E(N_t) = N_{t-1}$. This means that MQ =1. We set arbitrarily $M = 2$, $Q = 1/2$.

[48] Strictly speaking, we should choose ρ to match the difference in the R^2 between the above regression and the associated "fake" regression, where surnames are randomly allocated to households in a manner which maintains the same marginal distribution of surnames. However, the latter regression (in my simulations) has R^2 of (essentially) zero. Therefore, I'll ignore it.

[49] Note that this is because we've specified e_{jt} to be lognormal, not normal.



(a) Calibrated surname distribution



(b) Relationship between ICS and ρ imputed from the model

Figure 8: Calibration

Table 6: ICS. 50% Least Frequent Surnames.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.438_{(0.011)}$	$0.791_{(0.011)}$	$1.439_{(0.011)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.3303	0.3364	0.3709	0.3364	0.3693	0.3302
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.633	0.000	0.665

All regressions include individual and place of birth controls. Notes as in table 4. Source: Catalan Census. Sample: Individuals with 50% Least Frequent (first) Surnames of sample in table 4. Number of observations: 1,933,149. Number of surnames: 37,540.

- *Mutation.* We choose μ to match the concentration of surnames. A value of $\mu = 0.002$ generates a Gini coefficient of 0.82, which is maximal for our model, but less than the value of 0.903 that we observe in the data. This is indication that the process for fertility should be enriched in order for our model to be able to account for the concentration of surnames observed in the data. It is nevertheless interesting that the value of μ that we obtain is basically identical to the one that we (independently) got from the data of the Office of Public Records (see in page 11). Figure 8(a) shows the distribution of surnames generated.

- *ICS.* Given the parameters, we replicate the model of section 3.2. We can see the result in figure 8(b). From Table 5(a), the incremental contribution of surnames to the R^2 is 0.032. Thus, the calibrated value of the persistence parameter, ρ, is 0.47.

  The fact that this number (that should be considered a back–of–the–envelope calculation) is in the range that sounds "reasonable" reinforces our belief that by looking at the informative content of surnames we can infer where and when intergenerational mobility is larger.

4.5 The ICS is larger for less frequent surnames

Two individuals sharing a very frequent surnames are much less likely to have a close family linkage than two individuals sharing a very *unfrequent* surname. Thus, according to the model we should expect the ICS of surnames to be larger in a sample containing only individuals with very unfrequent surnames than in another that has also frequent surnames. In the measure that this is the case we can be confidant that the ICS that we recover is a consequence that the surname partition is informative on family linkages.

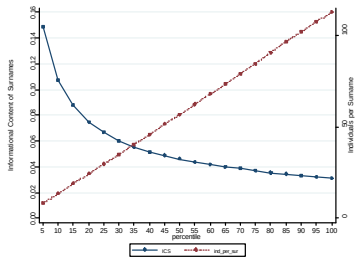


Figure 9: ICS is larger for less frequent surnames

Each point represents the ICS by percentile calculated as in table 4 (columns 3 & 4), where percentile x corresponds to the x% least frequent surnames. Source: Catalan Census. Sample: x% least frequent surnames of sample as in table 4.

Table 6 repeats the exercise of subsection 4.3 but including only the 45% of the population with the least frequent surnames. As expected, the ICS is the largest so far, 3.45%.

In order to reassure ourselves, we run a (large) series of regressions dividing the population by the frequency of their surnames. The regressions are identical to our main regressions (columns 3 and 4 in table 4), so we can extract the ICS as the difference of the R^2. They differ between them because for some of them we include only individuals with very unfrequent surnames, while for others we include *also* individuals with frequent surnames.

The results of this exercise are summarized in Figure 9. The horizontal axis represents the population ordered from the individuals with the less frequent surname (in the left) to the individuals with the more frequent one (in the right). They are marked by deciles. The decreasing line (with the scale on the left) draws the informative content of Surnames for a population with the individuals with surname frequencies that are smaller than the value of the x coordinate. This is, for $x = 5$ is the ICS for the 5% of individuals who have the least frequent surnames; for $x = 10$ is the ICS for the 10% of individuals who have the least frequent surnames; and so on. Notice that for $x = 100$ we have the whole population, and we refer again to table 4. The increasing line (with the scale on the right) is there in order to give a frame of reference. It tells the average number individuals per surname in each regression.

It is noticeable that the information contained in surnames decreases sharply. When we restrict the sample to individuals who have very unfrequent surnames (low values in the horizontal axis), we get much larger informative content of the surname. This is indication that our controls for ethnic issues are working, and that we are focusing on the fact that the surname provides information on the family linkages of the agents.

4.6 Richer Catalans have unfrequent surnames

As we have seen in the theory section the frequency of a surname (and not only the specific surname) should be expected to be informative about a person insofar the probabilities of lineage extinction/creation were different across social groups. In this section we show that in Catalonia frequency does contain information; richer people having less frequent surnames.[50]

[50]Notice that this is different from what we did in the previous section. There the specific surname was shown to be more informative depending on its frequency, but this information may indicate either poverty or wealth. Here we show that the *frequency itself* is informative, and that unfrequent surnames are more commonly associated to more education.

Table 7: Education and Surname Frequency. Complete Population.

LHS: years of education	(1)	(2)	(3)	(4)
FrequencySurname1	-30.157$_{(0.309)}$	-23.696$_{(0.309)}$		
FrequencyFakeSurname1			0.148$_{(0.301)}$	0.107$_{(0.299)}$
CatalanDegreeSurname2		1.636$_{(0.007)}$		1.692$_{(0.007)}$
R^2	0.3378	0.3449	0.3363	0.3440

All regressions include individual and place of birth controls. Notes: Standard errors in parenthesis. Fake-surnames have the same distribution as Surnames and are allocated randomly. Source: Catalan Census. Sample: as in table 4. Number of observations: 4,293,173.

In the first and second columns of table 7 in order to explain educational attainment in addition to the usual controls we add the *frequency* of the first surname. In the second column we also introduce the Catalan Degree Index of the mother's surname (the second surname). Clearly, the more frequent a surname is, the less education should you expect her holder to have achieved. In columns 5 and 6 we do the same with fake–surnames. Clearly the frequency of fake–surnames are not significant, which insures that it is really the frequency of the surname which produces the result. Richer people has less usual surnames even if we control for the degree of $CatalanDegree$ of the individual.

The point estimate of the effect of frequency of -23.696 can be qualitatively understood in the following way: an increase of one standard deviation in frequency translates into 0.15 years less of education. This is, a decrease of 3% of one standard deviation of education.[51]

Overall the results indicate that either the death rate of lineages is smaller among the more educated, or their birth rate is larger, or both. This is not surprising, as both effects are perfectly conceivable. It is more common to create new surnames (by hyphenation of first and second usually) among the newly rich.[52] It is also easy to imagine the existence of an "hereu" effect inducing better off families to have children until the point of insuring one male descendant. In the same manner notice that another effect that could go in this direction is if educated *males* were going to have more children than non educated *males*.[53] Notice nevertheless that we are excluding foreign immigrants, if we were to include them the results could very well change, as the effective mutation rate would be much larger for the poorly educated.

The summary is that even if this correlation is apparently interesting, it is difficult to learn about its causes. This makes it of relatively little use, even if worth reporting.

4.7 Siblings

As we saw in section 4.1.1 all Spaniards have two surnames, equivalent to the father's surname and the mother's maiden name in Anglo-Saxon countries. Thus, all siblings (irrespectively of sex and marriage status) share not one but two surnames and their ordering. This allows us to get a much more detailed representation of family linkages. In particular it allows us to determine with a large level of precision whether two agents are siblings. Lets call the set of two surnames in a certain specific order a "complete surname". This is, surname C3PO followed of surname

[51] For the sample of the table the mean of frequency of surname 1 is 0.00327 and its standard deviation 0.00620.

[52] We have not done it, but running education on the length of the surname would most likely be indicative of higher education.

[53] With females would be much more difficult to believe, but we refer to males. Even if educated males tend to marry educated females, the relatively widespread facts of divorce and second marriages points in the direction that educated males are more likely to have more children than less educated ones.

Table 8: ICS, complete surnames ("Siblings").

(a) All Spanish Citizens

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
Adj.$R^2_{\text{reg w/ surname dummies}}$	0.5486	0.5416	0.5375	0.5326	0.5283	0.4696
Adj.$R^2_{\text{reg w/ fake-surname dummies}}$	0.3244	0.3224	0.3218	0.3228	0.3232	0.3354
ICS	0.2242	0.2192	0.2157	0.2098	0.2051	0.1342
Observations	774,788	1,315,853	1,664,717	1,900,652	2,067,590	3,695,479
Number of complete surnames	387,394	567,749	654,965	702,152	729,975	811,502
Max number of people per complete surname	2	3	4	5	6	All sample

(b) 45% Most Catalan Surnames

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
Adj.$R^2_{\text{reg w/ surname dummies}}$	0.5371	0.5330	0.5301	0.5262	0.5225	0.4901
Adj.$R^2_{\text{reg w/ fake-surname dummies}}$	0.3086	0.3062	0.3084	0.3086	0.3085	0.3126
ICS	0.2285	0.2268	0.2217	0.2167	0.2140	0.1775
Observations	501,206	832,790	1,030,202	1,154,022	1,236,936	1,590,456
Number of complete surnames	250,603	361,131	410,484	435,248	449,067	478,691
Max number of people per complete surname	2	3	4	5	6	All sample

Surname dummies refer to each combination of 2 surnames. Each column displays the ICS by frequency x calculated as in table 4 (columns 3 & 4), where x = max number of people per complete surname, with $x = \{2, 3, 4, 5, 6, \text{all}\}$ in columns (1) to (6), respectively. Source: Catalan Census.
For 8(a) Sample: Spanish citizens in Catalonia aged ≥ 25 with frequency of complete surname> 1.
For 8(b) Sample: Individuals with 45% Most Catalan (first) Surnames of sample in table 8a.

R2D2 determine a specific complete surname (C3PO-R2D2). It is different from the one of C3PO followed by Vader (C3PO-Vader). It is also different from R2D2 followed by C3PO (R2D2-C3PO).

Complete surnames generate a partition of the population that is very informative of the sibling relationship. This is, if we group all individuals with their siblings, or if we group them by complete surnames we are bound to have very similar groupings. Particularly if we restrict the population to individuals who share both surnames with a small number of people.

The reason is straight forward. Take two agents who share the same two unfrequent surnames in the same order. The probability that they do not have the same parents is extremely small. It is extremely unlikely that two males with the same unfrequent surname would marry two females with another extremely unfrequent surname.

To a first degree we can approximate the sibling relationship by running a regression similar to the ones that we have done in previous sections but including one dummy not of one, but of two surnames for each individual. Thus, we place a dummy for the complete surname. For this exercise we have dropped individuals that are the only ones with the same complete surname, a number substantially larger than those who have an unique first surname.

Table 8(a) shows the results of a series of regressions with dummies for complete surnames. Each column reports the R^2 of two regressions. One with complete surnames dummies, and another with "fake–complete–surname" dummies. It also calculates the imputed ICS.

The first column restricts the population to individuals whose complete surname is shared only by two people. The second column includes also individuals whose complete surname is shared by three individuals. The third one the ones with four, and so on. The sixth column includes all the individuals whose complete surname is share with at least one other person.

There are several remarkable features. (1) The R^2 of the regressions are much larger than when controlling

Table 9: ICS, "invented" Catalonias. Complete Population.

LHS: years of education	(1)	(2)	(3)
$\text{Ad}R^2_{\text{reg w/ surname dummies}}$	0.3653	0.3661	0.3646
$\text{Ad}R^2_{\text{reg w/ fake-surname dummies}}$	0.3440	0.3429	0.3452
ICS	0.0213	0.0232	0.0194
Sample	All surnames	"First letters"	"Last letters"
Observations	4,293,173	2,187,084	2,106,089

ICS calculated as in table 4 (columns 3 & 4). Source: Catalan Census. Samples: Column (1) as in Table 4; Columns (2) and (3) are the first and second half, respectively, of sample in column (1) alpahebetically ordered.

with only one surname dummy. Now the regressors explain more than 50% of the total variance. (2) The R^2 with "fake–complete–surnames" do not change significally[54] with respect to the ones that we have seen before. (3) Thus, the ICS are now much larger. Complete surnames explain by themselves between a 13% and a 22% of the variance. (4) Finally, the ICS is larger the smaller is the frequency of the complete surname. The more people shares the same two consecutive surname, the less likely it is that they are siblings. The ICS is of 22.42% for complete surnames with only two individuals and of 13.42% for the whole population.

Notice that we are not controlling for ethnicity in this regression, as at least one of the surnames would capture both the ethnicity and the family components of surname information. We do omit the variable in order to avoid possible multicollinearity problems. To make sure that we are not capturing ethnicity effects we run the same regressions in an ethnically homogeneous group. Table 8(b) displays the same results as table 8(a) concentrating on the most Catalan surnames. It showsthe same qualitative results.

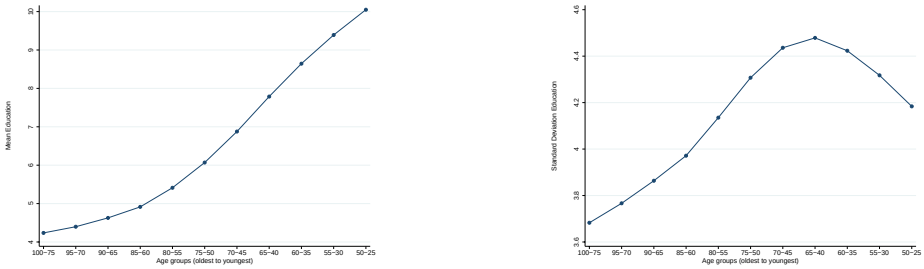
The fact that the ICS is so large when restricting to siblings is indication that we are capturing (and quite well) family links. Of course some people who are not siblings are categorized as siblings, even when there are only two of them sharing the same surname, this being more likely the more individuals are sharing the surname, but whenever there are two siblings living in Catalonia we assign to them the same dummy. Thus, if we were going to take this as a direct measure of the effect of common ancestry we would be under-estimating it. Nevertheless, like when we were using only one surname we are interested not as much on the number as in its meaning. In the following sections we will use this fact to look at the evolution of mobility.

4.8 Invented Catalonias

As a further robustness check we run a series of regressions where we restrict the surnames of the population in a random manner. If our method were correct the results should not be affected be centering in surnames starting with an "A" or with a "B". Thus, in table 9 we look at the ICS of random subsets of surnames.

The first column looks at the ICS of all the population (like in table 4) for comparison purposes. The second and third column do the same exercise with two "invented Catalonias". For convenience we have divided the population in two halves of around the same size. The first with the surnames starting by the first letters of the alphabet, the second with the ones whose surname starts by the last letters of the alphabet.[55]

[54] The small differences are accounted because there are different populations.
[55] We have done the experiment with other random groupings, and obtained the same result.



(a) Average (b) Standard Deviation

Figure 10: Evolution of years of education over (a moving window of) cohorts. Complete Population.

Source: Catalan Census. Sample: As in table 4.

The remarkable thing is that neither the R^2 of the regressions nor the ICS change with the populations. This shows that our findings are structural, that depend on deeply rooted social and economic mechanisms, and that they may shed some light on these mechanisms.

4.9 The ICS has grown over time

Census data permits to divide the population by cohorts, and this allows to look at how the degree of intergenerational mobility has evolved over time. In the model we have made only an steady state analisys, but it is clear that given that you start from a skewed distribution, an exogenous increase in the value of inheritance (ρ) would translate into an increase of the ICS, as the correlation between family members would increase. With a census we can determine the ICS for individuals born in different cohorts, comparing their values we obtain a measurement on the evolution of intergenerational mobility and the degree of inheritance.

For our general analysis we divide the population in two cohorts: those born before 1950 and those born after. This division is not arbitrary, as it has been noted migration flows exploded from the end of the fifties to the mid seventies and were very low before this period.

Access to education was very limited for Spaniards born before the 50's. This was a consequence of both the general poverty of the country and of the lack of investment in public education. Starting in the late 50's the economic situation dramatically changed in Spain, resulting in a much wider access to education and an enormous increase in the provision of public education, particularly from the 1975 and onwards. Figure 10 shows that average years of education have increased while the standard deviation has been much more stable. We show it over a moving windows of cohorts, like in figure 6.

Given this, one would be forgiven to think that surnames should be more informative on the educational attainment of older individuals (born before 1950) than among younger ones, but it would be wrong. Surnames are actually much more informative for the young than for the old individuals, suggesting a *decrease* in mobility *in spite* of the much wider access to education.

Table 10(a) presents the results of the same regressions than table 4 but restricting the population to those

Table 10: ICS over cohorts. Complete Population.

(a) Born before 1950 ("old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		0.896	0.594	0.897		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.2313	0.2335	0.2533	0.2335	0.2524	0.2313
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.679	0.000	0.688

(b) Born after 1950 ("young")

CatalanDegreeSurname2		2.143	1.271	2.145		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.1002	0.1183	0.1534	0.1184	0.1481	0.1002
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.260	0.000	0.421

All regressions include individual and place of birth controls. Notes as in table 4. Source: Catalan Census.
For 10(a). Sample: Individuals born before 1950 of sample in table 4. Number of observations: 2,052,725. Number of surnames: 31,237.
For 10(b) Sample: Individuals born after 1950 of sample in table 4. Number of observations is 2,232,102. Number of surnames is 31,847.

individuals who were born *before* 1950.[56] The results are similar to the ones of using the whole population. The adjusted R^2 of the regression with real surname dummies is 0.2533, while the for the regression with fake–surnames is 0.2335, resulting in an informative content of surnames of 1.98%.

In Table 10(b) we run the same set of equations, but this time restricting the population to those born *after* 1950. For this younger cohort the educational attainment of an individual is more difficult to explain with the right hand side variables, probably reflecting the much wider access to education, and much more equitable across genders and geographical locations. The Adjusted R^2 are much lower both for the regression with real surnames (0.1534) and the one with fake–surnames (0.1184).

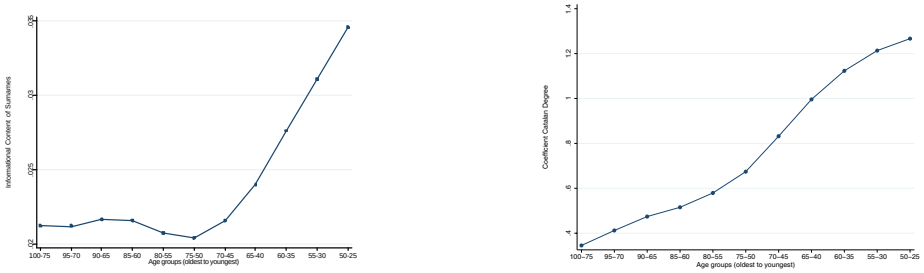
There are two remarkable features in the comparison of the tables. First that the parameter of $CatalanDegree$ has increased dramatically over time. Thus, ethnicity has become more important for determining outcomes.

Nevertheless what we find most important (and perhaps surprising) is that the surnames, *the explanatory variable that relates to inheritance and the past increases* its explanatory power, the ICS being of 3.5%, about 75% larger than for the previous generation. We read this as a strong indication that the degree of intergenerational mobility has decreased in Catalonia during this period, even after controlling for the increasing effect of ethnicity. A more dramatic way of looking at these results is in Figure 11.

Figure 11(b) shows the evolution of the parameter of $Catdegree$ over time, while figure 11(a) shows the evolution of ICS. We calculate them for subsamples of the population chosen according to their age in the same moving window manner that in figure 10. It is clear that the ICS of surnames increases dramatically, and does so monotonically, starting with the cohorts who started to gain access to public education, even after one controls for the increase in the incidence of ethnicity.

This is a potentially important and surprising result, so we run a series of further checks.

[56] It includes people born in Catalonia and people born outside Catalonia. If we were going to restrict to individuals born in Catalonia (like in table 5(a)) we would include the children of the emigrants in the sample of "young" but we would not include their parents in the sample of the "old". Thus, to look at time trends could be misleading.



(a) Evolution of ICS over time

(b) Evolution of parameter of *Catdegree*

Figure 11: Evolution of ICS and *Catalandegree* parameter over (a moving window of) cohorts. Complete Population.

Source: Catalan Census. Sample: As in table 4.

Table 11: ICS over cohorts. 45% Most Catalan Surnames.

(a) Born before 1950. ("old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$0.491_{(0.015)}$	$0.421_{(0.015)}$	$0.500_{(0.015)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.1973	0.1982	0.2237	0.1984	0.2231	0.1975
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.194	0.000	0.189

(b) Born after 1950. ("young")

CatalanDegreeSurname2		$0.989_{(0.014)}$	$0.973_{(0.014)}$	$0.991_{(0.014)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.0650	0.0702	0.1071	0.0704	0.1026	0.0652
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.086	0.000	0.105

All regressions include individual and place of birth controls. Notes as in table 4. Source: Catalan Census.
For 11(a). Sample: Individuals born before 1950 with 45% Most Catalan (first) Surnames of sample in table 4. Number of observations: 978,822. Number of surnames: 20,839.
For 11(b). Sample: Individuals born after 1950 with 45% Most Catalan (first) Surnames of sample in table 4. Number of observations: 947,573. Number of surnames: 21,879.

Table 12: ICS over cohorts. 50% Least Frequent Surnames.

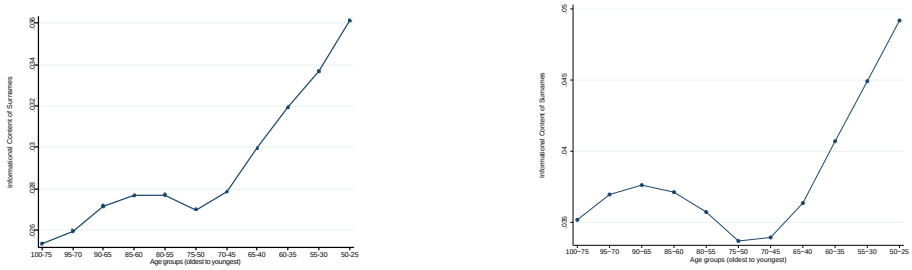
(a) Born before 1950. ("old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$0.730_{(0.016)}$	$0.449_{(0.017)}$	$0.732_{(0.017)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.2268	0.2284	0.2642	0.2283	0.2637	0.2267
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.698	0.000	0.716

(b) Born after 1950. ("young")

CatalanDegreeSurname2		$1.872_{(0.014)}$	$1.040_{(0.015)}$	$1.870_{(0.014)}$		
Surname Dummies			Yes		Yes	
Fake-Surname Dummies				Yes		Yes
Adjusted R^2	0.0945	0.1102	0.1616	0.1104	0.1576	0.0947
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.183	0.000	0.154

All regressions include individual and place of birth controls. Notes as in table 4. Source: Catalan Census.
For 12(a). Sample: Individuals born before 1950 with 50% Least Frequent (first) Surnames of sample in table 4. Number of observations: 924,094. Number of surnames: 30,721.
For 12(b) Sample: Individuals born after 1950 with 50% Least Frequent (first) Surnames of sample in table 4. Number of observations: 1,004,925. Number of surnames: 31,404.



(a) 45% Most Catalan Surnames

(b) 50% Least Frequent Surnames

Figure 12: Evolution of ICS over (a moving window of) cohorts, in subpopulations.

Source: Catalan Census. Samples: Individuals with 45% Most Catalan (first) Surnames and individuals with 50% Least Frequent (first) Surnames of sample in table 4, respectively.

Table 11 does the same exercise that table 10 but restricting the sample to the 45% of the population with the most catalan surname. As before, the ICS of surnames increases (in this case from 2.53% to 3.67%) even if the other RHS variables explain much less for the young than for the old. Notice that this is a much more homogeneous group in the ethnic dimension.

Table 12 repeats the same exercise but restricting the sample to the 50% of the population with the less frequent surnames. The ICS of surnames increases again (from 3.59% to 5.12%) while the other RHS variables still explain much less for the young than for the old. Notice that the values of the ICS are much larger than before, as smaller frequencies capture family links better.

Figure 12 shows the same growth of the ICS using the same moving windows of 25 years than figure 11 once we restrict the population to the individuals with high *Catdegree* (12(a)) and individuals with low surname frequencies (12(b)).

Finally, with complete surnames (like in subsection 4.7) the same pattern arises. Table 13 shows the tables for

Table 13: ICS, complete surnames ("Siblings") over cohorts. Complete Population.

(a) Born before 1950 ("old")

LHS: years of education	(1)	(2)
Surname Dummies	Yes	
Fake-Surname Dummies		Yes
Adjusted R^2	0.3664	0.2247
Surnames jointly significant*	Yes	No
(p-value)	0.000	0.248

(b) Born after 1950 ("young")

LHS: years of education	(1)	(2)
Surname Dummies	Yes	
Fake-Surname Dummies		Yes
Adjusted R^2	0.3158	0.0995
Surnames jointly significant*	Yes	No
(p-value)	0.000	0.547

All regressions have individual controls and place of birth dummies. Source: Catalan Census. Notes as in table 4.
For 13(a). Sample: Individuals born before 1950 of sample in table 4. Number of observations: 1,606,685. Number of complete surnames: 397,869.
For 13(b). Sample: Individuals born after 1950 of sample in table 4. Number of observations: 1,902,912. Number of complete surnames: 468,995.

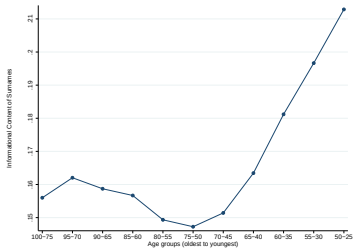
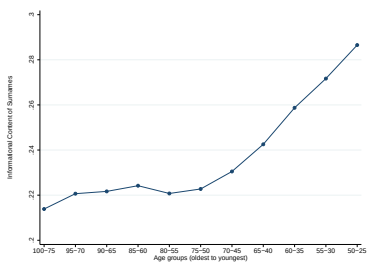
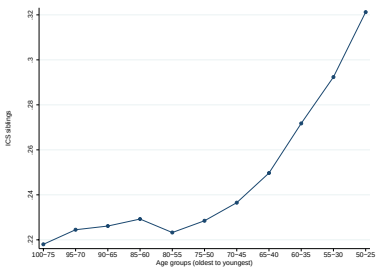


Figure 13: Evolution of ICS over (a moving window of) cohorts, complete surnames ("Siblings"). Complete Population.

Source: Catalan Census. Sample: As in table 4.



(a) 45% Most Catalan Surnames



(b) 50% Least Frequent Surnames

Figure 14: Evolution of ICS over (a moving window of) cohorts, complete surnames ("Siblings") in subpopulations.

Source: Catalan Census. Samples: Individuals with 45% Most Catalan (first) Surnames and individuals with 50% Least Frequent (first) Surnames of sample in table 4, respectively.

young and old. Figure 13 shows the evolution for the population, and figure 14 for subpopulations with the more Catalan and less frequent surnames.

The figure that emerges is very robust. The value of ethnicity has increased, and ICS has increased, even controlling for ethnicity. ICS has increased even in very ethnic groups, among siblings, among people with unfrequent surnames. There has been a general increased in the access of education, and very large, but background matters now much more than it did in the past in order to determine how much education you acquire.

It is not only that the value of being of Catalan background (versus with background form the rest of Spain) has increased, which undoubtedly has. It is that the specific characteristics of your family matter more. There has been growth in average educational attainment, but intergenerational mobility has decreased. There is an apparent contradiction, but is an illusion, as there exists no reason for which growth in itself should cause more mobility.

Almost everybody has a better material life than their parents did, but that has nothing to do neither with mobility nor with inheritance. Nothing that affects "to almost everybody" could affect mobility. It is an obvious consequence of long term growth. Nevertheless it produces the misperception that one has beaten the odds, that one has done better than his parents, and has done better than expected. As one generalizes, this results in a upward bias in the perception that a society has on the degree of intergenerational mobility (and to consider that inheritance is less important than actually is).

It is an illusion, it is just growth. Mobility works at its own path. It is defined only in relative terms. To measure mobility it is not enough to compare my welfare with my parent's welfare; it is necessary to compare it with the welfare of the children of the individuals whose parents were richer (and poorer) than mine. Intergenerational mobility is defined relative to others. Mobility may well be decreasing, while everybody lives better than their parents.

Our results are extremely suggestive that this is exactly what happened in Catalonia. The average educational attainment increased across the board, product of both an increase in the provision of public education and the general economic improvement. Nevertheless along this general increase in the average level of education attainment there arise three remarkable features: a very large decrease of the variance explained by gender, age and place of birth, and an increase in the variance explained by surnames, both as a proxi of ethnicity (in $CatalanDegree$) and once ethnicity has been controlled for. Thus, there are three prominent features of our results.

Generalized increase in educational attainment. The first feature is that to be female, or to be born in a rural environment became much less important in order to explain the education of an individual. This can be observed in the very large fall in the R^2 of the all the regressions when we use only the younger population. Particularly, it is clear from column (1) in tables 10(b), 11(b) and 12(b) that the percentage of the total variance explained by the individual controls is much lower. The most likely explanations for this are (1) the widespread increase in education attainment and investment in public education over the period, and (2) the big improvement of the status of females. Thus, nowadays people from different locations and different genders have much less dissimilar fates than generations ago.

This does not refer to the parents of the individuals. It does not refer to the specifics of their upbringing and their parent's status. In our analysis this is reflected in the variance explained by surnames, both in $CatalanDegree$ (as a proxy of ethnicity) and directly (as an approximation to family networks).

Increase in the importance of ethnic background. The second feature of the data are the increase of the importance of the ethnic component, reflected both in the point value of $CatalanDegree$. Notice that the increase of the importance of $CatalanDegree$ can not be a direct and obvious consequence of the migration process. In our regressions we include not only the children of the immigrants, but also the immigrants themselves. Thus, $CatalanDegree$ is more important to determine the education of the second than of the first generation of emigrants. It is not the case that it increases because there are more emigrants. The explanation must lie in the workings of the education system and the social composition of Catalonia. Notice that this does not mean that low $CatalanDegree$ agents have obtained less education, but that their difference vis a vis high $CatalanDegree$ agents has increased. The increase in educational attainment has been large across the board, for both ethnic groups, but has affected Catalan speakers more than Spanish speakers.

In Catalonia there are two linguistic communities, Catalan and Castillian (Spanish), each represent roughly half the population. Catalan speakers have enjoyed substantially larger incomes and larger levels of educational attainment during all the period of our study (this is true for both the born before and after 1950). Nevertheless before the final years of the 70's there was no obvious formal linguistic advantage toward Catalan speakers, as the language of the administration, commerce and education (both public and private, even if there were a few exceptions on the latest) was overwhelmingly Spanish. From the late 70's onwards the increasing political power of the Catalan nationalism[57] has translated into a series of drastic legal and administrative reforms that have turned upside down the relative importance of both languages in society while changing only marginally its overall language composition. In the respect that we care, since the beginning of the 80's all education is provided *exclusively* in Catalan in all public and practically all private schools. Catalan is now the sole language of administration, and proficiency in Catalan has been the key requirement to work in public administration since the beginning of the 80's. Further legal change has made Catalan an important (albeit perhaps not the main) business language. Still, even if being suggestive of possible causalities this is at most a partial explanation. A deep understanding of the increase in the value of ethnicity is beyond the scope of the present paper and demands of further work. We can nevertheless add some insight by looking at assortative mating. In section 4.10 we show that assortative mating across ethnic lines increased and can explain at least partially this result.

Decrease of Intergenerational Mobility. The third and most notorious (from the viewpoint of this paper) feature of the data are that the variance explained by surnames has increased *even once we have conditioned on* $CatalanDegree$. This is probably the most important of our empirical results. Particularly remarkable in the light

[57] The increasing power of Catalan Nationalism is explained (1) by the larger levels of income and education of the Catalan speaking community and (2) because Spanish electoral law has allowed Catalan nationalism to operate as a third party in Spanish politics, allowing it to obtain high leverage from its successive alliances with either left or right leaning governments. See Miley (2004) for a study of the politics of nationalism and language in Catalonia.

of the enormous increase of spending in public education during the period. The informative content of surnames has increased a 30%. This is strongly suggestive of a fall in the degree of intergenerational social mobility, and an increase in the value of family background.

There are two possible explanations for this. The first one would be that the decrease in mobility is a consequence of the increase in the provision of public education. This would be in line with Checchi, Ichino, and Rustichini (1999) who suggest it to explain differences between Italy and the US. It could also be the result of misperceptions on the return to education by the poor (like in Piketty (1995)). In those models an increase in the provision of public education may induce a decrease in the degree of intergenerational mobility. The intuition is as follows. An increase in the provision of public education can be reinterpreted as a subsidy on education, reducing its price across the board. The decrease is the same for all individuals; the rich and the poor; the children of the highly educated, and of those who got no education at all. This should be reflected in an average increase of educational attainment. Nevertheless not all individuals have the same demand of educational services. In particular, given a price of education, the demand of education is likely to be larger for the children of the already educated. Not just because they are richer, but because they have more interest, motivation, incentive, role models, etc. The result is that the children of the (rich and) educated are more likely to make a good use of the subsidy. They just get more out of it. Thus, even if the average individual gets more education, the fact that your parents were rich, or poor, might become more important in order to determine your education.

We have no evidence in favor or against this explanation. It might be the case, but our methodology does not offer a way of testing for it. There is though a second possible explanation, and our methodology does provide with a way of testing it.

The second explanation is the one suggested in section 3.5: that the increase in the ICS is a consequence of an increase in the degree of assortative mating one generation before. On this we have direct evidence which is quite conclusive that this mechanism is at least an important explanation of the decrease of intergenerational mobility. As we will see in section 4.10 making use of the specifics of the Spanish naming convention allows us to measure an increase in the degree of assortative mating timed one generation before the increase of the ICS.

4.10 Assortative Mating

In this section we show that an increase in assortative mating is a possible explanation of the observed increase in the importance of background. Both, for what respect to the increase in importance of ethnicity and for those things that are family specific.

Our strategy consists in checking the amount of information that the characteristics of the first surname of an individual has on the characteristics of the second. Given that the first is inherited from the father (and is informative on his characteristics), and the second from the mother (and refers to her characteristics), we can measure how similar are the characteristics of the father to the characteristics of the mother.

As we have seen surnames provide with information about its holder. Thus, we can reinterpret this result

Table 14: Assortative Mating in Education and Catalan Degree over cohorts.

(a) AM in Education

	EduSurname2	
	"Old"	"Young"
EduSurname1	$0.170_{(0.001)}$	$0.303_{(0.001)}$
Observations	2,041,044	2,222,917
R^2	0.3410	0.1997

(b) AM in CatalanDegree

	CatDegreeSurname2	
	"Old"	"Young"
CatDegreeSurname1	$0.217_{(0.001)}$	$0.328_{(0.001)}$
Observations	2,041,044	2,222,917
R^2	0.5110	0.2778

All regressions include individual and place of birth controls. Notes: Standard errors in parenthesis. Source: Catalan Census. Sample: Individuals born before ("old") and after ("young") 1950, respectively, among Spanish citizens in Catalonia aged ≥ 25 with frequency of first surname and second surname> 1.

as indicating that the first surname of an Spaniard provides with information about his father, and the second surname provides with information about his mother.

For doing the regressions that in Table 14 we associate to each surname the average education of its holders (in subtable 14(a)) and to the the *CatalanDegree* of the holder (in subtable 14(b)). We then run two sets of regressions.

In table 14(a) we run the average education of the first surname of the individual again the usual individual controls and the average education of the second surname. The larger is the correlation, the more information has the measurable characteristics of the mother of an individual on the measurable characteristics of the father. Notice that we are referring not to what the individual does, but to what her parents did. Nominally, to marry each other instead of somebody else. A large positive value of this correlation indicates that among the parents of the agents there was a large degree of assortative mating along the educational dimension.

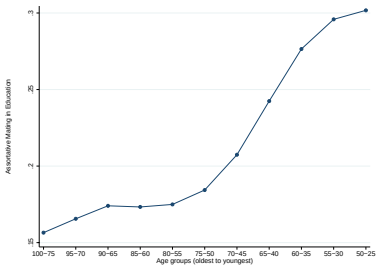
In the first column we run the regression for the individuals born before 1950. Thus, the correlation that we obtain (0.17) refers to the parents of those individuals. In he second column we run a regression with the individuals born after 1950. The correlation is much larger (0.303). Thus the parents of the younger group of agents were more prone to marry individuals of similar educational level than their grandparents (the parents of those born before 1950). Notice that this is an effect that refers to the parents of the individuals, not to the individual herself. Of course it has effects on the individual, but they are the ones that we have analyzed in the previous sections.

In table 14(b) we do the same exercise with *CatalanDegree* instead of education. We obtain a similar result. There is an increase in the degree of assortative mating across ethnic lines from the parents of the "old" to the parents of the "young", even though this increase is less pronounced than in the educational dimension.

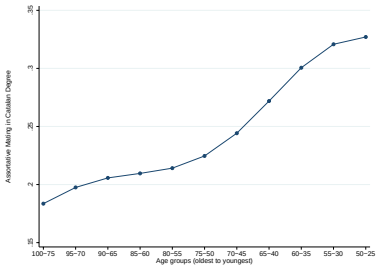
In order to give a graphical representation of these results we plot in figure 15 the value of the parameters for the regression of education (figure 15(a)) and for the regression on *CatalanDegree* (figure 15(b)) on the usual moving window of cohorts. Clearly, they are both increasing, and education has a timing which particularly resembles the timing of the increase in ICS.

In order to make sure that these results are not driven by ethnicity we run the same regressions on ethnically homogeneous populations (figure 16(a)) and with very unfrequent surnames (figure 16(b)). The results are qualitatively identical.

Thus, an increase in the degree of assortative mating may explain the increase in the value of inheritance that



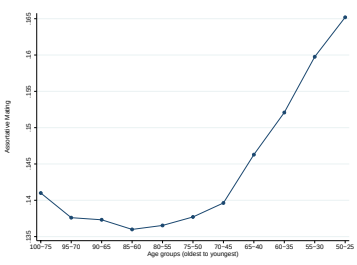
(a) Assortative Mating in Education



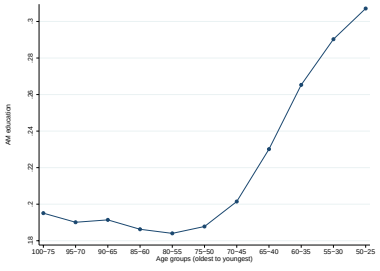
(b) Assortative Mating in *Catdegree*

Figure 15: Evolution of AM in Education and *CatalanDegree* over (a moving window of) cohorts. Complete Population.

Source: Catalan Census. Sample: Spanish citizens in Catalonia aged ≥ 25 with frequency of first surname and second surname> 1.



(a) 45% Most Catalan Surnames



(b) 50% Least Frequent Surnames

Figure 16: Evolution of AM in Education and *CatalanDegree* over (a moving window of) cohorts, in subpopulations.

Source: Catalan Census. Samples: Individuals with 45% Most Catalan Surnames and individuals with 50% Least Frequent Surnames of sample in figure 15(a), respectively.

we have observed in the previous sections.

5 Summary, assessment and conclusions.

The goal of this paper is to demonstrate that the amount of information contained in surnames is indicative of the degree of intergenerational mobility of an economy; and thus, that it is possible to obtain meaningful comparisons of intergenerational mobility by measuring it. In this way we hope to help freeing the study of intergenerational mobility from the slavery of longitudinal data.

To this aim we have developed a theoretical model relating ICS and inheritance, and have studied the information contained in the surnames of the inhabitants of a large Spanish region. Doing this, we believe to have shown that surnames do capture family links in a manner that allows them to be used to extract longitudinal information from census (cross sectional) data. It is not only ethnicity, but also (and most critically) information related to the immediate surroundings, and specificities of background. This is because surname distributions are necessarily skewed, implying that for a large percentage of the population the surname is informative on family connections.

The point of the theoretical exercise is to show *why* the distributions are skewed, and how does this relate to ICS. Notice that even if we did not know these reasons, conditional on having an skewed surname distribution we would know that ICS would increase with the value of inheritance. As a matter of fact surname distributions *are* skewed, which allows us to make an empirical study.

In this respect, to show (1) that in Catalonia intergenerational mobility has decreased, and (2) that this has been caused (at least partially) by an increase in the degree of assortative mating, has a value beyond the inherent one of learning what is going on in Spain.[58] We know very little about mobility trends around the world, and much less about its causes. There is an ongoing debate on what are the trends (if any) in the US, and what might be causing them. We believe that our results for Catalonia may help by indicating that in at least in another developed economy there has been a clear decrease in mobility caused by an increase in assortative mating.

Still, our preferred reading of the results is an indirect one. They show that our method is workable. We are able to give a coherent view of the evolution of mobility in a complex, changing society. Outside Spain it would be difficult to replicate our complete empirical exercise, as the second surname would be missing; but what we want to remark that our results are the same when we use only the first surname (once controlling by ethnicity). The rest of our results, and the evidence on assortative mating, are just reassurances necessary in this, the first paper using this method. They show that (1) looking at the correlation of education between siblings does not change the qualitative results, and that (2) we should not be surprised of finding a decrease in mobility, as it is explained by a previous increase in assortative mating. Had we looked only to the first surname, the same decrease of mobility would have been reported.

[58] Which is of enormous interest to us, but unlikely to excite the average reader.

5.1 Some ongoing work and possibilities of future research.

There are several directions in which we plan to expand the present work.

On one hand there is the obvious expansion of the area studied. There is an obvious interest in understanding how intergenerational mobility relates to other macroeconomic outcomes, like wage/asset inequality or per capita GDP growth. A relatively easy way of looking at that is studying the evolution of this variables across Spanish regions, as there is substantial variance to be exploited and the database is (1) easily accessible and (2) very coherent and compressive. We are presently realizing work in this direction.

Our long term objective is to provide with comprehensive and comparable measures of intergenerational mobility across countries and time. This is going to be useful only in the measure that helps us understand how intergenerational mobility affects and is affected by economic activity and policy. The obvious next step is to get access to census data from OECD countries (that will be relatively easy to compare) and develop the work of the present paper across countries. We believe that it would be even more useful to determine the degree of ICS in LDC and compare it with the one obtained in OECD. The objective would be to make an assessment of how intergenerational mobility moves with growth and economic development, and how it varies across countries with different levels of income, openness, etc.

Whenever we make comparisons across countries or regions we need to deal with the fact that the surname distributions will be different. We believe that there are two ways of tackling this problem. One is to calibrate a more extensive (OLG) model to the data, and to impute from it the degree of mobility. In general this would allow to use more complex (non linear) income transmission processes, and perhaps look beyond ρ.

A second approach would be to realize in the model Monte Carlo simulations to see what is the size of the queue of the surname distribution that allows to make coherent comparisons. This would exploit to the maximum the skewness of the distribution, and would make the comparisons only with the individuals who have surnames unfrequent enough.

As we have stated, surnames (their informative content) can also be used to determine the degree and speed of integration of immigrants in a recipient country. In this paper we have presented a measure of it in the form of our variable $CatalanDegree$, but further theoretical work is necessary on this respect. Likewise in order to study the degree of discrimination of different ethnic groups. Presently we are working in both of these issues.

Finally, there are two further directions in which we consider possible to develop our work. First, by looking for alternative measures (comparing the distribution of surnames across professions, income groups, etc.) in the manner already explained. Second, exploring the possibilities of our methodology in measuring the degree of inheritability of non economic traits, particularly for health affections using medical data. We are at a preliminary data stage in both.

References

Aaronson, D. and B. Mazumder, 2007. Intergenerational economic mobility in the U.S., 1940 to 2000. Federal Reserve Bank of Chicago WP 2005-12.

Angelucci, M., G. De Giorgi, M. Rangel, and I. Rasul, 2007. Family networks and school enrolment: Evidence from a randomized social experiment. Mimeo UCL.

Bagüés, M. F., 2005. ¿Qué determina el éxito en unas oposiciones? FEDEA, Documento de Trabajo 2005-01.

Becker, G. S., 1967. Human capital and the personal distribution of income: An analytical approach. *Woytinsky Lecture, Institute of Public Administration, University of Michigan* (1).

Becker, G. S. and N. Tomes, 1979. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy 87*(6), 1153–1189.

Becker, G. S. and N. Tomes, 1986. Human capital and the rise and fall of families. *Journal of Labor Economics 4*(3 part 2), S1–S39.

Behrman, J. and P. Taubman, 1985. Intergenerational earnings mobility in the United States: Some estimates and a rest of becker's intergenerational endowments model. *Review of Economic and Statistics 67*, 144–51.

Behrman, J. and P. Taubman, 1990. The inergenerational correlation between children's adults earnings and their parents' income: Results from the Michigan Panel Survey of Income Dynamics. *Review of Income and Wealth 2*(36), 115–27.

Bertrand, M. and S. Mullainathan, 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review 94*(4), 991–1013.

Björklund, A., T. Eriksson, M. Jäntti, O. Raaum, and E. Österbacka, 2002. Brother correlations in earnings in Denmark, Finland, Norway, and Sweden compared to the United States. *Journal of Population Economics 15*(4), 757–772.

Björklund, A. and M. Jäntti, 1997. Intergenerational income mobility in Sweden compared to the United States. *The American Economic Review 87*(5), 1009–1018.

Blanden, J., A. Goodman, P. Gregg, and S. Machin, 2004. *Changes in Intergenerational Mobility in Britain.* Cambridge University Press. in Miles Corak (ed.) Generational Income Inequality.

Cabré, A., 2004. La aportación de los 'otros' catalanes. *El País (Edición Barcelona)* (06/09/2004), 6.

Chadwick, L. and G. Solon, 2002. Intergenerational income mobility among daughters. *The American Economic Review 92*(1), 335– 44.

Checchi, D., A. Ichino, and Rustichini, 1999. More equal but less mobile? Education financing and intergenerational mobility in Italy and in the U.S. *Journal of Public Economics 74*(3), 351–93.

Collado, M. D., I. Ortuño-Ortín, and A. Romeo, 2006. Vertical transmission of consumption behavior and the distribution of surnames. Mimeo.

Comi, S., 2003. Intergenerational mobility in Europe: evidence from ECHP. Mimeo.

Couch, K. A. and T. A. Dunn, 1997. Intergenerational correlations in labor market status: A comparison of the United States and Germany. *Journal of Human Resources*, 210–32.

Dahan, M. and A. Gaviria, 2001. Sibling correlations and intergenerational mobility in Latin America. *Economic Development and Cultural Change, University of Chicago Press* (3), 537–54.

Darwin, G. H., 1875. Marriages between first cousins in England and their effects. *Journal of the Statistical Society*, 153–84.

Dearden, L., S. Machin, and H. Reed, 1997. Intergenerational mobility in Britain. *Economic Journal 107*, 47–64.

Duncan, O. D., D. Featherman, and B. Duncan, 1972. *Sociological Background and Achievement.* New York: Seminar Press.

Dunn, C., 2004. The intergenerational transmission of earnings: Evidence from Brazil. Mimeo.

Ermisch, J., M. Francesconi, and T. Siedler, 2006. Intergenerational mobility and marital sorting. *Economic Journal 116*, 659–679.

Ferreira, S. G. and F. A. Veloso, 2004. Intergenerational mobility of wages in Brazil. Mimeo.

Fertig, A. R., 2007. Trends in intergenerational earnings mobility in the US. *Journal of Income Distribution 12*, 108–130.

Fryer, R. and S. Levitt, 2004. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics 119*(3), 767–805.

Grawe, N. D., 2004. *Intergenerational mobility for whom? The experience of high- and low-earnings son in international perspective.* in Miles Corak (ed.) Generational Income Inequality, Cambridge University Press.

Haider, S. and G. Solon, 2006. Life-cycle variation in the association between current and lifetime earnings. *The American Economic Review 96*(4), 1308–1320.

Hertz, T., 2007. Trends in the intergenerational elasticity of family income in the United States. *Industrial Relations 46 (1)*, 22–50.

Hertz, T. N., 2001. Education, inequality andeconomic mobility in South Africa. Ph.D. thesis, University of Massachusetts.

Holmlund, H., 2007. Intergenerational mobility and assortative mating: Effects of an educational reform. Mimeo, CEP, LSE.

Lam, D. and R. F. Schoeni, 1993. Effects of family background on earnings and return to schooling: evidence from Brazil. *Journal of Political Economy 101*(4), 710–40.

Lasker, G. W., 1985. *Surnames and genetic structure.* Cambridge: Cambridge University Press.

Levine, D. I. and B. Mazumder, 2007. The growing importance of family: Evidence from brothers' earnings. *Industrial Relations 46 (1)*, 7–21.

Levitt, S. and S. Dubner, 2005. *Freakonomics.* HarperCollins.

Lillard, L. A. and M. R. Kilburn, 1995. Intergenerational earnings links: Sons and daughters. Papers 95-17, RAND - Labor and Population Program.

Manrubia, S. C. and D. H. Zanette, 2002. At the boundary between biological and cultural evolution: The origin of surname distributions. *Journal of Theoretical Biology 261*(4), 461–477.

Mayer, S. E. and L. M. Lopoo, 2005. Has the intergenerational transmission of economic status changed? *Journal of Human Resources 40*(1), 169–85.

Miley, T. J., 2004. The politics of language and nation: The case of the catalans in contemporary Spain. Ph.D. thesis, Department of Political Science at Yale University.

Ng, I., 2007. Intergenerational income mobility of young singaporeans. in YouthSCOPE. Ho, Kong Chong (ed.). Singapore: National Youth Council.

Ortega, J., 2007. Determinantes del nivel de catalán de los inmigrantes en cataluña: un análisis de sección cruzada a nivel comarcal. *Cuadernos de ICE*.

Osterbacka, E., 2001. Family background and economic status in Finland. *Scandinavian Journal of Economics 103*(3), 467– 84.

Osterberg, T., 2000. Intergenerational income mobility in Sweden: What do tax data show? *Review of Income and Wealth. 46*(4), 421–36.

Page, M. E. and G. Solon, 2003. Correlations between brothers and neighboring boys in their adult earnings: The importance of being urban. *Journal of Labor Economics 21*, 831–55.

Piketty, T., 1995. Social mobility and redistributive politics. *The Quarterly journal of economics 110*(3), 551–584.

Solon, G., 1992. Intergenerational income mobility in the United States. *The 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.

Solon, G., 2004. *A Model of Intergenerational Mobility Variation over Time and Place.* Cambridge University Press.

Solon, G., M. Corcoran, Roger, and L. Deborah, 1991. A longitudinal analysis of siblings correlations in economic status. *Journal of Human Resources 26*, 509–34.

Wiegand, J., 1997. Intergenerational earnings mobility in Germany. Mimeo.

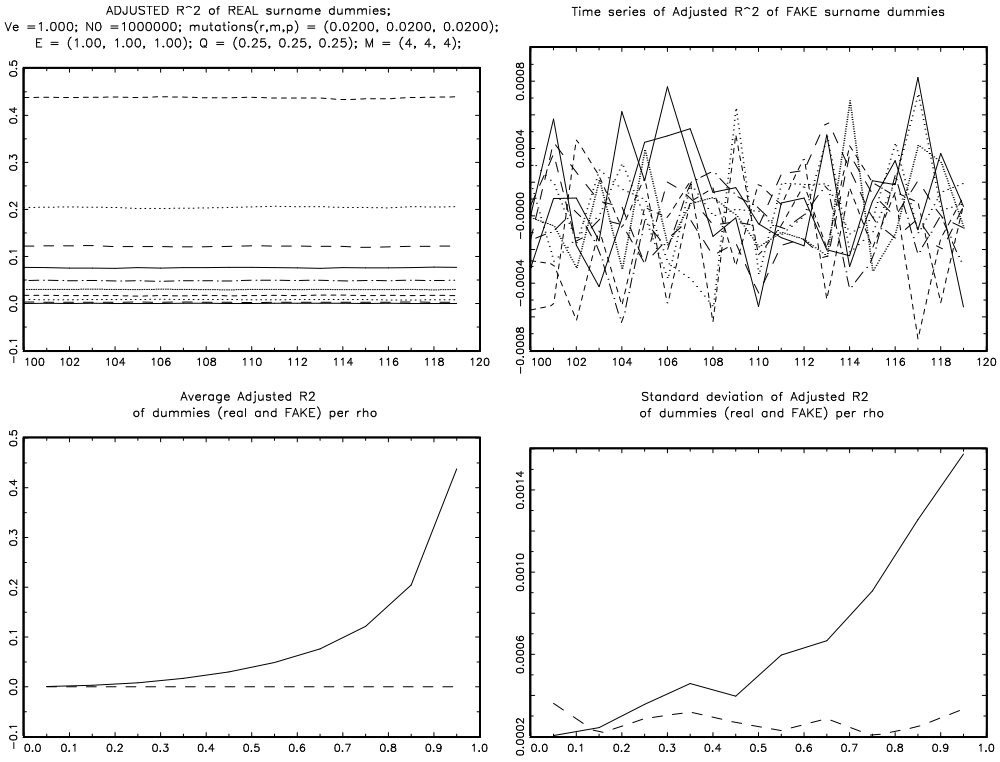


Figure 17: High Family Variance

A The relationship between the value of inheritance and ICS is robust to different values of the conditional variance of income, the mutation rate of surnames and family size.

In figure 17 we plot the equivalent to figure 1 for a fertility process with higher family variance, finding no qualitative differences.

Figure 18(a) plots the equivalent to figure 1 for an income process where the conditional variance of income is much larger (V_e is 10, instead of 1), while figure 18(b) does it for a much smaller value of the conditional variance. Things are not different from figure 1 in any qualitative manner.

Finally in figure 19(a) we plot the same graph than figure 1 except for having a larger mutation rate (again ten times larger) while figure 19(b) uses a much lower mutation rate; again there are no qualitative differences. With larger mutation rates the magnitude of the effects is larger (in particular for low values of ρ), as there are more uncommon surnames, but the qualitative results are the same. Notice that the results are robust *even with very small values of* μ, as this generates enough surname variation.

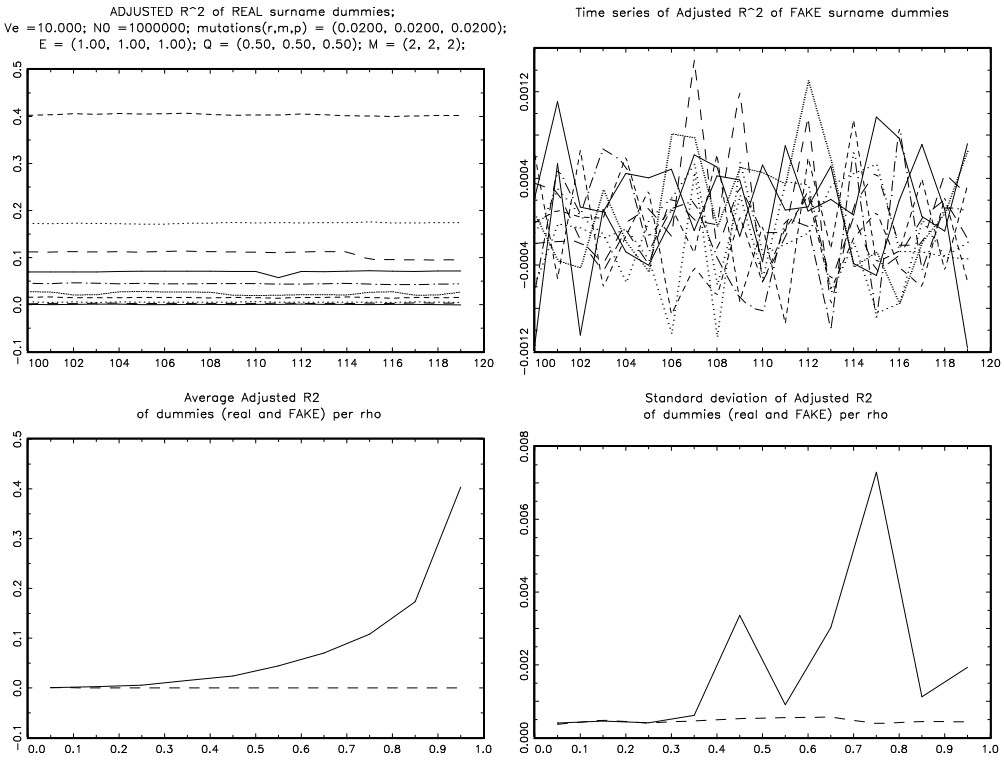
B Difference in fertility across income groups

In figures 20(a), 20(b) and 20(c) we show the results of a simulation that the only thing that changes with respect to our benchmark simulation is that the expected number of children differs among the income groups (even if the probability of having male offspring is the same for all of them, $Q_j = \frac{1}{2} \quad \forall j$). In this simulation $E_r = 1.5; \quad E_m = 1; \quad E_p = \frac{1}{2}$.

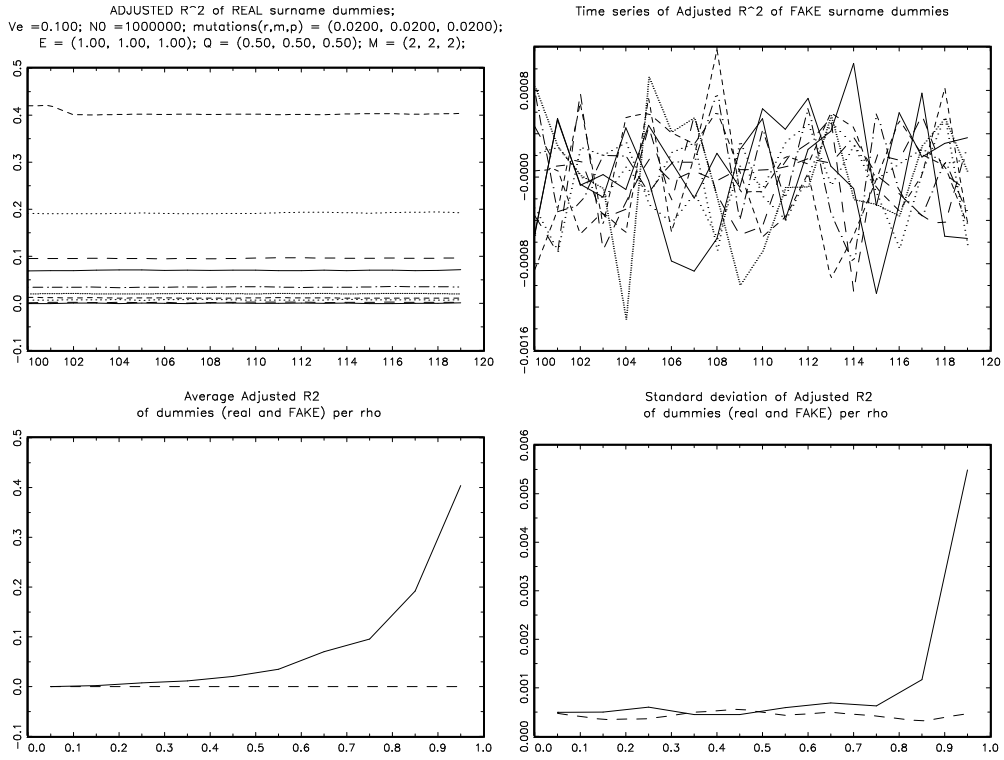
In figure 20(b), we observe that the t-statistic of frequency is always positive, significantly different from zero, and it increases with inheritance. The reason being that rich people have more kids, which makes surname more common.

Notice that also in this case the distribution of surnames is affected by inheritance. In figure 20(c) more inheritance implies a larger Gini index and a smaller number of surnames per person. This is because with more inheritance rich people lineages become large. Of course they can not be all rich (as the definition of "rich" and "poor" is relative), so the less fortunate between them moves down to lower incomes. *Their* lineages do not disappear, even if the probabilities of having male descendants decrease substantially, as their rich cousins share their surname with them. The mutations that happen among the poor would be short living, the mutations among the rich will survive by making their surname large.

Finally in figure 20(a) we meet again with our main result. Irrespectively of if frequency of the surname is positively (as in this case) or negatively (as in the previously) associated to inheritance, it is always the case that more inheritance translates into a larger informative content of surnames. This is because ICS refers to family bonds, while frequency has information because the shape of the distribution of surnames is a function of income distribution once lineage birth/death probabilities depend on the income of the agents.

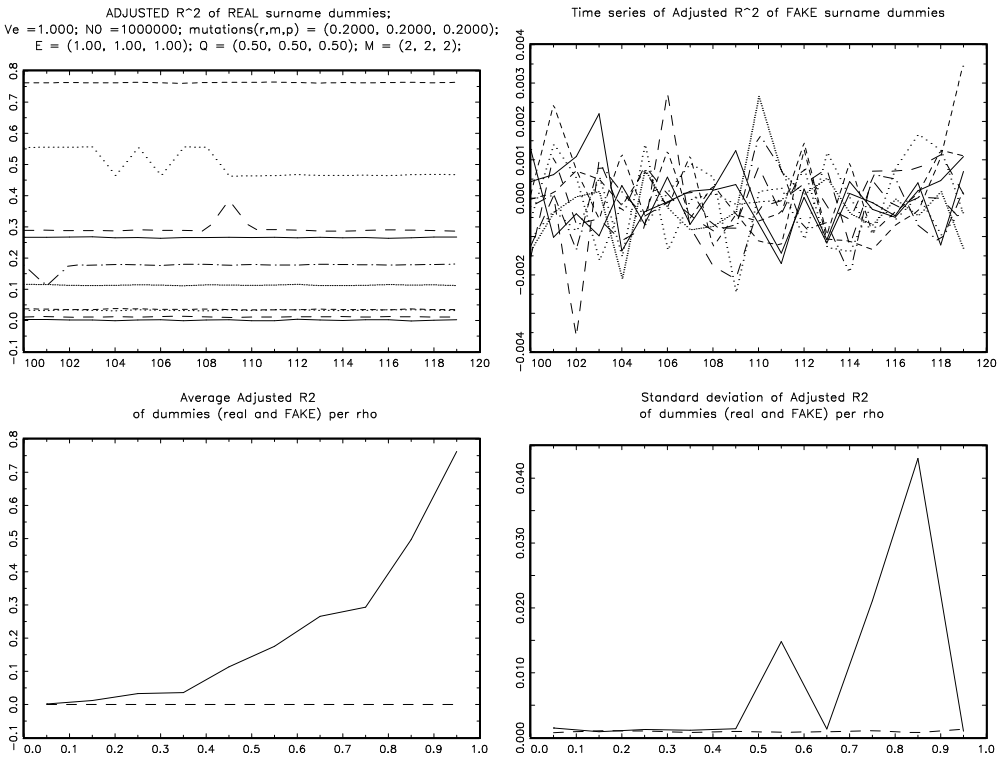


(a) High Conditional Variance

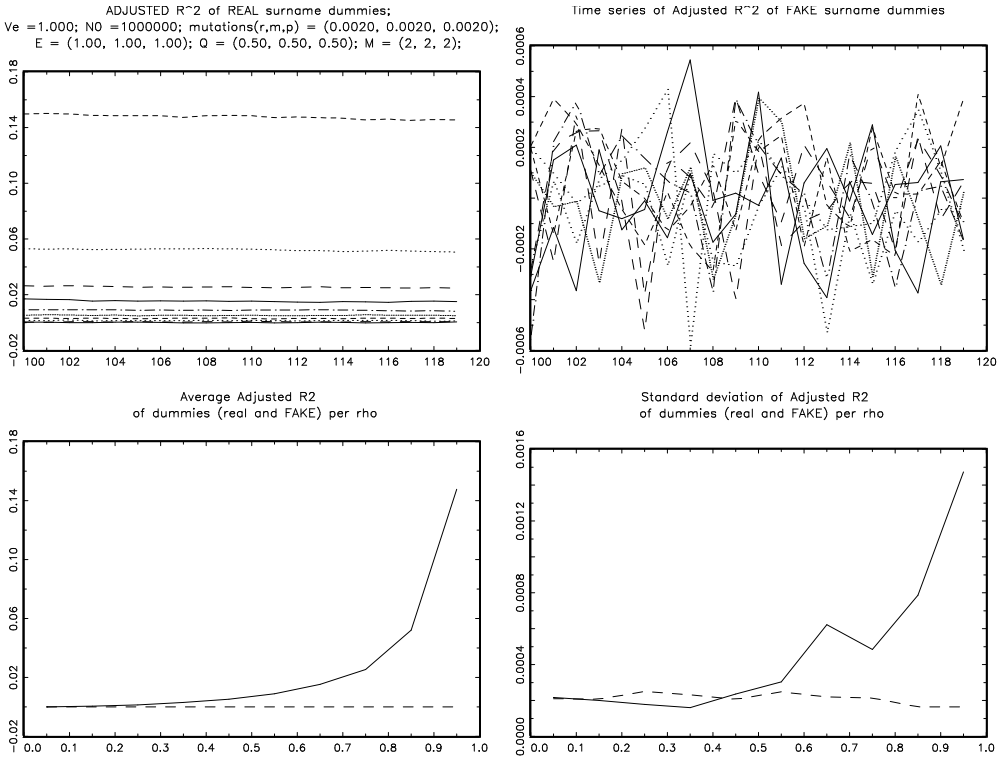


(b) Low Conditional Variance

Figure 18: Differences in V_e

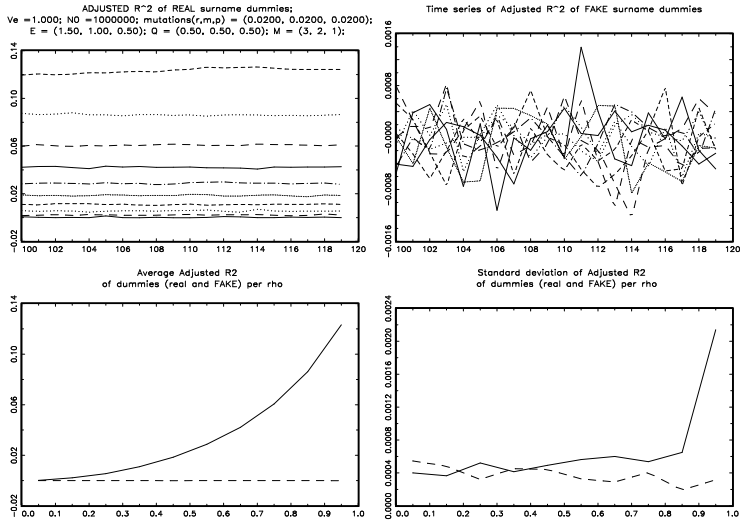


(a) High Mutation Rate

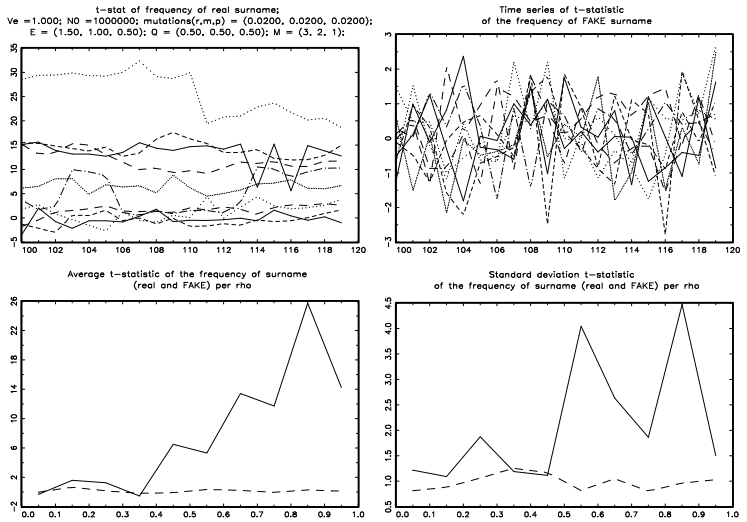


(b) Low Mutation Rate

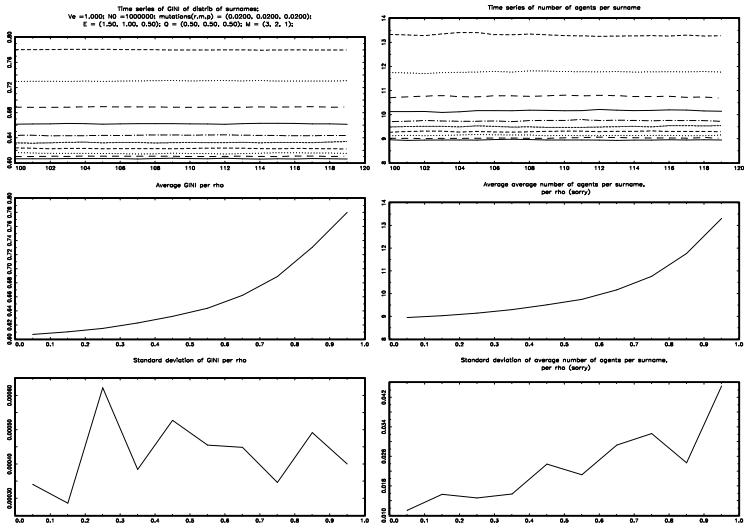
Figure 19: Differences in μ



(a) Fertility differences Adjusted R2



(b) Fertility differences frequency



(c) Fertility differences distribution

Figure 20: Differences across socieconomic groups in the Average Fertility

C Proof of Result 1

We will show this in steps. We start by assuming (later to be proved) the existence of a well defined steady state, and V (the value of the unconditional variance of both x and y in (7)).

First notice that in spite of what might appear from (6), the expected income of the mother conditional on the income of the father is not equal to the income of the father, but smaller than it. The reason is that one needs to apply Bayes Rule to the two sources of information available (the unconditional process (7) and the matching process (6)). The income of the wife conditional on the husband's is also normally distributed, but is "mean reverting" with respect to the husband's income:

$$x|y \sim N\left(\frac{V}{V+\sigma_u}y, \frac{V\sigma_u}{V+\sigma_u}\right) \tag{12}$$

if the variance of the matching process is much smaller than the variance in the population, then the expected income of the wife is almost as large as the income of the husband. If the variance of incomes in the population is large (relative to the mating process), then the expected value is close to the population mean (that in our case is zero).

We can rewrite (12) as $x = \frac{V}{V+\sigma_u}y + \omega$; with ω being noise with variance $\frac{V}{V+\sigma_u}\sigma_u$.

Second, notice that the income of a male child is related to the income of his father via the direct link (5) and via the knowledge that the father's income has on the mother's income (12).

$z|y \sim N\left(\frac{1+\frac{V}{V+\sigma_u}}{2}y, \frac{1}{4}\frac{V}{V+\sigma_u}\sigma_u\right)$ or, alternatively: $z = \frac{1+\frac{V}{V+\sigma_u}}{2}y + \frac{1}{2}\omega$

And given inheritance (5), $y'|y \sim N\left(r\frac{1+\frac{V}{V+\sigma_u}}{2}y, \frac{r^2}{4}\frac{V}{V+\sigma_u}\sigma_u + \sigma_e\right)$. This can be rewritten as :

$$y' = \frac{r}{2}\left(1 + \frac{V}{V+\sigma_u}\right)y + \frac{r}{2}\omega + e \tag{13}$$

which is identical to (8) with $\rho = \left(1 + \frac{V}{V+\sigma_u}\right)\frac{r}{2}$ and $\sigma_\varepsilon = \frac{r^2}{4}\frac{V}{V+\sigma_u}\sigma_u + \sigma_e$.

Third, notice that in steady state the distribution of income across males is the same in all generations, and has to be equal to the one that we have assumed (7).

From (13) we see that the unconditional distribution of y is a normal with zero mean and a certain unconditional variance that in steady state this needs to be equal to the unconditional variance that we have assumed (V):

$$V = \frac{\frac{r^2}{4}\frac{V}{V+\sigma_u}\sigma_u + \sigma_e}{1 - \left(1 + \frac{V}{V+\sigma_u}\right)\frac{r}{2}} \tag{14}$$

Fourth, We can substitute V for ρ in (14). With which after some manipulation we get:

$$\left[\rho - \frac{r}{2}\right] \times \left[\frac{1-\rho^2}{r-\rho} - \frac{r}{2}\right] = \frac{\sigma_e}{\sigma_u} \tag{15}$$

The LHS of (15) is monotonously increasing for $\rho < r$; it is equal to zero for $\rho = \frac{r}{2}$; it has an asymptote at $\rho = r$, approaching (positive) infinity as ρ approaches r from the left; and it is negatively valued for $\rho > r$.

Thus, there exists an unique value of ρ that solves (15). This is the steady state value of the correlation of income between a father and his children. Clearly, $\frac{r}{2} \leq \rho < r$, it is increasing in r (the LHS is decreasing in r for the relevant values of ρ) and σ_e, and it is decreasing in the variance of the mating process σ_u.

Fifth, merging ρ and σ_ε we get: $\frac{\sigma_e}{\sigma_u} = \frac{\frac{r}{2}\left(\rho - \frac{r}{2}\right)}{\frac{\sigma_\varepsilon}{\sigma_e} - 1}$ and substituting in (15) we get:

$$\left[\frac{\sigma_\varepsilon}{\sigma_e} - 1\right] = \frac{r}{2}\left[\frac{1-\rho^2}{r-\rho} - \frac{r}{2}\right]^{-1} \tag{16}$$

The RHS of (16) is positively valued at the solution, and decreasing in ρ. Thus, σ_ε is increasing in σ_u and larger than σ_e.

Finally, using ρ we get $V = \frac{1-\rho^2}{r-\rho}\sigma_u$. This determines that, as it was assumed, there exists an unique unconditional distribution of y (and by symmetry, of x).

QED.