UNIVERSIDAD DE MURCIA

Departamento de Fundamentos del Análisis Económico
Working Paper Series
Number 02, November, 2013

# Long Run Intergenerationa Social Mobility and the Distribution of Surnames

M.Dolores Collado (Universidad de Alicante)
Ignacio Ortuño-Ortín (Universidad Carlos III)
Andrés Romeu (*Universidad de Murcia*)

# Long-run intergenerational social mobility and the distribution of surnames*

M.Dolores Collado* Ignacio Ortuño-Ortín**
Andrés Romeu***
*Universidad de Alicante
**Universidad Carlos III de Madrid
***Universidad de Murcia

November 4, 2013

**Abstract**

We develop a novel methodology to analyze intergenerational social mobility over long periods of time when the precise ancestors of the individual cannot be identified. We base our approach on the incorporation of surnames in the analysis of social mobility, applying our methodology to assess the degree of intergenerational social mobility within two Spanish regions from the late 19th century to the beginning of the 21st century. The results show that the probability of having a high educational level, or of belonging to a high-status socioeconomic group, still depends on the socioeconomic status of the great-great-grandparents. Our analysis suggests, however, that such dependence will vanish in the next-to-present generation.

## 1 Introduction

Intergenerational socioeconomic mobility (ISM, henceforth) determines the degree to which individuals change their relative positions in the social hierarchy with respect to those of their

*We thank Manuel Arellano, Stéphane Bonhomme, Gregory Clark, Pedro Mira, and Luis Ubeda for very useful suggestions and comments. A previous version of this paper, entitled "Intergenerational social mobility in Spain since the late 19th Century" was presented in seminars at CEMFI, Salamanca, Collegio Carlo Alberto (Turin), Cardiff and Institute Juan March. The authors gratefully acknowledges financial support from Instituto Cántabro de Estadística and the Spanish MEC through grants ECO2010-19596, ECO2011-29751, ECO2010-19830 and Fundación Séneca 11998/PHCS/09. Collado also acknowledge financial support from Generalitat Valenciana grant PROMETEO/2013/037.

parents or ancestors. The degree of ISM can be seen as an indicator of the level of equality of opportunity in society. The absence of intergenerational mobility in socioeconomic status might arise through the effect of parental educational level and parental wealth on the offspring's educational levels and wealth. Other possible reasons are genetic inheritance and group effects. A large body of research tries to measure the level of intergenerational mobility. Economists have focused mostly on the measurement and analysis of intergenerational mobility in income or wealth–Solon (1999) provides one of the first surveys on the literature, Black and Devereux (2010) a more recent one, and Piketty (2000) provides an excellent survey on the theoretical models– and usually estimate an intergenerational elasticity coefficient, which measures the strength of the statistical correlation between parental income and offspring's income. Alternatively, sociologists have focused their analysis on the mobility in educational levels and occupations, with the standard approach using transition matrices to describe and measure social mobility (for a survey of this literature, see Erikson and Goldthorpe 2002).

Most of the estimates of ISM level use "one-generation data", i.e., data on some socioeconomic variable from a sample of individuals in a certain cohort and from their corresponding parents. Thus, for an already considerable number of countries there are good estimates of one-generation mobility in income and education during the second half of the 20th century (see the mentioned surveys and, for example, Solon 2002, Björklund 2009, Blanden 2011, and for cross-national comparisons in education mobility Hertz *et al.* 2007). In this paper we analyze intergenerational social mobility over long periods of time when the precise ancestors of the individual cannot be identified. We base our approach on the incorporation of surnames in the analysis of social mobility. The study of ISM across several generations is important for two reasons, first because we can learn whether the degree of social mobility has changed over time, and second because, as explain later, it can help us to understand if the standard one-generation estimates of social mobility are correct.

Our approach differs from most of the existing work on intergenerational mobility in two

main aspects. First, we try to measure the correlation between the socioeconomic status of individuals in the current generation and the socioeconomic status of their ancestors (through the paternal family line) at the end of the 19th century. Thus, we do not focus on the "one-generation case" although we will be able to say something about it and compare our findings with the existing literature. Second, we have data on the socioeconomic status of individuals in a population at the end of the 19th century and on the status of their descendants at the end of the 20th century. We do not know, however, who the descendants of any specific individual are. Nonetheless, we know the full name of all individuals from both generations. Thus, we develop a novel methodology for estimating the statistical association between the status of individuals in a population and the status of their corresponding descendants based on the use of information contained in the surnames. Surnames are useful in our analysis because their distribution and the distribution of socioeconomic characteristics among people are not independent, that is, there is a bias in the distribution of surnames among different socioeconomic groups. Collado et al (2008) characterize and quantify such bias in the distribution of surnames in Spain during the last years of the 20th century and also at the end of the 19th century. Güell et al. (2007) provide a similar result regarding the surnames in the region of Catalonia at the end of the 20th century.

We apply our methodology to the analysis of social mobility in the Spanish regions of Cantabria and Murcia. We argue, however, that our methodology can be applied to study long-run social mobility in any other country that also has the type of data used here, as long as there is a socioeconomic bias in the distribution of surnames. Our main data sets are the electoral census of 1898 and the 2001 population census of Cantabria. We will carry out some robustness exercises using a different data set, namely, the Yellow Pages of the telephone directory of 2004. In a second exercise we apply our methodology to study social mobility in the region of Murcia, for which we have only the electoral census of 1898 and the 2004 Yellow Pages but not the 2001 population census.

We will consider only two socioeconomic classes (although our methodology allows for

any number of classes and we have replicated all our results for the three socioeconomic classes case) corresponding basically to an upper class covering around 19% of the population with the highest educational level (or economic level in some cases), and a low class with the rest of the population. We estimate a transition matrix which indicates for individuals in each social class the probabilities that their descendants belong to any specific class. Our main conclusion is that our estimations strongly support the argument that the probability of belonging to the high-status class is still correlated with the socioeconomic status of the great-grandfathers and great-great-grandfathers. We show that for a (male) person at the end of the 19th century, the relative probability of having a descendant, via paternal line, of high-status class at the beginning of the 21st century over that of having a descendant of low-status class is on average around 30% higher for people of high-status class than for people of low-status class. We also compute, under certain assumptions, the average "one-generation" level of social mobility that is compatible with our findings on the three to four generation social mobility level. We compare such average one-generation level of social mobility with those reported by other authors for Spain (Klakbrenner and Villanueva 2006) and get approximately equal results. Comparing the one-generation transition matrix estimates with those obtained for other countries suggests that these Spanish regions have enjoyed levels of social mobility that are between the estimates for the United States and for Italy (Checchi et al. 1999). In section 6 we conduct an additional exercise with a methodology similar to that in Clark *et al.* (2012) and consider a linear equation to estimate an intergenerational elasticity coefficient. We obtain values around 0.4-0.5 for the average one-generation income or education elasticities. Thus, our findings here are within the range of values usually estimated for the one-generation elasticities (see Hertz *et al.* 2207). Although given our data constraints these are very rough estimates, they are in any case far away from those elasticities reported by Clark (0.7 for income and 0.78-0.81 for education). Thus, we argue that, contrary to Lindahl *et al.* (2012) and Clark *et al.* (2012), the elasticities estimated using data on several generations do not suggest that the standard one-generation

approach overestimates the long-run intergenerational social mobility.

The number of papers dealing with intergenerational mobility during the first part of the 20th century or earlier times is very limited. Important exceptions are the works by Ferrie (2005) and Ferrie and Long (2009), Lindahl *et al.* (2012), Clark (2012a, 2012b) and Clark *et al.* (2012). The first two authors study ISM in the United States and the United Kingdom since the late 19th century. However, they only compare occupational mobility from fathers to sons in the second half of the 19th century with occupational mobility from fathers to sons in the second half of the 20th century. Lindahl *et al.* (2012) analyze intergenerational mobility in income and education across several generations in Sweden. They use a data set containing information on individuals from four generations of the same family. Their results on the correlation between the educational level of individuals in the current generation and the educational level of their great-grandparents are broadly similar to ours, although they use a smaller sample size (around 800 Swedish families) and we analyze one more generation (great great-grandparents). They find that estimates of mobility from one generation overestimate the true mobility over more generations, suggesting that intergenerational transmission of genetic or behavioral factors–which can not be measured directly–is important and lasts more than one generation.[1] This is also the main finding in Clark (2012a, 2012b) and Clark *et al.* (2012), where it is argued that mobility in a series of countries has remained relatively constant over the last few generations, but that the degree of social mobility is much lower than the one implied by existing one-generation estimates. This is something that we can not directly test because we have no data on parents and children for all those generations, but our estimates of ISM for four generations are consistent with what other have found for the one-generation case. Thus, our results do not seem to give much weight to such genetic or behavioral factors[2] found in Lindahl *et al.*

[1] Sauder (2006) and Maurin (2002) also analyze the direct effect of grandparents on granchildren education.

[2] We focus primarily on educational mobility. Osborne (2005) and Mood et al. (2012) show that transmission of genetic and behavioral factors, as personality traits and physical characterisitics, are of little importance to understand transmission of eduactional status. Mood et al. (2012) argue, however, that personality traits play an important role in the labour market and part of the earnings transmission might

(2012), Clark (2012a, 2012b) and Clark *et al.* (2012).

The closest work to ours comes from Clark *et al.* (2012) and Clark (2012a, 2010b), who also use surname distributions in different centuries to measure long-run social mobility. Their methodology, however, is quite different since, contrary to our findings for Spain, in England the socioeconomic bias in the distribution of surnames disappeared as early as in the middle of the 17th century. Therefore, since surnames in England convey no information about social status, Clark has to rely on a different methodology based on linking different generations through rare surnames. One potential problem with this approach is the sample size, which in general is much smaller than in the case of considering all the surnames in a region. As mentioned above, in section 6 we conduct an additional exercise with a methodology similar to that of Clark and obtain results (which are consistent with the main findings of our paper) showing a higher degree of long run social mobility than the one reported in Clark *et al.* (2012) and Clark (2012a, 2010b).

Webber (2004) analyzes the geographic distribution of different types of surnames in some British regions to estimate that individuals of Scottish descendent have experienced more upward mobility than have descendants of Irish migrants. Thus, his work does not estimate the values of individual social mobility, and although it uses surnames it has a different goal and methodology from those of the present work.

Güell et al. (2007) use surnames to analyze long-run social mobility. Their approach is entirely different from that in this paper, as the former uses only one-generation data on the distribution of surnames and income. They also have to impose very strong assumptions on the dynamics and on the parameters of the model, which are difficult to verify in empirical work.

An additional difference from the standard approach studying intergenerational social mobility is that we also take into account the reproduction rates of the different social classes. Furthermore, our estimates of the parameters of social mobility together with the

be explained by this behavioral factor.

estimates of reproduction rates allow us to study how the current class composition depends on the social classes in the past. We show that after three or four generations, there is still an excess of agents with ancestors from the upper class among the current upper class population. Our analysis suggests, however, that the individual composition of social classes beyond the year 2030 will be basically independent of the individual class composition in 1898. Thus, our results can be seen as showing that there is still a 19th-century influence in today's society, but this influence will disappear shortly. It is important to recall that we focus our analysis exclusively on the influence via paternal lines. One would suspect that the total influence of all ancestors (throughout the maternal and paternal lines) is higher than the one we are able to quantify here.

# 2 The Model

Consider a society with no migration flows from outside. Suppose we just have data for two specific years, which we denote as year 1 and year $T$. In the empirical part of the paper 1 and $T$ will correspond to the years 1898 and 2001, respectively. Since we are considering an isolated society, all the individuals in year $T$ are descendants of the individuals in year 1.

Let $Y^1$ be the cohort of adult male individuals in year 1, i.e., the set of adult **male** individuals within a certain age bracket in year 1 such that none of them has an **adult** descendant in year 1. Define in a similar way $Y^T$ as the set of **adult** individuals in year $T$. Notice that $Y^T$ does not exclude female individuals. In some of our empirical findings we will also consider the case in which $Y^T$ contains exclusively male individuals, as in $Y^1$, and the case of only female individuals. We denote by $N^1$ and $N^T$ the total number of individuals in $Y^1$ and in $Y^T$, respectively. In the empirical part we approximate such set of "young adults" by the age bracket (22,46) for year $T$ and (25,45) for year 1.[3] The individuals in these two

[3]We use 25 as the minimum age in year 1 since individuals younger than 25 are not included in the electoral census. We choose a wider age bracket in year $T$ since individuals in the current generation tend to have children when they are older than their ancestors were a century ago.

sets $Y^1$ and $Y^T$ will be the object of study throughout the paper.

For each individual in $Y^T$ his/her *paternal ancestry lineage* (PAL) is given by his/her father, grandfather, great-grandfather, and so on (all of them in the paternal line). Note that for each individual in $Y^T$ there is only one individual in $Y^1$ belonging to his/her PAL, because if there were two individuals in $Y^1$ it would follow that one of them should be the son of the other and this cannot happen since the individuals in $Y^1$ have no descendant in $Y^1$. Therefore, for any individual in $Y^T$ there exists a unique individual in $Y^1$ in his/her PAL that we name his/her *ancestor*. Correspondingly, for any individual in $Y^1$ we consider as his *descendants* the set of individuals in $Y^T$ who have that individual in $Y^1$ as their ancestor. Importantly, we assume that the only data available that may help to link descendants and ancestors are the full names of all the individuals in $Y^1$and $Y^T$.

Our goal is to analyze the link between certain socioeconomic characteristics of the individuals in $Y^T$ and the characteristics of the corresponding ancestors in $Y^1$. Suppose that individuals in $Y^T$ and in $Y^1$ can be classified as belonging to one of the two social classes: high class ($H$) and low class ($L$). In the empirical section the criteria to define the two classes will be the level of education and the type of profession. Since, in general, years 1 and $T$ might be very distant in time the criteria to define the social classes in each of these two periods might be different. At this point, the restriction to two classes is a simplification to keep consistency with the empirical part of the paper where individuals are classified as belonging to one of two groups, but the model can be easily generalized to an arbitrary number of social classes or income groups. We assume that individuals remain in the same class during their entire lifetime, leaving aside issues on intra-generational social mobility.

Let us define the *reproduction rate* of an individual in $Y^1$ as the number of his descendants in $Y^T$. The reproduction rate of an individual in class $i$ is a random variable with expectation $r_i$ for $i \in \{H, L\}$. Notice that the reproduction rate is not equal to the fertility rate. In fact, $H$ type individuals could have a lower fertility rate than those of type $L$ and still the reproduction rate of the former could be higher due to a lower mortality rate.

The type of any individual in $Y^T$ is also a random variable that might depend on the type of his/her ancestor. We denote by $p_{ij}$ the conditional probability that an individual with ancestor of type $i$ is itself of type $j$. Note that $p_{ij}$ is also read as the probability that a given descendant of a person of type $i$ is of type $j$. Notice that since $p_{HL} = 1 - p_{HH}$ and $p_{LL} = 1 - p_{LH}$ it is enough to know the two probabilities $p_{HH}$ and $p_{LH}$. One important difference from the standard approach in the literature on social mobility is that one-generation studies aim to estimate the probability of being of type $j$ given that the father (or mother) is of type $i$. In our case, the probability is conditional on the type of the ancestor, which in general is not the father (in our empirical part the ancestry would be the great-grandfather and sometimes the great-great-grandfather). Obviously, when the ancestor coincides with the father our probabilities coincide with the one-generation measures of intergenerational mobility.

**Definition 1** *We say that a society displays perfect intergenerational social mobility if* $p_{HH} = p_{LH}$.

One way to assess the degree of social rigidity, or the lack of social mobility, is by measuring the odds ratio (OR) of those probabilities[4]:

$$\text{OR} = \frac{p_{HH}/p_{HL}}{p_{LH}/p_{LL}}.$$

Clearly, perfect intergenerational social mobility implies an odds ratio of 1.

Let $n_H^T$ and $n_L^T$ be the expected number of descendants of type $H$ and type $L$ of any individual in $Y^1$. Given the conditional probabilities $p_{HH}, p_{LH}$ and the expected reproduction rates $r_H, r_L$ we may compute $n_H^T$ and $n_L^T$ as:

$$n_H^T = p_{HH}\ r_H\ d_H + p_{LH}\ r_L\ d_L \tag{1}$$

[4]In the case of more than two social classes there would be more odds ratios. Ferrie et al. (2009) show how to asses the degree of social rigidity in that case.

$$n_L^T = p_{HL}\ r_H\ d_H + p_{LL}\ r_L\ d_L \tag{2}$$

where $d_H$ ($d_L$) is a dummy variable that takes value 1 if the ancestor is of type $H$ (type $L$) and zero otherwise. If we aggregate equations (1) and (2) for the entire population in $Y^1$, we have

$$N_H^T = p_{HH}\ r_H\ N_H^1 + p_{LH}\ r_L\ N_L^1 \tag{3}$$

$$N_L^T = p_{HL}\ r_H\ N_H^1 + p_{LL}\ r_L\ N_L^1 \tag{4}$$

where $N_H^1$ ($N_L^1$) is the number of individuals of type $H$ (type $L$) in $Y^1$, and $N_H^T$ ($N_L^T$) is the expected number of individuals of type $H$ (type $L$) in $Y^T$.

Then, we can define an alternative measure of social mobility that takes into account reproduction rates and is based on the previous equations.

**Definition 2** *We define the expected outflow ratio $F_{HH}$ ($F_{HL}$) as the percentage of individuals in class $H$ (class $L$) with ancestor of type $H$:*

$$F_{HH} = \frac{p_{HH}\ r_H\ N_H^1}{N_H^T} \times 100 \tag{5}$$

$$F_{HL} = \frac{p_{HL}\ r_H\ N_H^1}{N_L^T} \times 100 \tag{6}$$

Using the outflow ratio we can study how the current individual class composition depends on the individual composition of the social classes in the past:

**Definition 3** *A society presents history independent class composition (HICC) if both outflow ratios coincide with the percentage of individuals of type $H$ in $Y^1$, i.e.:*

$$\frac{N_H^1}{N^1} \times 100 = F_{HH} = F_{HL}$$

Note that perfect intergenerational social mobility does not imply HICC. Thus, in a society with equal conditional probabilities ($p_{HH} = p_{LH}$) and different reproduction rates

the proportion among $H$-type individuals in $Y^T$ of individuals with ancestor of type $H$ could be different from the proportion of individuals of type $H$ in $Y^1$. In this case, even though the society presents perfect intergenerational social mobility, the individual composition of the different classes in year $T$ would not be independent of the individual class composition in year 1.[5]

If we had data to determine the class of each individual in $Y^T$ and $Y^1$ (or data on a large enough sample of them) and we knew the ancestor of each individual in $Y^T$, we could easily estimate the probabilities $p_{ij}$ and the reproduction rates $r_i$ by applying indirect ordinary least squares (OLS) to equations (1) and (2). Then we could use these estimates to assess the degree of intergenerational social mobility and whether HICC holds. Unfortunately, for large values of $T$, data on the ancestry of agents in $Y^T$ are rarely available. In some cases, however, there are data on the surnames of the individuals in $Y^1$ and in $Y^T$. Such information may come, for instance, from the population census. Population census data are available for many countries for several periods and our method is specially designed to estimate the parameters of interest of the ISM using this type of data. Thus, we propose an "indirect" way of estimating the degree of ISM based on the use of surnames.

## 2.1 Surnames

Suppose that all the individuals in society bear a unique surname[6] that is inherited from the father. Thus, all the individuals in the same PAL bear the same surname. In most cases, however, the surname is not enough to identify the ancestor in $Y^1$ of an individual in $Y^T$ as the same surname may be shared by different PAL. Nonetheless, the surname for each individual in $Y^T$ delimits the set of that individual's potential ancestors. Now we show that,

[5] Clark (2007) argues that for many generations in England the rich had a higher reproduction rate than the poor had. This implied a downward social mobility that made possible the emergence of the Industrial Revolution.

[6] In Spain people have two surnames. The first one is the first surname of the father and the second one is the first surname of the mother. Here we focus mainly on the first surname, that coming from the father, since the second one is "lost" in the second generation.

if there is enough variety of surnames and the surnames are not independently distributed across classes we can use them to estimate $p_{ij}$, and $r_i$.

Figure 1: An example of social class and surname inheritance: three generations and 12 PALs.

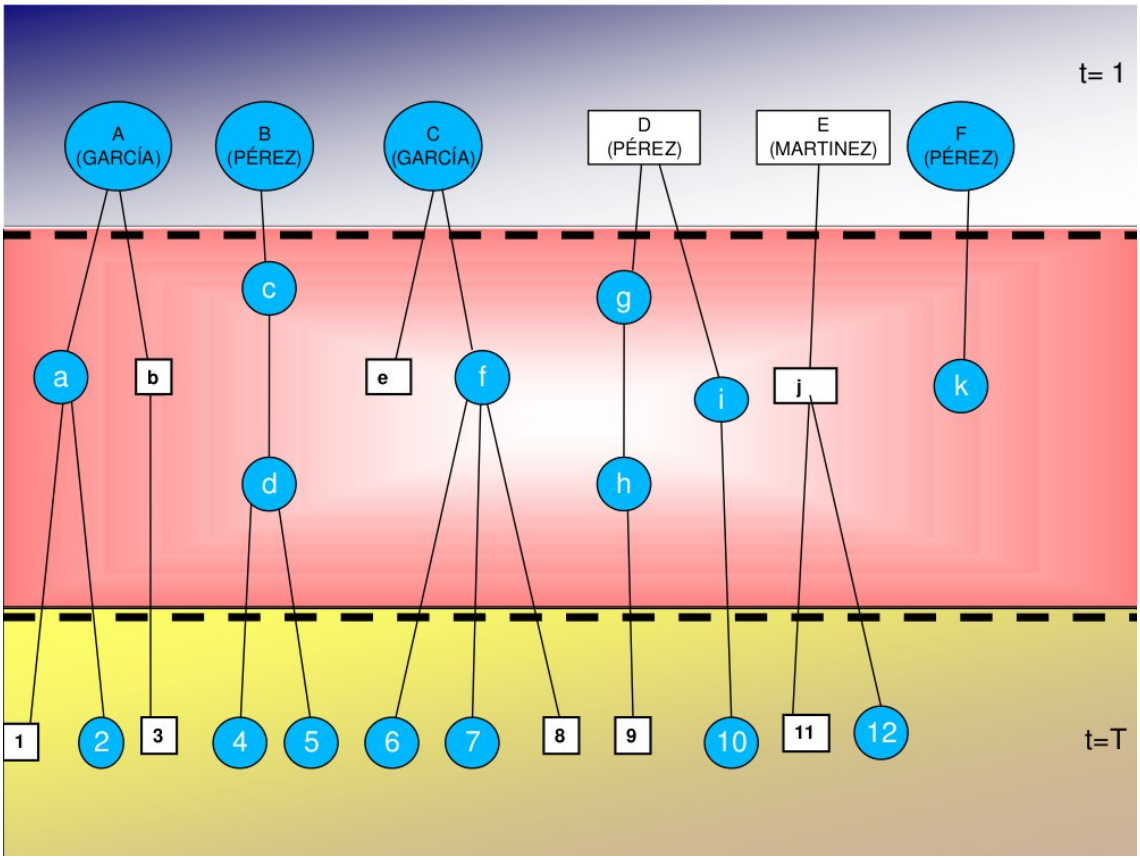



Figure 1 shows a simple example to illustrate the way we will use the "information" contained on surnames. In the first period ($t = 1$) the set $Y^1$ contains six persons $\{A, B, C, D, E, F\}$. We indicate in parenthesis the surname of each of these six persons. Each line represents a descendant. Notice that person $F$ does not have descendants in period $T$. In the last period, $T$, 12 persons form our set $Y^T$. The figure shows the PAL of all the individuals in $Y^T$. Thus, the PAL of, for example, person 1 is $\{a, A\}$, and the PAL of person 4 is $\{d, c, B\}$. Individuals with PAL that coalesce must have the same surname. For example, person 1 and person 2 both have the surname "García". However, the PAL of person 1 and the PAL of person 6

do not coalesce and they have the same surname "García". The figure also shows the class to which each person belongs. Individuals within a circle belong to the low class and those within an square belong to the high class.

If for all the individuals we knew their PAL and their social class, estimating the parameters of the model would be easy. However, we deal with situations of "incomplete" information as in the example of the figure where we do not observe the intermediate individuals between $Y^1$ and $Y^T$, i.e., we do not observe the individuals in the red area of the figure. How can we estimate the parameters $p_{ij}$, $r_i$ in this case? Knowing the surnames of individuals in $Y^1$ and $Y^T$ **delimits the set of possible ancestors of individuals in** $Y^T$. In Figure 1, person 1 can be a descendant of just two people, either $A$ or $C$. The case of person 11 is even better since he has to be a descendant of the unique person with the surname "Martínez" in $Y^1$. Thus, in our analysis of intergenerational social mobility we know that the ancestor of person 1 is for sure a low-type person, whereas the ancestor of person 11 is a high-type person. For person 4, however, his surname (Pérez) does not fully identify the class of his ancestor, who could be of any type.

At this point it might be useful to comment on two polar situations regarding the variety of surnames in the population. Suppose there is only one surname in the whole population $Y^1$. It is clear that in such a case surnames are of no help in our problem. The opposite would hold in a society where there are as many surnames as individuals in $Y^1$, in which case, we would know with certainty the ancestor of each individual in $Y^T$ just by observing his surname. The real world lies between those two polar situations. Thus, many societies present a few very common surnames, a large number of surnames borne by very few persons, and most of the population bearing surnames of intermediate frequency.

In addition to the requirement of enough variety of surnames there is a second necessary condition for surnames to be useful in our analysis. If for all surnames in period $t = 1$ the percentage of people of high type were the same, surnames would not include any useful information for the study of social mobility. Thus, the distribution of surnames and the

distribution of social classes among the population cannot be independent. Collado et al. (2008) show indeed that surnames and individuals among socioeconomic groups in Spain are not independently distributed. Moreover, they find a specific "bias" on the distribution of surnames. The more uncommon surnames appear in higher frequencies among groups of high socioeconomic status (see also Güell et al. 2007 for a similar result).

## 2.2 Estimation Method

We can write equations (1) and (2) as

$$n_H^T = \gamma_{HH}\ d_H + \gamma_{LH}\ d_L \tag{7}$$

$$n_L^T = \gamma_{HL}\ d_H + \gamma_{LL}\ d_L \tag{8}$$

where $\gamma_{ji} = p_{ji}\ r_j$ represent the (expected) number of descendants of type $i$ of an individual of type $j$. We will first estimate the parameters in equations (7) and (8). Then, we will recover the structural parameters, i.e., the conditional probabilities and the reproduction rates, by solving the equations $\gamma_{ji} = p_{ji}\ r_j$, for $i, j \in \{H, L\}$.

If, for any individual $k$ in $Y^1$, we could observe his type, the set of his descendants and their types, we could consistently estimate the parameters by estimating

$$n_{H,k}^T = \gamma_{HH}\ d_{H,k} + \gamma_{LH}\ d_{L,k} + \varepsilon_{H,k} \tag{9}$$

$$n_{L,k}^T = \gamma_{HL}\ d_{H,k} + \gamma_{LL}\ d_{L,k} + \varepsilon_{L,k} \tag{10}$$

by OLS, where $d_{H,k}$ ($d_{L,k}$) is a dummy variable that takes the value 1 if individual $k$ is of type $H$ (type $L$) and zero otherwise, $n_{H,k}^T$ ($n_{L,k}^T$) denote the observed number of type $H$ (type $L$) descendants of $k$, and $\varepsilon_{H,k}$ and $\varepsilon_{L,k}$ are the error terms. These two equations, however, cannot be directly estimated because we do not have information on the descendants of each particular individual. What we observe is the type and the surname of each individual in

$Y^1$ and $Y^T$. Notice that, as mentioned, the surnames delimit the set of potential ancestors of each individual in $Y^T$. Our identification strategy consists in aggregating the equations by surname.

Let $m^1_{i,s}$ ($m^T_{i,s}$) be the number of individuals in class $i$ with surname $s$ in $Y^1$ ($Y^T$). Then, from (9) and (10) we have

$$m^T_{H,s} = \gamma_{HH}\ m^1_{H,s} + \gamma_{LH}\ m^1_{L,s} + u_{H,s} \tag{11}$$

$$m^T_{L,s} = \gamma_{HH}\ m^1_{H,s} + \gamma_{LH}\ m^1_{L,s} + u_{H,s} \tag{12}$$

Since in our data we observe the number of individuals in class $i$ with surname $s$ in $Y^1$ and $Y^T$, by aggregating to the surname level we overcome the unobservability problem that equations (9) and (10) present.

The properties of the errors terms in equations (11) and (12) depend on the assumptions made on the original error terms in equations (9) and (10). If we assume that $\{(\varepsilon_{H,k}, \varepsilon_{L,k})\}$ are iid, i.e.,

$$\begin{pmatrix} \varepsilon_{H,k} \\ \varepsilon_{L,k} \end{pmatrix} \sim \text{iid} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_H & \sigma_{HL} \\ \sigma_{HL} & \sigma^2_L \end{pmatrix} \right]$$

then, the error term of the aggregated equations displays heteroskedasticity of a known form:

$$\begin{pmatrix} u_{H,s} \\ u_{L,s} \end{pmatrix} \sim \text{iid} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} m^1_s\sigma^2_H & m^1_s\sigma_{HL} \\ m^1_s\sigma_{HL} & m^1_s\sigma^2_L \end{pmatrix} \right]$$

where $m^1_s$ denotes the size of surname $s$, i.e., the number of people with surname $s$ in $Y^1$. Since the form of the heteroskedasticity is known, we may reach efficiency estimating equations (11) and (12) by generalized least squares (GLS) dividing both sides of the equation by the square root of the surname size in $Y_1$. Moreover, to account for the potential heteroskedasticity of the errors in equations (9) and (10), we calculate a robust-to-heteroskedasticity

variance matrix for our GLS estimator to compute the standard errors.

# 3 The Data

We apply the estimation methodology described in the previous section to the Spanish regions of Cantabria and Murcia. These regions are located in almost opposite sides of the country and are quite different in their climate, in their orographic and geophysical conditions, and in their socioeconomic and productive structures. Cantabria has a current population of about 589,000 and Murcia about 1,446,000. The GDP per capita in Cantabria is just slightly above the average of the whole country and in Murcia it is 82% of such average. In this empirical application, period $t = 1$ for the region of Cantabria corresponds to the year 1898 and period $T$ to 2001 (in some of the robustness exercises we also use data from the year 2004). We focus mostly on the region of Cantabria because, as we will explain later, this is the only region for which we have two census data sets (in electronic format) with a time separation of more than one century. In the case of Murcia we only have census data for the period $t = 1$ (year 1890), with the data used for period $T$ obtained from the telephone directory. Thus, our benchmark case will deal exclusively with the region of Cantabria. The estimations regarding the region Murcia will be used just as a robustness check.

Two notes of caution must be raised at this point. First, it is not clear that these regional samples are representative of the whole country.[7] Second, our theoretical model assumes no immigration flows and this can introduce a bias in the results. However, the region of Cantabria was a net exporter of migrants during the 20th century with very reduced immigration flows. Thus, 1980-90 was the only decade with a positive net migration flow, consisting of only 6,500 people, 1.2% of the original population (see Alcaide 2007). In any case, this recent immigration is discounted in the analysis since our 2001 data contain information on the birthplace of each individual and we discard all the individuals born in a

[7] We hope that the data of the 1898 census corresponding to the rest of the Spanish regions will be available soon in electronic format.

different region from Cantabria. In the case of Murcia, the region has been even a stronger net exporter of migrants[8] during the period. Despite all the precautions, we cannot rule out that such immigrants might have some influence on our result. In any case, we should bear in mind that our reproduction rates $r_i$ should be understood as the number of descendants living in Cantabria in 2001. Thus, the total number of descendants in the whole country might be different from such rate. In the same way, our estimates of intergenerational social mobility refer exclusively to the lineages remaining in the region.

## 3.1 Period 1: Data on Year 1898 and 1890

The data for the year 1898 (or 1890 for Murcia) come from the Spanish electoral census of that year. Such census lists the full name, age, address, and occupation of the person, and whether the person is illiterate or not, for the entire male population over 25 years old. This is a nationwide census but currently is only available in electronic format for the regions of Cantabria and Murcia.

The number of (male) people in Cantabria in the census is 59,000 and in Murcia almost 100,000. We select the individuals within the age range of 25-45 years.[9] These individuals form our set $Y^1$ which contains 31,908 people in Cantabria and 49,789 in Murcia.

We would like to classify the individuals in such sets according to their educational level. A natural classification would be to consider any person to be of high type if he is able to read and write. Unfortunately, such information in Cantabria's census contains too many errors, so we decided not to use it. Thus, we instead classify the people according to the socioeconomic status of their professions. We clustered all the 330 professions listed in the census into two groups: The high-class group ($H$) contains professions that can be seen as denoting a high socioeconomic status and covers 19.62% of the population in Cantabria

[8] The process changed in the 1980's when a significant number of the people who emigrated during the previous decades returned to the region.

[9] Our results are robust to small changes in the age range considered. This and the rest of robustness checks not provided in the paper are available from the authors upon request

(19.31% in Murcia), whereas the low-class group ($L$) contains all the other professions. It is important to note that this profession-based classification is probably very highly correlated with the classification we would obtain with the mentioned literacy criteria. In fact, we classified the population in the electoral census of Murcia first according to this profession-based criteria and second according to whether a person is able to read and write (contrary to the situation with Cantabria, this information in the census of Murcia is very reliable) obtaining that the two classifications are highly correlated (the percentage of people in class $H$ who are able to read and write is almost 70%, whereas the percentage of people in class $L$ is about 17%). Thus, we are confident that the groups used for the socioeconomic classification of the population of Cantabria in 1898 also classify people according to their human capital level.

## 3.2 Period T: Data on Year 2001

The main data set used here is the 2001 population census of Cantabria. In a second exercise we will use the data from the Yellow Pages of the telephone directory corresponding to Cantabria and Murcia as a way of checking the robustness of our results.

The 2001 population census of Cantabria contains information, among other variables, on the full name, age, occupation, and educational level of all individuals, both males and females. The set $Y^T$ is now obtained by selecting all the individuals, men and women, from the census in the specified age range of 22-46[10]. To test the robustness of our results to the sample selection, we also consider the case of only male and only female individuals. After excluding people born in other regions the total population in $Y^T$ is 150,832 (73,837 women).

We use two criteria to distinguish the $H$ and $L$ types. First, we classify people according to the type of education acquired. We include in the high class all the individuals with a bachelor's degree or a higher educational level. This class covers 17.44% of the population in

[10] We also analyzed what happens when the set $Y^T$ is given by the individuals between 46 and 70 years old. The results are consistent with the findings reported in this paper.

$Y^T$ (14.85% among men and 20.15% among women). Notice that the share of the population in the high class is relatively similar to the share of the population in the high class in $Y^1$. In the low class we include all the remaining individuals.

Second, we also consider classifying people in $Y^T$according to the socioeconomic status of their professions. The census provides a classification of professions in 18 groups and we classify those groups in two socioeconomic classes, high and low. Table A in the Appendix provides the classification of those 18 groups. Since the profession is only reported for those who are working, we have to drop inactive and unemployed individuals; thus, the population size is smaller than when we use the educational level. According to this socioeconomic classification of people in $Y^T$ we have 101,133 individuals (38,000 women) with 25.96% of them belonging to the high class (24.73% among men and 28.02% among women).

It can be argued that, to study issues of intergenerational social mobility, the use of educational level as the variable that generates our social classes is more correct than using the socioeconomic status of the professions. Thus, a problem with our data on professions is that for each individual the profession that appears in the census is the one the individual had at the time the census was performed. Many individuals change their profession throughout their lives and only considering the profession at a moment in time can bias the results significantly. This problem is similar to the one Solon (1992) emphasized in the case of intergenerational income mobility. The educational level, by contrast, seems to be safer from these changes over the adult life of an individual. In any case, we believe that, despite this problem, it is interesting to carry out all of our estimates, both with the type of profession and with the level of education.

A potential critique to defining social classes based on the level of education is the possible inconsistency with our definition of classes in the 19th century, which is based on the socioeconomic status of the professions. In fact, as we have argued, our social classes in the 19th century could be highly correlated with levels of education. However, we think that even if this is not the case the results are interesting because there is nothing inconsistent in

such approach. Thus, we could ask questions like, "How many of those with higher education today are descended from people with a high-status profession in the 19th century?" This is interesting even though the definitions of social classes in each period are different. Aware of the advantages and disadvantages of each approach and for the sake of completeness, we will carry out all of our analysis using both the profession and the level of education as separating criteria.

The data of the population census in the 21st century are available only for Cantabria, not for Murcia. This forced us to use an alternative data source that is available in electronic format for any Spanish region: the 2004 business section of the telephone directory[11] (Yellow Pages). We compiled information from the Yellow Pages for both Cantabria and Murcia. For Cantabria this business section contains 15,991 numbers registered under the names of persons[12] (32,177 in the case of Murcia) and provides information on the name and address of the subscriber and the type of business or professional activity. The number of different professions is about 1,000. We classify the professions in the Yellow Pages according to the level of education required to practice such professions.[13] The high group contains professions that require a bachelor's degree or a higher educational level. In the low class we include all the remaining professions. According to this classification of people, 31.10% in $Y^T$ belong to the high class (32.54% in the region of Murcia).

As the Yellow Pages provide no information on the age of people we cannot directly select a generation of people between the desired age interval of 22-46 years. We are confident that most people listed in the Yellow Pages belong to approximately the same generation and that the number of listed people whose father and/or mother is also listed is probably small. However, we are aware that it is only an approximation that may distort the results.[14]

[11] Notice that the population census refers to year 2001 and the telephone directory to 2004. We believe that this small date difference is of no consequence for our analysis.

[12] The telephone directory is available on a commercial CD-ROM (INFOBEL, http://www.infobel.com).

[13] The classification of the more than a thousand professions was done in a subjective manner by each of the three authors independently. The limited number of discrepancies was solved by consulting different information sources. The classification is available from the authors upon request.

[14] An advantage with respect to the census, though, is that the list of professions in the Yellow Pages is

Therefore, we consider that our main results are those obtained using the population census of Cantabria. The use of the Yellow Pages is motivated for two reasons: in the case of Cantabria, as a robustness analysis of the results obtained with the census, and in the case of Murcia, as an exercise of inter-regional comparison.

# 4 Main Empirical Results

We apply the methodology developed in section 2 to estimate the parameters $\gamma_{ji}$ using equations (11) and (12). Then, as mentioned above, we will recover the structural parameters, i.e., the conditional probabilities and the reproduction rates, by solving the equations $\gamma_{ji} = p_{ji}\ r_j$, for $i, j \in \{H, L\}$. All the results presented in this section, except for those in the last subsection, refer to the region of Cantabria. In all of them $Y^1$ is given by all male individuals in the 1989 electoral census aged 25-45 years, and the classes are based on the socioeconomic status of the professions. We first present the benchmark case for which i) $Y^T$ is given by all individuals age 22-46 years in the 2001 population census of Cantabria and (ii) the classes in 2001 are based on the educational level criterion.

The cases presented subsequently differ in the type of data used to generate $Y^T$ (only the male population or only the female population, or the telephone directory) or in the classification used to obtain the classes in $Y^T$ (socioeconomic status of the profession instead of educational level ). In the last subsection we present the results for the region of Murcia using the telephone directory to generate the population in $Y^T$.

## 4.1 Benchmark Case

Tables 1 and 2 show the results for the region of Cantabria when $Y^T$ is based on the 2001 population census and the corresponding two classes, $H$ and $L$, are generated by educational

very large (more than a thousand possible different professions) while the classification of occupations in the census contains only 19 different types.

level. This case contains the core results of the paper.

**Table 1**
*Parameters $\gamma_{ji}$*
*Source: Population census 2001.*
*Education groups.*

| Parameter | Estimate | SE |
|---|---|---|
| Equation 11 | | |
| $\gamma_{HH}$ | 1.058 | 0.072 |
| $\gamma_{LH}$ | 0.768 | 0.019 |
| Equation 12 | | |
| $\gamma_{HL}$ | 4.024 | 0.308 |
| $\gamma_{LL}$ | 3.873 | 0.089 |

The difference between the $\gamma_{ji}$ parameters clearly points in the direction of no HICC. A high-type person in the 19th century had on average 1.058 descendants belonging to the high-class group in 2001, and 4.024 descendants belonging to the low-class group, whereas for a low-type person in the 19th century those figures are 0.768 and 3.873.

Given the estimated values of the parameters $\gamma_{ji}$ we can easily compute the conditional probabilities $p_{HH}$ and $p_{LH}$, the reproduction rates $r_H$ and $r_L$, the odds ratio $OR$ and the outflow ratios $F_{HH}$, $F_{HL}$. Table 2 presents all these values, with the corresponding standard errors, with a last column showing the percentage of people of type $H$ in $Y^1$.

**Table 2**
*Mobility parameters and reproduction rates.*
*Source: Population census 2001. Education groups.*

| $p_{HH}$ | $p_{LH}$ | $r_H$ | $r_L$ | $OR$ | $F_{HH}$ | $F_{HL}$ | $(N_H^1/N^1) \times 100$ |
|---|---|---|---|---|---|---|---|
| 0.208 | 0.165 | 5.082 | 4.641 | 1.326 | 25.17 | 20.23 | 19.62 |
| (0.007) | (0.003) | (0.369) | (0.103) | (0.075) | | | |

These results show that the degree of social mobility has not been strong enough to erase the influence of the 19th-century ancestors on today's descendants. An odds-ratio of $OR = 1.326$ shows that the relative probability of having a descendant of type $H$ over having a descendant of type $L$ is around 32.6% higher for people of type $H$ than for people of type $L$. Note that the number of generations between people in $Y^1$ and people in $Y^T$,

i.e., the length of the PAL, must be three or four generations for the largest majority of the population (people in $Y^1$ are the great-grandfathers or great-great-grandfathers of people in$Y^T$). Thus, such bias on the relative probability of having descendants of type $H$ three or four generations forward seems quantitatively important.

A complementary approach to assess the degree of social mobility consists in comparing the outflow ratio $F_{HH}$ to the percentage $\frac{N_H^1}{N^1} \times 100$ of high-class ancestors. Our definition of HICC requires those two variables to take the same value. Our estimation shows, however, that 25.17% of $H$-type people in $Y^T$ have ancestors of type $H$ instead of the 19.62% required under HICC. In other words, after three/four generations, there is around a 28% "excess" of agents with ancestors of type $H$ among the current population in class $H$. It is interesting to note that the outflow ratio $F_{LH}$ is also bigger, although just slightly, than the proportion of high-class individuals in $Y^1$. This is due to the fact that the reproduction rate is higher for individuals of type $H$ than for individuals of type $L$, and therefore, the proportion of descendant of $H$-type individuals is larger than their share in $Y^1$.

This higher reproduction rate for $H$ type than for $L$ type (5.082 versus 4.641), although not statistically significant, appears in most of our estimates and is consistent with the findings of Clark (2007), who proves that reproduction rates in England have been higher for the high-class people than for the low-class people.

Thus, we conclude that our estimations show that the probability of belonging to the high-education group is still correlated with the socioeconomic status of the great-grandfathers and great-great-grandfathers. It is important to recall that we focus our analysis exclusively on the influence via paternal lines. One would suspect that the total influence of all ancestors (throughout the maternal and paternal lines) is still higher than the one we are able to detect here.

## 4.2 Gender Differences

Recall that the set $Y^T$ contains the entire population (male and female) within the age bracket 22-46 years. The set $Y^1$, however, contains no female population. It might be interesting to carry out the previous analysis, but first considering exclusively the male population in $Y^T$ and then the female population in $Y^T$. In a similar way to Tables 1 and 2 above, Tables 3 and 4 summarize the results for the two cases.

**Table 3**
*Parameters $\gamma_{ji}$ for men and women*
*Source: Population census 2001.*
*Education groups.*

| Parameter | Estimate | SE |
|---|---|---|
| Equation 11 | | |
| $\gamma_{HH}$, men | 0.479 | 0.035 |
| $\gamma_{HH}$, women | 0.579 | 0.041 |
| $\gamma_{LH}$, men | 0.329 | 0.010 |
| $\gamma_{LH}$, women | 0.439 | 0.011 |
| Equation 12 | | |
| $\gamma_{HL}$, men | 2.121 | 0.163 |
| $\gamma_{HL}$, women | 1.903 | 0.148 |
| $\gamma_{LL}$, men | 2.039 | 0.046 |
| $\gamma_{LL}$, women | 1.834 | 0.044 |

**Table 4**
*Mobility parameters and reproduction rates. Men and women.*
*Source: Population census 2001. Education groups.*

| | $p_{HH}$ | $p_{LH}$ | $r_H$ | $r_L$ | $OR$ | $F_{HH}$ | $F_{LH}$ | $(N_H^1/N^1)$ |
|---|---|---|---|---|---|---|---|---|
| Men | 0.184 | 0.139 | 2.600 | 2.367 | 1.40 | 26.23 | 20.25 | 19.62 |
| | (0.008) | (0.003) | (0.191) | (0.052) | (0.095) | | | |
| Women | 0.233 | 0.193 | 2.482 | 2.273 | 1.27 | 24.35 | 20.21 | 19.62 |
| | (0.009) | (0.003) | (0.181) | (0.052) | (0.083) | | | |

The general picture is the same for men and for women, and it coincides with that obtained in the aggregated case. However, there are some differences worth mentioning. The odds ratio is 1.40 for men and only 1.27 for women (even though the difference is not statistically significant). The proportion of men in group $H$ with ancestors of type $H$ is

26.23, whereas for women the proportion is 24.35. Thus, it seems that the influence of (male) ancestors is somewhat stronger for men than it is for women.

## 4.3 Telephone Directory

In this case the set $Y^T$ is constructed using the 2004 business section of the telephone directory in Cantabria. Tables 5 and 6 present the results.

**Table 5**
*Parameters $\gamma_{ji}$*
*Source: Telephone directory.*

| Parameter | Estimate | SE |
|---|---|---|
| Equation 11 | | |
| $\gamma_{HH}$ | 0.237 | 0.019 |
| $\gamma_{LH}$ | 0.136 | 0.005 |
| Equation 12 | | |
| $\gamma_{HL}$ | 0.407 | 0.031 |
| $\gamma_{LL}$ | 0.330 | 0.009 |

**Table 6**
*Mobility parameters and reproduction rates.*
*Source: Telephone directory.*

| $p_{HH}$ | $p_{LH}$ | $r_H$ | $r_L$ | $OR$ | $F_{HH}$ | $F_{HL}$ | $(N_H^1/N^1) \times 100$ |
|---|---|---|---|---|---|---|---|
| 0.368 | 0.292 | 0.643 | 0.467 | 1.41 | 29.77 | 23.11 | 19.62 |
| (0.016) | (0.007) | (0.046) | (0.012) | (0.126) | | | |

Note that the values of the $\gamma_{ji}$ parameters and the reproduction rates are lower than in the benchmark case due to a smaller population in the telephone directory than in the population census. However, the odds ratios are very similar (1.41 here and 1.326 in the benchmark case), thus confirming that the main results of the previous section are robust to the use of a different data source.

## 4.4 Socioeconomic Groups

We use the 2001 population census of Cantabria for $Y^T$, as in the benchmark case, but now the two classes, $H$ and $L$, are determined by the socioeconomic status of the professions. Tables 7 and 8 present the results.

**Table 7**
*Parameters $\gamma_{ji}$*
*Source: Population census 2001.*
*Socioeconomic groups.*

| Parameter | Estimate | SE |
|---|---|---|
| Equation 11 | | |
| $\gamma_{HH}$ | 1.023 | 0.071 |
| $\gamma_{LH}$ | 0.774 | 0.020 |
| Equation 12 | | |
| $\gamma_{HL}$ | 2.335 | 0.172 |
| $\gamma_{LL}$ | 2.349 | 0.051 |

**Table 8**
*Mobility parameters and reproduction rates.*
*Source: Population census 2001. Socioeconomic groups.*

| $p_{HH}$ | $p_{LH}$ | $r_H$ | $r_L$ | $OR$ | $F_{HH}$ | $F_{HL}$ | $(N_H^1/N^1) \times 100$ |
|---|---|---|---|---|---|---|---|
| 0.305 | 0.248 | 3.359 | 3.123 | 1.33 | 24.40 | 19.53 | 19.62 |
| (0.008) | (0.003) | (0.234) | (0.067) | (0.069) | | | |

It must be noted that the census only reports the profession of adult individuals if they are working; therefore, a large number of observations concerning profession are missing. Thus, the set $Y^T$ here contains less individuals (101,133) than in the benchmark case (150,832) which explains why the values of the $\gamma_{ji}$ parameters and the reproduction rates are lower here than in section 4.1 (see Appendix 1 for details on the classification of professions). When this is taken into account, the main conclusion obtained when people are classified according to their educational level also holds in this case. The probability of having a profession in group $H$ is higher for people whose great-grandfathers, or great-great-grandfathers, had themselves high-type professions.

Notice the similarity between the odds ratio here (1.33) and in the benchmark case (1.326). The outflow ratios $F_{HH}$ in both cases are also similar (24.40 and 25.17). Considering men and women separately yields a similar type of results to those in section 4.2. The odds ratio is a little bit higher for men (1.389) than for women (1.239), reinforcing the idea that the socioeconomic status of women depends less on their ancestors than does the socioeconomic status of men.

## 4.5 A Different Region

Here we focus on the region of Murcia. As explained above we construct $Y^1$ with data from the 1890 electoral census and $Y^T$using the 2004 business section of the telephone directory. Tables 9 and 10 give the results. It is important to note that Murcia and Cantabria are very different geographically and socially. Even so the two have relatively similar odds ratios (1.864 and 1.41). Murcia has been a more backward region with great emigration flows to other regions, which could explain why the odds ratio is somewhat higher. It is also interesting to note that the reproduction rate is higher for the low type, indicating perhaps a bias in emigration (the higher probability of the most skillful emigrating).

**Table 9**
*Parameters $\gamma_{ji}$*
*Source: Telephone directory (Murcia).*
*Education groups.*

| Parameter | Estimate | SE |
|---|---|---|
| Equation 11 | | |
| $\gamma_{HH}$ | 0.260 | 0.023 |
| $\gamma_{LH}$ | 0.198 | 0.007 |
| Equation 12 | | |
| $\gamma_{HL}$ | 0.326 | 0.039 |
| $\gamma_{LL}$ | 0.462 | 0.013 |

**Table 10**

*Mobility parameters and reproduction rates.*
*Source: Telephone directory, Region of Murcia. Education groups.*

| $p_{HH}$ | $p_{LH}$ | $r_H$ | $r_L$ | $OR$ | $F_{HH}$ | $F_{HL}$ | $(N_H^1/N^1) \times 100$ |
|---|---|---|---|---|---|---|---|
| 0.444 | 0.300 | 0.586 | 0.661 | 1.864 | 23.91 | 14.42 | 19.31 |
| (0.023) | (0.005) | (0.056) | (0.019) | (0.207) | | | |

# 5 Validating the main results

We conduct several additional extensions and modifications of the model to check the robustness of the main results obtained in the previous section. In all the cases, the results here are very similar to the ones reported in the benchmark case. In what follows, we present the most prominent results though for the sake of briefness not all the tables are shown.

## 5.1 Mobility among Bearers of Unique Surnames

Unique surnames in $Y^1$ are a very special case in our sample, because these are the only cases in which we have an exact identification of the ancestor. As explained in section 2.2 when this happens, we can consistently estimate the parameters by estimating equations (9) and (10). Obviously, there may be, and there is in fact, a sample selection bias issue here due to the already explained socioeconomic bias on the distribution of surnames. For instance, among people with unique surnames in 1898 about 35.8% of them are of $H$ type. Nevertheless, this exercise has the advantage that an exact identification of the ancestors is possible. Thus, we build the subsample of individuals in $Y^1$ bearing unique surnames (1,915 individuals) and their descendants in $Y^T$ (6,557 individuals), using the Cantabria data and taking the classes based on the education level. Tables 11 and 12 show the results. The odds ratio is now 1.339, basically the same we obtain in the benchmark case (1.326). Reproduction rates are now lower than in previous cases, but this might be due to a genetic bias. Many of these people have unique surnames because their ancestors had a low reproduction rate, and they may have inherited it. The men-only and women-only cases (not reported here) obtain very similar parameters in both cases. The odds ratio is again higher for men (1.407)

than for women (1.279), although it is not statistically significant.

**Table 11**
*Parameters $\gamma_{ji}$*
*Source: Population census 2001.*
*Unique surnames. Education groups.*

| Parameter | Estimate | SE |
|---|---|---|
| Equation 9 | | |
| $\gamma_{HH}$ | 0.774 | 0.087 |
| $\gamma_{LH}$ | 0.650 | 0.047 |
| Equation 10 | | |
| $\gamma_{HL}$ | 2.528 | 0.284 |
| $\gamma_{LL}$ | 2.842 | 0.199 |

**Table 12**
*Mobility parameters and reproduction rates.*
*Source: Population census 2001. Unique surnames. Education groups.*

| $p_{HH}$ | $p_{LH}$ | $r_H$ | $r_L$ | $OR$ | $F_{HH}$ | $F_{HL}$ | $(N_H^1/N^1) \times 100$ |
|---|---|---|---|---|---|---|---|
| 0.234 | 0.186 | 3.302 | 3.492 | 1.339 | 39.92 | 33.17 | 35.82 |
| (0.012) | (0.008) | (0.358) | (0.235) | (0.114) | | | |

## 5.2 Additional robustness checks

People in Spain bear two surnames, the first being the father's first surname and the second being the mother's first surname. In all our previous results, only first surnames were used. This is consistent with the idea that we follow the paternal ancestry lines, but leaves open the question of whether intergenerational transmission of social status may also act through the maternal channel. For this reason, we repeat all the estimations of the previous sections using the second surname instead, i.e., individuals in $Y^T$ are identified by their second surnames. Notice that now the ancestry linage is not the paternal but the mother-grandfather-great-grandfather, etc., lineage. The results obtained are very similar to those reported in section 4.1. The odds ratio now is 1.298 (s.e. 0.072), a bit smaller than in the benchmark case (1.326) but with the difference not statistically significant. Even more, if we now use the women-only subsample in $Y^T$ our estimates of the odds ratio using either the first surname or the second surname almost coincide (1.270 and 1.272 and standard errors of 0.083 and

0.080, respectively). Thus, the results using the second surname are very consistent with those using the first surname, although they suggest a possible difference between maternal and paternal influence on social mobility.

As additional checks of the robustness of our results we carry out the following exercises: i) we modify the age bracket to 30-54 years in the definition of $Y^1$and $Y^T$; ii) we drop the five most common surnames; iii) we drop the most unusual surnames; and iv) we use some slightly different classification of professions in the 1898 electoral census and in the telephone directory. The results obtained in all these cases are consistent with the main findings reported in the paper and are available upon request.

# 6 Comparisons with previous empirical findings

One of the most prominent features of our approach is that it is a low-cost method because it uses census data that can be readily obtained for many countries. The "price" to pay is the need to make some assumptions; thus, we would like to know if these assumptions are acceptable. One way of knowing this is to compare our findings with the estimates that other authors have found using different methods and data. However, making these comparisons is not easy because, as already noted, there are very few contributions that try to measure social mobility during long periods of time.

As already mentioned, an exception is the work of Lindahl *et al.* (2012), who estimate the correlation between educational level of individuals and their great-grandparents. Their educational group classification is different from the one we use here, so it is not easy to compare their results with ours. In any case, we take their transition probabilities (Table 4c, pp 17) and aggregate the four categories in two groups. In particular, we aggregate the two lower and the two higher educational groups for the great-grandparents generation and the three lower educational groups for the great-grandchildren generation. With this aggregation their high education group containing 4.5% of the individuals in the first generation and 25%

of the individuals in the last generation and the odds-ratio is 2.15. In our case, considering a high educational group classification different from that used until now and which contains 6.3% of the population in $Y^{115}$ and 17.4% in $Y^T$, our odds-ratio is 1.61. These two odds-ratios have the same order of magnitude, and since Lindahl *et al.* use a much smaller sample size, and in our case many of the observations correspond to individuals and their great-great-grandparents, we can not rule out that the two corresponding levels of educational mobility in this period were similar.

We can also try to compare our findings with those of one-generation studies of ISM. Since we only have data on generations 1 and $T$ we cannot provide an estimate of one-generation social mobility unless we make some additional assumptions. We assume that the length of PAL is three generations, which approximately covers the 103-year span between 2001 and 1898.[16] Thus, the individuals in $Y^1$ are assumed to be the great-grandfathers of individuals in $Y^T$. We need further to assume constant reproduction rates and constant conditional probabilities of social mobility along the different generations. Furthermore, to obtain estimations on the one-generation reproduction and mobility parameters consistent with our previous findings we have to restrict the set $Y^T$ to the male population.

Under these assumptions, estimates of the one-generation parameters may be obtained by viewing the dynamic process as a Markov chain.[17] Denote by $\tilde{r}_i$ the one-period reproduction rate of individuals of type $i$, i.e., the expected number of (male) children of a person of type $i$ who reach adult life. Let $\tilde{p}_{ij}$ be the probability that a son of a person of type $i$ is of type $j$. Thus, given the previously estimated values of $r_i$ and $p_{ij}$ the one-generation parameters are the solution to

[15] This classification refers to the benchmark case of Santander, and is not provided here, but is available upon request.

[16] Taking three generations is based on a rough calculus. Assuming that each generation takes 30 years and that most of the children are born when their fathers are 25-30 year old, for most people in $Y^T$ of age 22 the PAL has at least length four. For individuals age 46 we could assume that the average length is three. Since the median age in $Y^T$ is 34 years it is difficult to establish with certainty if the average length of all the PALs is closer to three or to four, but we find it conservative to take it as three. Assuming four generations yields lower levels of one-generation social mobility.

[17] Lindahl et al. (2012) call into question this assumption. See also Sauder (2006) and Maurin (2002).

$$\begin{pmatrix} \tilde{p}_{HH}\tilde{r}_H & \tilde{p}_{LH}\tilde{r}_L \\ (1-\tilde{p}_{HH})\tilde{r}_H & (1-\tilde{p}_{LH})\tilde{r}_L \end{pmatrix}^G = \begin{pmatrix} p_{HH}r_H & p_{LH}r_L \\ (1-p_{HH})r_H & (1-p_{LH})r_L \end{pmatrix} \tag{13}$$

where $G$ denotes the number of generations, three in our case. The one generation odds-ratio can be now computed as

$$\widetilde{OR} = \frac{\frac{\tilde{p}_{HH}}{\tilde{p}_{HL}}}{\frac{\tilde{p}_{LH}}{\tilde{p}_{LL}}}$$

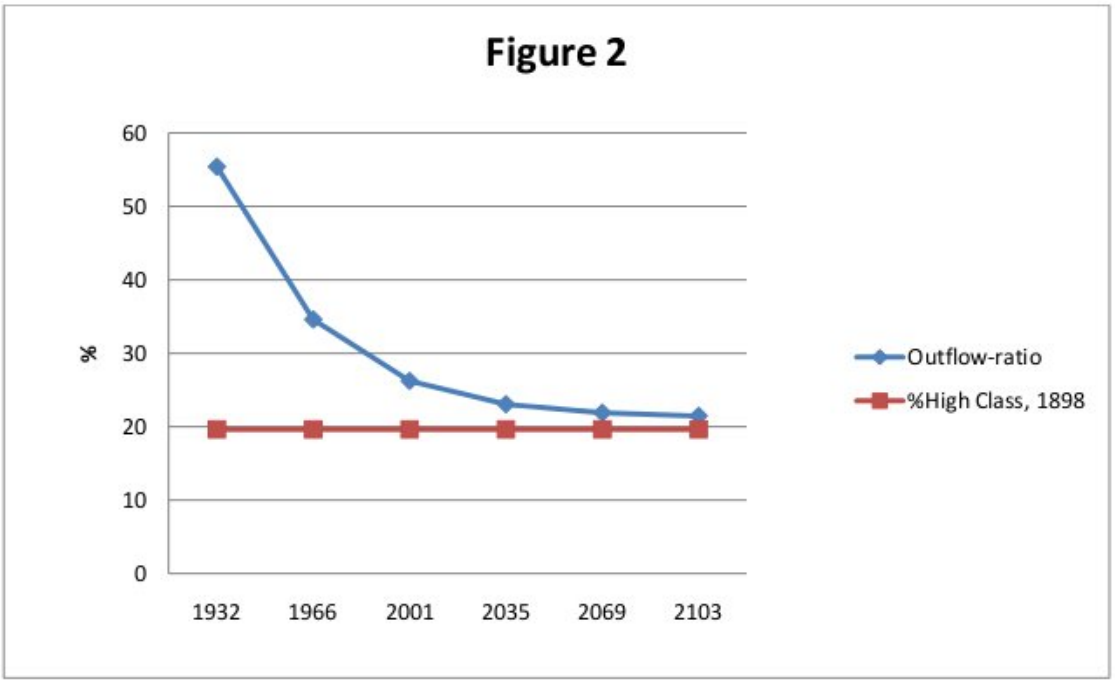
Using the parameters reported in Table 4 we solve equation (13) and obtain $\tilde{p}_{HH} = 0.441$, $\tilde{p}_{LH} = 0.092$, $\tilde{r}_H = 1.413$ and $\tilde{r}_L = 1.326$. The one-generation odds ratio is $\widetilde{OR} = 7.76$ (with standard error of 1.095). This ratio has a magnitude within the range of other studies on one-generation social mobility. In particular, Klakbrenner and Villanueva (2006) find a transition matrix from no-college to college education among Spanish father-son pairs in 1990 with a odds-ratio[18] of 10.5. These odds ratios are between the estimations for United States (6) and for Italy (25) reported in Checchi et al. (1999) in similar matrices of educational intergenerational mobility. Thus, our results suggest that, contrary to the assertions in Lindahl *et al.* (2012) and Clark *et al.* (2012), the standard estimates of mobility from one generation do not overestimate the true mobility over more generations.

We can also compute the theoretical class composition that should have prevailed from those reproduction rates and transition probabilities for each of the three generations between our first year and the final year $T$ (between 1898 and 2001). Let's identify those three generations with years 1932, 1966, and 2001. Thus, we can calculate the total male population and the shares of $H$ and $L$-type people for each of those three years. Similar to what we did in our previous sections we can compute the outflow ratio $F_{ij}^g$ for each generation $g \in \{1932, 1966, 2001\}$, i.e., the proportion of individuals in generation $g$ in class $j$

[18] If we assume that the length of PAL is four generations instead of three generations, the one-generation odds ratio would be $\widetilde{OR} = 11.43$.

with ancestor[19] of type $i$. Furthermore, by assuming that the parameters $\tilde{p}_{HH}, \tilde{p}_{LH}$, $\tilde{r}_H$, and $\tilde{r}_L$ remain always constant we can also estimate such outflow ratios for future generations. Figure 2 provides the outflow ratios $F^g_{HH}$ for our three generations (1932, 1966, 2001) and for the next three future ones (2035, 2069, 2103). The figure shows, for each generation, the proportion of high-type people with ascendant of high type. Recall that HICC requires the outflow ratio to be equal to the proportion of high-type individuals in $Y^1$, which is 19.62%. Thus, our analysis suggests that i) we have not reached a HICC yet, ii) but the individual composition of social classes beyond the year 2035 is basically independent of the class composition in 1898 so that four/five generations are enough to erase the traces of the past.

Figure 2: Proportion of High status descendants for each generation




As we mentioned above, one of the assumptions behind the one-generation results presented in this section is that the conditional probabilities have remained constant along the 20th century. If this assumption were true, the correlation in the socioeconomic status of

[19] The ancestors are individuals in 1898, i.e. in $Y^1$

brothers today should coincide with the correlation in the socioeconomic status of brothers at the end of the 19th century. Unfortunately, we do not precisely observe which individuals are brothers in our sets $Y^1$ and $Y^T$. However, we can approximately identify a subset of brothers by using individuals with very rare surnames. In particular we consider as very rare those surnames borne by just two people, 519 surnames in $Y^1$ and 825 in $Y^T$, and we claimed that this pairs of individuals are mainly brothers or at most cousins. We compute the correlation in the socioeconomic status of these brothers, obtaining a value of 0.67 for the 19th century and 0.63 in the 20th century.[20] These two numbers are very similar and this provides some evidence in favour of the assumption that the conditional probabilities have remained constant along the 20th century.

## 6.1 Estimating intergenerational elasticities

So far we have focus on the estimation of mobility matrices, and therefore it is difficult to compare our results with the standard approach in economics, which focus on estimating income and wealth elasticities. For this reason in this section we attempt to estimate the intergenerational elasticity of socioeconomic status for our benchmark case. Suppose that we want to estimate the intergenerational elasticity of income or education for our individuals in $Y^1$ and $Y^T$. Consider the standard linear intergenerational equation

$$y_i^T = \alpha + \beta y_i^1 + \varepsilon_i \tag{14}$$

where $y_i^1$ denotes the permanent income (in logs) of an individual $i$ from the 1898 generation ($Y^1$) and $y_i^T$ the income of his descendant in the current generation ($Y^T$), and $\varepsilon_i$ is a disturbance term. We would like to estimate the parameter $\beta$. In principle this task can not be performed satisfactorily in our case because: i) we do not have information on

[20] Notice that the correlation is very high since the variable socioeconomic status is a dummy. Bjorklund et al 2009 find lower correlations in income in Sweden, from 0.34 for the cohorts born in the early 1930s to 0.23 for the cohorts born around 1950.

the descendants of each particular individual and, ii) we only have categorical data for the variables $y_i^T$ and $y_i^1$ . However, we can overcome these two problems by using the technique proposed by Clark *et al.* (2012). The method is based on a double aggregation and the use of rare surnames. We first select all the surnames with less than 40 bearers.[21] Next we compute for each (rare) surname the average income of the individuals bearing such surname, and second we aggregate surnames in two groups according to such average incomes. Thus, consider first the following modified version of (14)

$$y_s^T = \alpha^* + \beta^* \; y_s^1 + \varepsilon_s \tag{15}$$

where $y_s^1$ ( $y_s^T$ ) is now the average log income across individuals with surname $s$ in $Y^1$ ($Y^T$). This aggregation by surnames is similar to the aggregation we do in section 2. Clark et al (2012) explains under which conditions $\beta^*$ is the same as $\beta$ in (14), in particular we have to assume that the expected reproduction rates are the same for all individuals.[22] The next step requires aggregating the surnames into two groups according to their average income. Consider the individuals and surnames in the first generation. We can order the surnames by their average (log) income, i.e. we order the surnames by their values $y_s^1$ . The first group contains the richest surnames, and we denote such set of surnames by $R$. The second group contains the rest of surnames and we denote it by $P$. We define $y_R^1$ and $y_P^1$ as the weighted average of the $y_s^1$ in the first and second group, namely

$$\begin{aligned} y_R^1 &= \frac{1}{\sum_{s\in R} m_s^1} \sum_{s\in R} m_s^1 y_s^1 \\ y_P^1 &= \frac{1}{\sum_{s\in P} m_s^1} \sum_{s\in P} m_s^1 y_s^1 \end{aligned} \tag{16}$$

[21] We select the rare surnames because we need variability in their average income. Most surnames beared by a large number of people have a similar average income. Our results do not depend on the precise definition of "rare" surname.

[22] In our previous results we found different reproduction rates for the two social classes. However, in most of the cases the differences were not statistically significant. Thus, we believe that assuming here constant reproduction rates does not invalidate our main qualitative results on income elasticities.

where $m_s^1$ is the number of individuals in generation 1 with surname $s$. We choose these groups of surnames, $R$ and $P$, such that each of them contains approximately half of the total individuals in the first generation. Given these two sets of surnames, we next compute the corresponding weighted average of the (log) income in the last generation, and we denote them by $y_R^T$ and $y_P^T$. The parameter $\beta^*$ in (15) can be estimated simply as the solution to

$$y_R^T - y_P^T = \widehat{\beta}^*(y_R^1 - y_P^1) \tag{17}$$

As explained in Clark *et al* (2012), assuming that the reproduction rates do no depend on income, the expected value of $\widehat{\beta}^*$ will be the intergenerational elasticity $\beta$ in (14).

Returning to our benchmark case, in order to use equation (17) we still need to find for each rare surname $s$ its average (log) income $y_s^1$. We propose the following way to compute such average incomes. Remember that the individuals in $Y^1$ were classified in a high-class group ($H$) and low-class group ($L$). However, in this section we will use a classification of individuals according to their occupations into three groups,[23] the high-class ($H'$), the middle-class ($M$) and the low-class ($L'$). The population shares of each of these three groups are, 6.26%, 13.36% and 80.38% . The reason we use here a finer classification is to obtain more accurate individual income estimates. We assume that the individuals in the high-class group $H'$ are the ones in the top 6.26% of the income distribution, the individuals in the middle-class group $M$ are the ones in the interval (6.26%, 80.38%), and individuals in the low-class group $L'$ are the ones in the bottom 80.38% of the income distribution. For the current generation ($Y^T$) we also classify individuals in three groups with population shares of 3.45%, 11.40% and 85.15%. However, we do not know the particular income of any individual in a group. For this reason, we next assume that all individuals in a group have the same expected income. Thus, we identify an individual's income with the average income in his group, $H', M$ or $L'$. To do this we need to know, for the first generation, the average

[23] The classification with the list of occcupations in each group is available from the authors upon request.

income among the 6.26% richest people in $Y^1$, the average income in the interval (6.26%, 80.38%) and the average income among the poorest 80.38% of the population. Assuming that income follows a Lognormal distribution, and using the parameters from Prados de la Escosura (2008) and Carreras *et al.* (2005), we compute that the average income (in 1990 dollars) among the high-class group[24] $H$ in 1898 is $I_H^1 = 4,435$, for the middle-class group the average income is $I_M^1 = 2,694$ and for the low-class group $L$ is $y_L^1 = 1,226$. The average income for each of the three corresponding groups in $Y^T$ (year 2001), are $I_H^T = 48,489$ , $I_M^T = 28,531$ and $I_L^T = 11,749$. Given this information about the expected income of each individual is easy to calculate for each rare surname $s$ the average of the log income in the first generation, $y_s^1$, and in the second generation, $y_s^T$

Now we can easily calculate $y_R^1$ and $y_P^1$ using the corresponding formula in (16), and we obtain $y_R^1 = 7.4405$ and $y_P^1 = 7.1440$. In a similar way we can obtain the corresponding values for the current generation, $y_R^T = 9.5460$ and $y_P^T = 9.5272$. Finally, from equation (17) we obtain the intergenerational elasticity $\widehat{\beta} = 0.0633$. Notice that this elasticity $\widehat{\beta}$ refers to the association between incomes of individuals in the current generation ($Y^T$) and incomes of their ancestors in year 1898 ($Y^1$). If we assume, as we did above for the case of transition matrices, that individuals in $Y^1$ are the great-grandfathers of individuals in $Y^T$, and we further assume constant elasticities along the different generations, the associated one-generation elasticity would be

$$\beta_1 = \widehat{\beta}^{1/3} = 0.3985$$

If instead we assumet that individuals in $Y^1$ are the great-great-grandfathers of individuals

[24] According to Prados de la Escosura (2008) the Gini coefficients for the income distribution in Spain in years 1898 and 2000 were 0.32, and 0.33 respectively. The mean income for those two year were 1623 dollars and 14928 dollars. If we know the Gini coefficient of a lognormal distribution, $G$, we can determine its variance $\sigma$, by solving $G = 2\Phi(\frac{\sigma}{\sqrt{2}}) - 1$. Our results are robust to considering a Gini coefficient for year 1898 as high as 0.5

in $Y^T$,the associated one-generation elasticity would be

$$\beta_1 = \widehat{\beta}^{1/4} = 0.5016$$

Thus, the one-generation income elasticity should be in the interval $[0.3985, 0.5016]$. This intergenerational income elasticity is within the range of the one-generation standard estimates for many countries (see Blanden 2011) and the income elasticity estimates for Spain by several authors, 0.4 in Cervini-Plá (2011) and 0.4-0.66 in Sanchez-Hugalde (2004). Thus, we argue that, contrary to Lindahl *et al.* (2012) and Clark *et al.* (2012), the elasticity estimated using data on several generations do not suggest that the standard one-generation approach underestimates the long-run intergenerational social mobility.

This estimate of the income elasticity is based on some strong assumptions, and that's why we should take the results with certain skepticism. Still, we believe it is a good approximation which is consistent with our previous results obtained with transition matrices. Moreover, we have cheeked that this double-aggregation methodology works quite well when applying it to the data of the 2001 population census of Cantabria. Such census, as explained above, contains information, among other variables, on the full name, age, occupation, and educational level of all individuals, both males and females. It also contains information about the family relationships in the household. Thus, we can identify parents and sons living together in the same house. We can then estimate the intergenerational educational elasticity in the standard way, i.e. as in (14), and also using the double aggregation methodology proposed by Clark et al (2012). Both methods give almost identical results,[25] which makes us more confident about the validity of our estimate obtained using the double-aggregation method.

[25]The results and details of both procedures are availbale from the authors upon request. The elasticies obtained here are different from the ones reported by other authors and the the one we estimate in this paper. This is due to a well-known sample selection problem related to the offspring's and parens' income when they live together.

# 7 Conclusions

We have developed a novel methodology that makes it possible to study long-run intergenerational mobility using census data from different years. We link individuals in the different census data sets by using their surnames. A necessary condition for our methodology to work is that surnames must convey socioeconomic information, i.e., there must be some bias in the distribution of surnames among the different socioeconomic groups. The existence of such bias has been established for the Spanish case by Collado et al. (2008) and Güell et al. (2007).

We have applied our methodology to study intergenerational mobility in two Spanish regions during the 20th century. Our econometric analysis suggests that for a male born in the middle of the 19th century, the probability that any of his adult descendants (in the patrilineal line) at the end of the 20th century would have a high status, compared with the probability of having a low status, is 32.6% higher if he has a high status himself than if he has a low status. Thus, we still detect a significant imprint of the past. We argue, however, that if we assume stability of the mobility coefficients, the link between socioeconomic classes basically disappears after five generations.

Our results are consistent with those found by other authors examining short-run socioeconomic mobility in the second half of the 20th century. We also found that the socioeconomic link with ancestors is somewhat weaker for women than for men. It is important to stress that our methodology only analyzes the paternal line. These results could be different in the more general case considering both paternal and maternal influence, and in this case is reasonable to expect a higher influence of the past. At the same time incorporating the paternal and the maternal lines in the analysis of long run social mobility makes it difficult to talk about the social class of the ancestors. After a few generations, backwards in time, the set of ancestors of any given person becomes large and probably many of them belong to different social classes. In our case, with three generations, the number of ancestors of

any individual can be 16, a number already large enough to expect that all of them belong to the same social class. In the case of even more generations the problem becomes much more complicated by the large number of ancestors that are shared.[26] Thus, the analysis of long-run social mobility and social classes might not be well defined.

Finally, our methodology also permits estimation of the reproduction rate of different social classes. We have shown that there is a reproductive advantage for individuals in higher socioeconomic groups, although it is not always statistically significant. This result is in concordance with the thesis defended by Clark (2009) for the case of England.

# References


[1] Alcaide Inchausti, Julio, 2007. *Evolución de la población española en el siglo XX por provincias y comunidades autónomas*, vol.2, Fundación BBVA: Bilbao.

[2] Björklund, A. and M. Jäntti, 2009, "Intergenerational mobility and the role of family background." In W. Salverda, B. Nolan and T. Smeeding (eds), *Oxford Handbook of Economic Inequality*, pp 491-521, Oxford University Press: Oxford, UK.

[3] Björklund, A., Jäntti, M. and M. Lindquist, 2009, "Family Background and Income during the Rise of the Welfare State: Brother Correlations in Income for Swedish Men Born 1932-1968." *Journal of Public Economics* 93(5-6), 671-680.

[4] Black, Sandra E. and Paul J. Devereux, 2010. "Recent developments in intergenerational mobility." IZA DP No. 4866.

[5] Blanden, Jo, 2011. "Cross-country rankings in intergenerational mobility: A comparison of approaches from Economics and Sociology," *Journal of Economic Surveys*.doi:10.1111/j.1467-6419.2011.00690.x


[26] See Derrida et. al (2002) for an analysis of the statistical properties of genealogical trees.

[6] Carreras, Albert, Prados de la Escosura, Leandro and Joan Roses, 2005, "Renta y riqueza", chapter 17, pp 1297-1376, in A. Carreras and X. Tafunell (eds) *Estadísticas históricas de España: siglos XIX-XX, Volumen1*, Bilbao, Fundación BBVA.

[7] Cervini-Plá, María, 2011, "Intergenerational earnings and income mobility in Spain", MPRA paper No 34942, posted 22. November 2011.

[8] Checchi, Daniel A., Ichino, Andrea and Aldo Rustichini, 1999. "More equal but less mobile?: Education financing and intergenerational mobility in Italy and in the US", *Journal of Public Economics*, 74 (3), 351-393.

[9] Clark, Gregory, 2007. *A Farewell to Alms: A Brief Economic History of the World*. Princeton University Pres: Princeton.

[10] Clark, Gregory, 2009. "The indicted and the wealthy: Surnames, reproductive success, genetic selection and social class in pre-industrial England", mimeo.

[11] Clark, Gregory, 2010, "Regression to mediocrity? Surnames and social mobility in England, 1200-2009," mimeo.

[12] Clark, Gregory and Neil Cummins, 2012, "What is the true rate of social mobility? Surnames and social mobility in England, 1800-2012", mimeo.

[13] Clark, Gregory, 2012 a, "What is the true rate of social mobility in Sweden? A surname snalysis, 1700-2012", mimeo.

[14] Clark, Gregory, 2012 b, "Social mobility rates in the USA, 1920-2010: A surname analysis", mimeo.

[15] Collado-Vindel, M. Dolores, Ortuño-Ortín, Ignacio and Andrés Romeu, 2008. "Surnames and social status in Spain", *Investigaciones Economicas*, 32(3): 259-287.

[16] Corak, Miles, 2006. "Do poor children become poor adults? Lessons from a cross country comparison of generational earnings mobility." IZA DP No. 1993.

[17] Derrida, Bernard, Manrubia, Susana and Damian H. Zanette, 2002. "Statistical properties of genealogical trees", *Physical Review Letters*, 82:99, 1987-1990.

[18] Erikson Robert, and Jhon H. Goldthorpe, 1992. *The constant flux: A study of class mobility in industrial societies*, Clarendon Press: Oxford.

[19] Erikson Robert, and Jhon H. Goldthorpe, 2002. "Intergenerational inequality: A sociological perspective", *Journal of Economic Perspectives*, 16(3): 31-44.

[20] Ferrie, Joseph P., 2005. "History lessons: The end of American exceptionalism? Mobility in the United States since 1850", *Journal of Economic Perspectives*, 19(3): 199-215.

[21] Ferrie, Joseph P. and Jason Long, 2009, "Intergenerational occupational mobility in Britain and the U.S. Since 1850", forthcoming, *American Economic Review*.

[22] Güell, Maia, Rodrigue Mora, Jose Vicente and Chris Telmer, 2007. "Intergenerational mobility and the informative content of surnames", CEPR Discussion paper No 6316.

[23] Hertz, Tom, Jayasundera, Tamara, Piraino, Patricio, Selcuk, Sibel, Smith, Nicole and Alina Verashchagina, 2007. "The inheritance of educational inequality: International comparisons and fifty-year trends", *The B.E. Journal of Economic Analysis & Policy*, 7(2): article 10.

[24] Kalbrenner, Esther and Ernesto Villanueva, 2006. "Intergenerational mobility in income and education in Spain", mimeo.

[25] Lindahl, H., Palme , M., Massih, S. and A. Sjogren, 2012, "The intergenerational persistence of human capital: An empirical analysis of four generations", *IZA DP* No. 6463.

[26] Maurin, E. (2002), "The impact of parental income on early schooling transitions: A re-examination using data over three generations", *Journal of Public Economics* **85(3)**, 301-332.

[27] Mood, C., J.O. Jonsson, and E. Bihagen, 2012. "Socioeconomic persistence across generations: Cognitive and noncognitive processes", Chapter 3, pp. 53-83, in J. Ermisch, M. Jäntti, and T.M. Smeeding (eds.), *From Parents to Children. The Intergenerational Transmission of Advantage*. New York: Russell Sage.

[28] Osborne Groves, Melissa, 2005. "Personality and the intergenerational transmission of economic status", Chapter 7, pp 208-231, in S. Bowles, H. Gintis and M. Osborne Groves (eds.), *Unequal chances. Family background and economic success*. New York: Russell Sage.

[29] Piketty, Thomas, 2000. "Theories of persisitent inequality and intergenerational mobility", *Handbook of Income Distribution*, A.B. Atkinson and F. Bourguignon (eds). Elsevier Science.

[30] Prados de la Escosura, 2008. "Inequality, poverty and the Kuznets curve in Spain, 1850\2000", *European Review of Economic History*, v. 12, n. 3 , pp. 287 - 324.

[31] Sánchez Hugalde, Adriana, 2004, "Movilidad intergeneracional de ingresos y educativa en España (1980-90)", Institut d'Economia de Barcelona, w.p. 2004/1

[32] Sauder, U., 2006. "Education transmission across three generations. New evidence from NCDS data", mimeo, Univeristy of Warwick.

[33] Solon, Gary, 1992. "Intergenerational income mobility in the United States", *American Economic Review*, 82:393-408.

[34] Solon, Gary, 1999, "Intergenerational mobility in the labor market". In: Ashenfelter, O., Card, D. (Eds.), *Handbook of Labor Economics*, Vol 3A. Elsevier, North-Holland: Amsterdam, pp. 1761-1800.

[35] Solon, Gray, 2002. "Cross-country differences in intergenerational earnings mobility," *Journal of Economic Perspectives*, **16**: 59-66.

[36] Webber, Richard, 2004. "Neighbourhood segregation and social mobility among the descendants of Middlesbrough's 19th century Celtic immigrants", wp. 88, CASA, UCL.

**Departamento de Fundamentos del Análisis Económico**
Facultad de Economía y Empresa
E-30100 Campus de Espinardo
Murcia (SPAIN)
Tel.: (34) 868 883 784
E-mail: nsb@um.es
Website: www.um.es/analisiseco