

# *Intergenerational social mobility in Spain since the late 19th Century*

M. Dolores Collado\*    Ignacio Ortuño-Ortín\*\*    Andrés Romeu\*\*\*

\*Universidad de Alicante

\*\*Universidad Carlos III de Madrid

\*\*\*Universidad de Murcia

March 28, 2008

# Introduction

- The degree of intergenerational socioeconomic mobility 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 education level and parental wealth on offspring's education levels and wealth. Other possible reasons are genetic inheritance and group effects.
- There is an important literature trying to measure the level of intergenerational mobility.

Economists → Income or wealth (Corak 2006, Piketty 2000).

Sociologists → Education levels and occupations (Erikson and Goldthorpe 2002, Hertz et al 2008).

# Introduction

- The degree of intergenerational socioeconomic mobility 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 education level and parental wealth on offspring's education levels and wealth. Other possible reasons are genetic inheritance and group effects.
- There is an important literature trying to measure the level of intergenerational mobility.

Economists → Income or wealth (Corak 2006, Piketty 2000).

Sociologists → Education levels and occupations (Erikson and Goldthorpe 2002, Hertz et al 2008).

# Introduction

- The degree of intergenerational socioeconomic mobility 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 education level and parental wealth on offspring's education levels and wealth. Other possible reasons are genetic inheritance and group effects.
- There is an important literature trying to measure the level of intergenerational mobility.

Economists → Income or wealth (Corak 2006, Piketty 2000).

Sociologists → Education levels and occupations (Erikson and Goldthorpe 2002, Hertz et al 2008).

- Economists → intergenerational income elasticity.  
Sociologists → transition matrices.
- Most of the estimates of the level of intergenerational socioeconomic mobility use “one generation data”.
- Empirical evidence for the second half of the 20th century → Good estimates of the one generation mobility in income for several developed countries.
- Empirical evidence for the first half of the 20th century → Very scarce (Ferrie (2005) and Ferrie and Long (2005)).

- Economists → intergenerational income elasticity.  
Sociologists → transition matrices.
- Most of the estimates of the level of intergenerational socioeconomic mobility use “one generation data”.
- Empirical evidence for the second half of the 20th century → Good estimates of the one generation mobility in income for several developed countries.
- Empirical evidence for the first half of the 20th century → Very scarce (Ferrie (2005) and Ferrie and Long (2005)).

- Economists → intergenerational income elasticity.  
Sociologists → transition matrices.
- Most of the estimates of the level of intergenerational socioeconomic mobility use “one generation data”.
- Empirical evidence for the second half of the 20th century → Good estimates of the one generation mobility in income for several developed countries.
- Empirical evidence for the first half of the 20th century → Very scarce (Ferrie (2005) and Ferrie and Long (2005)).

- Economists → intergenerational income elasticity.  
Sociologists → transition matrices.
- Most of the estimates of the level of intergenerational socioeconomic mobility use “one generation data”.
- Empirical evidence for the second half of the 20th century → Good estimates of the one generation mobility in income for several developed countries.
- Empirical evidence for the first half of the 20th century → Very scarce (Ferrie (2005) and Ferrie and Long (2005)).

- Our approach differs from the existing work in two main features:

- 1 We try to measure the correlation between the socioeconomic status of individuals in the current generation and the socioeconomic status of their ancestors at the end of the 19th century.

We do not focus on "one generation" mobility, even though, under some assumptions, we can say something about it and compare with the existing literature.

- 2 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. However, we do not know who the descendants of any specific individual are.

We know the full name of all individuals in both generations.

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.

- Our approach differs from the existing work in two main features:
  - ① We try to measure the correlation between the socioeconomic status of individuals in the current generation and the socioeconomic status of their ancestors at the end of the 19th century.

We do not focus on "one generation" mobility, even though, under some assumptions, we can say something about it and compare with the existing literature.

- ② 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. However, we do not know who the descendants of any specific individual are.

We know the full name of all individuals in both generations.

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.

- Our approach differs from the existing work in two main features:
  - ① We try to measure the correlation between the socioeconomic status of individuals in the current generation and the socioeconomic status of their ancestors at the end of the 19th century.

We do not focus on "one generation" mobility, even though, under some assumptions, we can say something about it and compare with the existing literature.

- ② 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. However, we do not know who the descendants of any specific individual are.

We know the full name of all individuals in both generations.

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.

- The distribution of surnames and the distribution of socioeconomic characteristic are not independent (Collado et al (2008), Güell et al (2008))  $\Rightarrow$  The surnames are useful in our analysis
- Our methodology can be applied to study “long-run” social mobility in many countries that also have the type of data used here.
- Our main data sets are the electoral census of 1898 and the 2001 population census of Cantabria.
- Main conclusion  $\longrightarrow$  The probability of belonging to high status group is correlated with the socioeconomic status of the great-grandfathers and great-great-grandfathers.
- We also compute, under certain assumptions, the average “one generation” level of social mobility  $\longrightarrow$  The level of social mobility found for Cantabria is closer to the US than to Italy (Checchia, Ichino and Rustichini (1999)).

- The distribution of surnames and the distribution of socioeconomic characteristic are not independent (Collado et al (2008), Güell et al (2008))  $\Rightarrow$  The surnames are useful in our analysis
- Our methodology can be applied to study “long-run” social mobility in many countries that also have the type of data used here.
- Our main data sets are the electoral census of 1898 and the 2001 population census of Cantabria.
- Main conclusion  $\longrightarrow$  The probability of belonging to high status group is correlated with the socioeconomic status of the great-grandfathers and great-great-grandfathers.
- We also compute, under certain assumptions, the average “one generation” level of social mobility  $\longrightarrow$  The level of social mobility found for Cantabria is closer to the US than to Italy (Checchia, Ichino and Rustichini (1999)).

- The distribution of surnames and the distribution of socioeconomic characteristic are not independent (Collado et al (2008), Güell et all (2008))  $\Rightarrow$  The surnames are useful in our analysis
- Our methodology can be applied to study “long-run” social mobility in many countries that also have the type of data used here.
- Our main data sets are the electoral census of 1898 and the 2001 population census of Cantabria.
- Main conclusion  $\longrightarrow$  The probability of belonging to high status group is correlated with the socioeconomic status of the great-grandfathers and great-great-grandfathers.
- We also compute, under certain assumptions, the average “one generation” level of social mobility  $\longrightarrow$  The level of social mobility found for Cantabria is closer to the US than to Italy (Checchia, Ichino and Rustichini (1999)).

- The distribution of surnames and the distribution of socioeconomic characteristic are not independent (Collado et al (2008), Güell et al (2008))  $\Rightarrow$  The surnames are useful in our analysis
- Our methodology can be applied to study “long-run” social mobility in many countries that also have the type of data used here.
- Our main data sets are the electoral census of 1898 and the 2001 population census of Cantabria.
- Main conclusion  $\longrightarrow$  The probability of belonging to high status group is correlated with the socioeconomic status of the great-grandfathers and great-great-grandfathers.
- We also compute, under certain assumptions, the average “one generation” level of social mobility  $\longrightarrow$  The level of social mobility found for Cantabria is closer to the US than to Italy (Checchia, Ichino and Rustichini (1999)).

- The distribution of surnames and the distribution of socioeconomic characteristic are not independent (Collado et al (2008), Güell et al (2008))  $\Rightarrow$  The surnames are useful in our analysis
- Our methodology can be applied to study “long-run” social mobility in many countries that also have the type of data used here.
- Our main data sets are the electoral census of 1898 and the 2001 population census of Cantabria.
- Main conclusion  $\longrightarrow$  The probability of belonging to high status group is correlated with the socioeconomic status of the great-grandfathers and great-great-grandfathers.
- We also compute, under certain assumptions, the average “one generation” level of social mobility  $\longrightarrow$  The level of social mobility found for Cantabria is closer to the US than to Italy (Checchia, Ichino and Rustichini (1999)).

# The framework

- Consider a society with no migration flows from outside.
- Let  $P^t$  be the adult population in year  $t = 1, 2, \dots, T$ .
- For each individual in year  $T$  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).
- We assume that for each individual in year  $T$  we observe at least one individual in year 1 belonging to his/her PAL.
- We suppose that there are age brackets for "young" adults in years 1 and  $T$  such that no individual within that age bracket has an adult descendant in that year.

We also assume that any "young" adult in year  $T$  has one ancestor among the "young" adults in year 1.

We denote by  $Y^t$  the set of "young" individuals in year  $t$  ( $t = 1, T$ ).

# The framework

- Consider a society with no migration flows from outside.
- Let  $P^t$  be the adult population in year  $t = 1, 2, \dots, T$ .
- For each individual in year  $T$  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).
- We assume that for each individual in year  $T$  we observe at least one individual in year 1 belonging to his/her PAL.
- We suppose that there are age brackets for "young" adults in years 1 and  $T$  such that no individual within that age bracket has an adult descendant in that year.

We also assume that any "young" adult in year  $T$  has one ancestor among the "young" adults in year 1.

We denote by  $Y^t$  the set of "young" individuals in year  $t$  ( $t = 1, T$ ).

# The framework

- Consider a society with no migration flows from outside.
- Let  $P^t$  be the adult population in year  $t = 1, 2, \dots, T$ .
- For each individual in year  $T$  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).
- We assume that for each individual in year  $T$  we observe at least one individual in year 1 belonging to his/her PAL.
- We suppose that there are age brackets for "young" adults in years 1 and  $T$  such that no individual within that age bracket has an adult descendant in that year.

We also assume that any "young" adult in year  $T$  has one ancestor among the "young" adults in year 1.

We denote by  $Y^t$  the set of "young" individuals in year  $t$  ( $t = 1, T$ ).

# The framework

- Consider a society with no migration flows from outside.
- Let  $P^t$  be the adult population in year  $t = 1, 2, \dots, T$ .
- For each individual in year  $T$  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).
- We assume that for each individual in year  $T$  we observe at least one individual in year 1 belonging to his/her *PAL*.
- We suppose that there are age brackets for "young" adults in years 1 and  $T$  such that no individual within that age bracket has an adult descendant in that year.

We also assume that any "young" adult in year  $T$  has one ancestor among the "young" adults in year 1.

We denote by  $Y^t$  the set of "young" individuals in year  $t$  ( $t = 1, T$ ).

# The framework

- Consider a society with no migration flows from outside.
- Let  $P^t$  be the adult population in year  $t = 1, 2, \dots, T$ .
- For each individual in year  $T$  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).
- We assume that for each individual in year  $T$  we observe at least one individual in year 1 belonging to his/her *PAL*.
- We suppose that there are age brackets for "young" adults in years 1 and  $T$  such that no individual within that age bracket has an adult descendant in that year.

We also assume that any "young" adult in year  $T$  has one ancestor among the "young" adults in year 1.

We denote by  $Y^t$  the set of "young" individuals in year  $t$  ( $t = 1, T$ ).

**Goal** → To analyze the link between certain socioeconomic characteristics of an individual in  $Y^T$  and the characteristics of his/her corresponding ancestor in  $Y^1$ .

Suppose that individuals in  $Y^T$  and in  $Y^1$  can be classified as belonging either to class "high" ( $H$ ) or to class "low" ( $L$ ).

**Definition:** The reproduction rate of any individual in  $Y^1$  is the number of descendants in  $Y^T$ .

**Assumption:** The reproduction rate of any individual in  $Y^1$  is a random variable. The mean of the reproduction rate might depend on the class the individual belongs to.

$f_H$  → Expected reproduction rate of any individual in class  $H$ .

$f_L$  → Expected reproduction rate of any individual in class  $L$ .

The reproduction rate is different from the fertility rate.

**Assumption:** The probability that any individual in  $Y^T$  is of *High* type might depend on the type of his/her ancestor.

$p_H \rightarrow$  The probability that an individual with ancestor of type  $H$  is his/herself of type  $H$ .

$p_L \rightarrow$  The probability that an individual with ancestor of type  $L$  is his/herself of type  $H$ .

Let  $n_K^T$  be a random variable that denotes the number of type  $K$  ( $K = H, L$ ) descendants in  $Y^T$  of any given person in  $Y^1$ .

Given the (conditional) probabilities and the reproduction rates

$$En_H^T = f_H p_H d_H + f_L p_L d_L \quad (1)$$

and

$$En_L^T = f_H (1 - p_H) d_H + f_L (1 - p_L) d_L \quad (2)$$

where  $d_H$  is a dummy variable that takes value one if the ancestor is of type  $H$  and zero if the ancestor is of type  $L$ ,  $d_L = 1 - d_H$ , and  $E$  is the expectation operator conditional on the type of the ancestor.

In a society with "perfect intergenerational social mobility"  $p_H = p_L$ .

# Measures of social rigidity

The odds-ratio of  $p_H$  and  $p_L$ :

$$OR_p = \frac{\frac{p_H}{1-p_H}}{\frac{p_L}{1-p_L}} \quad (3)$$

- A society presents perfect intergenerational **conditional** mobility if  $OR_p = 1$ .

We can also take into account the reproduction rate to assess intergenerational mobility.

If we aggregate equations (1) and (2) for the entire population in  $Y^1$ , we have

$$EN_H^T = f_H p_H N_H^1 + f_L p_L N_L^1 \quad (4)$$

$$EN_L^T = f_H (1 - p_H) N_H^1 + f_L (1 - p_L) N_L^1 \quad (5)$$

$N_K^t$  is the number of type  $K$  ( $K = H, L$ ) individuals in year  $t$  ( $t = 1, T$ ).

We define the expected *outflows-ratio* of individuals of type  $H$  in  $Y^T$  with ancestor of type  $H$  in  $Y^1$  as

$$P_{HH} = \frac{N_H^1 f_H p_H}{N_H^1 f_H p_H + N_L^1 f_L p_L} \quad (6)$$

and the expected *outflows-ratio* of individuals of type  $L$  in  $Y^T$  with ancestor of type  $H$  in  $Y^1$  is:

$$P_{LH} = \frac{N_H^1 f_H (1 - p_H)}{N_H^1 f_H (1 - p_H) + N_L^1 f_L (1 - p_L)} \quad (7)$$

- A society presents **full** intergenerational mobility if

$$P_{HH} = \frac{N_H^1}{N^1} = P_{LH}.$$

Perfect intergenerational **conditional** mobility  $\nRightarrow$  Perfect **full** intergenerational mobility.

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 in  $Y^1$  of each individual in  $Y^T$ , estimating the values of the probabilities  $p_H$  and  $p_L$  and the values of the reproduction rates  $f_H$  and  $f_L$  would be easy.

Unfortunately, for large values of  $T$ , this kind of data is rarely available. However, in some cases there is data on the surname and the class of every individual in  $Y^1$  and in  $Y^T$ .

We propose an "indirect" methodology to estimate those parameters based on the use of surnames.

# Surnames

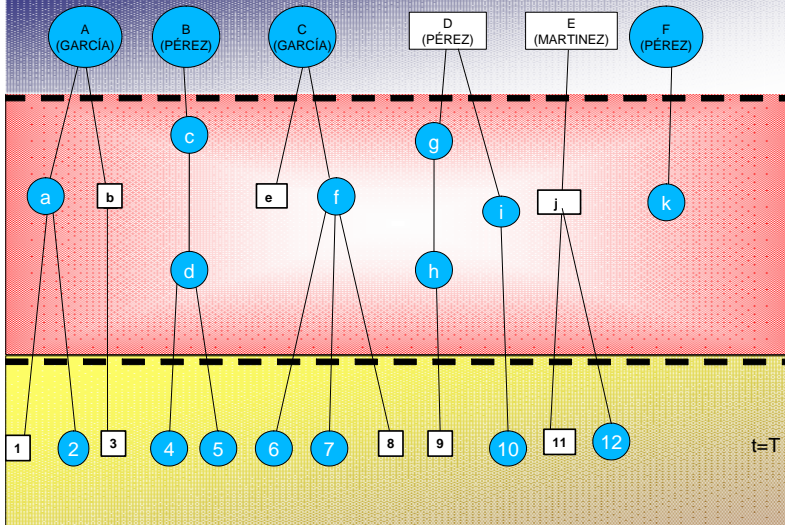
Suppose that all the individuals in society bear a unique surname which is inherited from the father. Thus, all the individuals in the same *PAL* bear the same surname.

However, in most cases the surname is not enough to identify the ancestors of an individual.

The identification problem arises because different people in  $Y^1$  might bear the same surname, and therefore, knowing the surname is not enough to identify for each person in  $Y^T$  who is his ancestor in  $Y^1$ .

Still, for each individual in  $Y^T$  his/her surname delimits the set of his potential ancestors. We will show that, if there is enough variety of surnames and the surnames are not independently distributed across classes we can use them to estimate  $p_H, p_L, f_H$  and  $f_L$ .

t=1



t=T

- Consider two polar situations regarding the variety of surnames in the population.
  - ① 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.
  - ② There are as many surnames as individuals in  $Y^1$ . In this case, we would know with certainty the ancestor of each individual in  $Y^T$  just by knowing the surname.
- Our methodology works if
  - ① The variety of surnames in the population is large enough.
  - ② The percentage of people of type  $i$  high varies across surnames.
- Collado et al 2007 find a specific "bias" in the distribution of surnames. The more uncommon surnames have higher frequencies among high socioeconomic status groups.

- Consider two polar situations regarding the variety of surnames in the population.
  - ① 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.
  - ② There are as many surnames as individuals in  $Y^1$ . In this case, we would know with certainty the ancestor of each individual in  $Y^T$  just by knowing the surname.
- Our methodology works if
  - ① The variety of surnames in the population is large enough.
  - ② The percentage of people of type 1 (high surnames) increases.
- Collado et al 2007 find a specific "bias" in the distribution of surnames. The more uncommon surnames have higher frequencies among high socioeconomic status groups.

- Consider two polar situations regarding the variety of surnames in the population.
  - ① 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.
  - ② There are as many surnames as individuals in  $Y^1$ . In this case, we would know with certainty the ancestor of each individual in  $Y^T$  just by knowing the surname.
- Our methodology works if
  - The percentage of personal types is high across social strata.
  - Collado et al 2007 find a specific "bias" in the distribution of surnames. The more uncommon surnames have higher frequencies among high socioeconomic status groups.

- Consider two polar situations regarding the variety of surnames in the population.
  - ① 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.
  - ② There are as many surnames as individuals in  $Y^1$ . In this case, we would know with certainty the ancestor of each individual in  $Y^T$  just by knowing the surname.
- Our methodology works if
  - ① The variety of surnames in the population is large enough.
  - ② The percentage of people of type High varies across surnames.
- Collado et al 2007 find a specific "bias" in the distribution of surnames. The more uncommon surnames have higher frequencies among high socioeconomic status groups.

- Consider two polar situations regarding the variety of surnames in the population.
  - ① 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.
  - ② There are as many surnames as individuals in  $Y^1$ . In this case, we would know with certainty the ancestor of each individual in  $Y^T$  just by knowing the surname.
- Our methodology works if
  - ① The variety of surnames in the population is large enough.
  - ② The percentage of people of type High varies across surnames.
- Collado et al 2007 find a specific "bias" in the distribution of surnames. The more uncommon surnames have higher frequencies among high socioeconomic status groups.

- Consider two polar situations regarding the variety of surnames in the population.
  - ① 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.
  - ② There are as many surnames as individuals in  $Y^1$ . In this case, we would know with certainty the ancestor of each individual in  $Y^T$  just by knowing the surname.
- Our methodology works if
  - ① The variety of surnames in the population is large enough.
  - ② The percentage of people of type High varies across surnames.
- Collado et al 2007 find a specific "bias" in the distribution of surnames. The more uncommon surnames have higher frequencies among high socioeconomic status groups.

- Consider two polar situations regarding the variety of surnames in the population.
  - ① 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.
  - ② There are as many surnames as individuals in  $Y^1$ . In this case, we would know with certainty the ancestor of each individual in  $Y^T$  just by knowing the surname.
- Our methodology works if
  - ① The variety of surnames in the population is large enough.
  - ② The percentage of people of type High varies across surnames.
- Collado et al 2007 find a specific "bias" in the distribution of surnames. The more uncommon surnames have higher frequencies among high socioeconomic status groups.

# Estimation methodology

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

$$En_H^T = \gamma_{HH}d_H + \gamma_{LH}d_L$$

$$En_L^T = \gamma_{HL}d_H + \gamma_{LL}d_L$$

where

$$\begin{aligned} \gamma_{HH} &= f_H p_H & \gamma_{LH} &= f_L p_L \\ \gamma_{HL} &= f_H (1 - p_H) & \gamma_{LL} &= f_L (1 - p_L) \end{aligned}$$

$\gamma_{JK}$   $\longrightarrow$  The (expected) number of descendants in class  $K$  of an individual of type  $J$ .

We first estimate the "aggregated" parameters  $\gamma_{JK}$  and we then compute the estimates of  $p_H, p_L, f_H$  and  $f_L$ .

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

$$n_{H,j}^T = \gamma_{HH}d_{H,j} + \gamma_{LH}d_{L,j} + \varepsilon_{H,j} \quad (8)$$

$$n_{L,j}^T = \gamma_{HL}d_{H,j} + \gamma_{LL}d_{L,j} + \varepsilon_{L,j} \quad (9)$$

by *OLS*. However, in our case we do not have such information.

What we observe is the type and the surname of each individual in  $Y^1$  and  $Y^T$ .

Notice that observing the surnames delimits the set of potential ancestors of each individual in  $Y^T$ .

Our identification strategy consists in aggregating equations (8) and (9) using the surnames.

Let  $m_{K,s}^t$  be the number of people of class  $K$  ( $K = H, L$ ) with surname  $s$  in  $Y^t$  ( $t = 1, T$ )

$m_{K,s}^1$  is the sum of  $d_{K,j}$  ( $K = H, L$ ) over all  $j$  with surname  $s$

$m_{K,s}^T$  is the sum of  $n_{K,j}^T$  ( $K = H, L$ ) over all  $j$  with surname  $s$ .

Then, aggregating equations (8) and (9) we have

$$m_{H,s}^T = \gamma_{HH}m_{H,s}^1 + \gamma_{LH}m_{L,s}^1 + u_{H,s} \quad (10)$$

$$m_{L,s}^T = \gamma_{HL}m_{H,s}^1 + \gamma_{LL}m_{L,s}^1 + u_{L,s} \quad (11)$$

The properties of the errors in equations (10) and (11) depend on the assumptions made about the errors in equations (8) and (9).

If

$$\begin{pmatrix} \varepsilon_{H,j} \\ \varepsilon_{L,j} \end{pmatrix} \sim iid \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_H^2 & \sigma_{HL} \\ \sigma_{HL} & \sigma_L^2 \end{pmatrix} \right)$$

then

$$\begin{pmatrix} u_{H,s} \\ u_{L,s} \end{pmatrix} \sim iid \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} n_s^1 \sigma_H^2 & n_s^1 \sigma_{HL} \\ n_s^1 \sigma_{HL} & n_s^1 \sigma_L^2 \end{pmatrix} \right)$$

where  $n_s^1$  denote the number of people with surname  $s$  in  $Y^1$ .

$\varepsilon_j = (\varepsilon_{H,j}, \varepsilon_{L,j}) \sim iid \implies u_j = (u_{H,j}, u_{L,j})$  are heteroskedastic.

The form of the heteroskedasticity is known  $\longrightarrow$  We estimate equations (10) and (11) by *GLS*.

# The data

Census data for Cantabria

$t = 1 \longrightarrow 1898$

$t = T \longrightarrow 2001$

## 1898 Electoral Census

Full name, age, address, occupation, and whether the person is illiterate or not, for all male population over 25 years old.

This is a nationwide census, however, it is only available in electronic format for the regions of Cantabria and Murcia  $\longrightarrow$  We use data for Cantabria.

- It is hard to say how representative of the whole country this regional sample is.
- Our theoretical model assumes no immigration flows. The region of Cantabria has been a net exporter of migrants during a large part of the 20<sup>th</sup> century with very reduce immigration flows.

The reproduction rate  $f_H$  ( $f_L$ ) should be understood as the number of descendants from a *High* (*Low*) type person in 1898 who live in Cantabria in 2001.

Sample size  $\rightarrow$  58,666

We select the individuals between 25 and 45 years old  $\rightarrow$  The set  $Y^1$  contains 32,557 men.

We classify people according to the socioeconomic status of their professions.

The High class group ( $H$ ) contains the professions that can be seen as denoting a high socioeconomic status (17.86% of the population) in  $Y^1$ . In the Low class group ( $L$ ) we include all the other professions.

This profession based classification is highly correlated with the education level.

## 2001 Census

Full name, age, occupation, and education level of all the individuals.

Sample size  $\rightarrow$  535,131

We select the individuals between 22 and 45 years old  $\rightarrow$  The set  $Y^T$  contains 150,962 people (men and women).

We classify people according to their education level. We include in the High class all the individuals with a Bachelor degree or a higher education level. This class contains 20.15% of the population in  $Y^T$  (17.63% among men and 22.98% among women).

We also consider an alternative classification of people in  $Y^T$  according to the socioeconomic status of their professions. According to this socioeconomic classification 23.72% belong to the High class (23.10% among men and 24.75% among women).

## 2004 Yellow pages

Name and address of the subscriber and the type of business or professional activity

Sample size → 18816

The number of different professions is about 1000.

We use the Yellow pages:

- As a robustness check

- To compare the results with those for other regions for which the 2001 Census is not available.

# Empirical results

Three Cases:

Case 1 → 2001 Census data, Education level.

Case 2 → 2001 Census data, Socioeconomic status of the profession

Case 3 → 2004 Yellow pages, Socioeconomic status of the profession

# Case 1.

**Table 1**

*Source: Population census 2001. Education groups.*

White's test: 18.02 (p=0.0029)

Parameter	Estimate	SE
$\gamma_{HH}$	1.1394	0.0772
$\gamma_{LH}$	0.8900	0.0204

White's test: 8.79 (p=0.1175)

$\gamma_{HL}$	3.5754	0.2622
$\gamma_{LL}$	3.7299	0.0798

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

$p_H$	$p_L$	$f_H$	$f_L$	$OR_p$	$P_{HH}$	$P_{LH}$	$(N_H^1/N^1) \times 100$
0.2416 (0.0081)	0.1926 (0.0027)	4.7147 (0.3275)	4.6199 (0.0948)	1.3354 (0.0727)	21.78	17.25	17.86

## Gender differences

**Table 3**

*Source: Population census 2001. Education groups.*

Parameter	Estimate	SE
$\gamma_{HH}$ , men	0.5306	0.0381
$\gamma_{HH}$ , women	0.6088	0.0435
$\gamma_{LH}$ , men	0.3894	0.0103
$\gamma_{LH}$ , women	0.5006	0.0119

Parameter	Estimate	SE
$\gamma_{HL}$ , men	1.8888	0.1398
$\gamma_{HL}$ , women	1.6867	0.1270
$\gamma_{LL}$ , men	1.9659	0.0416
$\gamma_{LL}$ , women	1.7641	0.0396

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

	$p_H$	$p_L$	$f_H$	$f_L$	$OR_p$	$P_{HH}$	$P_{LH}$	$N_H^1/N^1$
men	0.2193 (0.0087)	0.1653 (0.0032)	2.4193 (0.1703)	2.3552 (0.0478)	1.4181 (0.0912)	22.86	17.28	17.86
women	0.2652 (0.0104)	0.2210 (0.0031)	2.2954 (0.1615)	2.2646 (0.04841)	1.2719 (0.0827)	20.92	17.21	17.86

## Case 2.

**Table 5**

*Source: Population census 2001. Professions.*

White's test: 7.319 (p=0.1979)

Parameter	Estimate	SE
$\gamma_{HH}$	0.8876	0.0638
$\gamma_{LH}$	0.7049	0.0178

White's test: 8.221 (p=0.1449)

$\gamma_{HL}$	2.2328	0.1569
$\gamma_{LL}$	2.4020	0.0485

**Table 6**

*Mobility parameters and reproduction rates.*  
*Source: Population census 2001. Professions.*

$p_H$	$p_L$	$f_H$	$f_L$	$OR_p$	$P_{HH}$	$P_{LH}$	$(N_H^1/N^1) \times 100$
0.2844 (0.0087)	0.2268 (0.0030)	3.1204 (0.2122)	3.1069 (0.0627)	1.3545 (0.0722)	21.50	16.82	17.86

## Case 3.

**Table 7**

*Source: Telephone directory, 2004. Professions.*

White's test: 14.8818 (p=0.0108)

Parameter	Estimate	SE
$\gamma_{HH}$	0.2351	0.0194
$\gamma_{LH}$	0.1355	0.0052

White's test: 19.7412 (p=0.013)

$\gamma_{HL}$	0.3902	0.0310
$\gamma_{LL}$	0.3275	0.0082

**Table 8***Mobility parameters and reproduction rates.**Source: Telephone directory, 2004. Professions.*

$p_H$	$p_L$	$f_H$	$f_L$	$OR_p$	$P_{HH}$	$P_{LH}$	$(N_H^1/N^1) \times 100$
0.3760 (0.0173)	0.2926 (0.0066)	0.6253 (0.0451)	0.4629 (0.0120)	1.4564 (0.1391)	27.40	20.58	17.86

# One generation results (Case 1, Men).

Under some strong assumption we can recover the "one generation" parameters

- Constant number of generations in all *PAL*.
- Constant reproduction rates across generations.
- Constant conditional probabilities of social mobility across generations.

$$\begin{pmatrix} \gamma_{HH} & \gamma_{LH} \\ \gamma_{HL} & \gamma_{LL} \end{pmatrix} = \begin{pmatrix} \gamma_{HH}^1 & \gamma_{LH}^1 \\ \gamma_{HL}^1 & \gamma_{LL}^1 \end{pmatrix}^g$$

where  $g$  is the number of generations in each *PAL*

$$\begin{aligned} \gamma_{HH}^1 &= f_H^1 p_H^1 & \gamma_{LH}^1 &= f_L^1 p_L^1 \\ \gamma_{HL}^1 &= f_H^1 (1 - p_H^1) & \gamma_{LL}^1 &= f_L^1 (1 - p_L^1) \end{aligned}$$

**Table 10***One generation parameters.**Source: Population census 2001. Education groups.*

<i>3 generations.</i>				
$p_H^1$	$p_L^1$	$f_H^1$	$f_L^1$	$OR_p^1$
0.4841	0.1083	1.3521	1.3284	7.7272
(0.0230)	(0.0048)	(0.0608)	(0.0137)	(1.0113)
<i>4 generations.</i>				
$p_H^1$	$p_L^1$	$f_H^1$	$f_L^1$	$OR_p^1$
0.5696	0.0901	1.2554	1.2371	13.3694
(0.0229)	(0.0045)	(0.0470)	(0.0104)	(1.8095)

Checchia et al 1999.

Italy  $\longrightarrow OR_p^1 = 24.6$ US  $\longrightarrow OR_p^1 = 6.0$

## Last generation results

- Children from 22 to 46 years old.
- Fathers and children living in the same household.
- Fathers living in the same/different province they were born.

**Table 9**

*Last generation*

*Parents and children in the same household*

---

*Odds ratios*

---

<i>Cantabria</i>	<i>Spain (same prov.)</i>	<i>Spain (dif. prov.)</i>
5.0386	5.0455	5.3264
(0.1246)	(0.0455)	(0.1155)

---

# Numerical simulations

- Using the individual data for 1898 and the estimated "one generation" parameters for men in case 1 (reproduction rates and conditional probabilities), we have simulated the individual data for the last generation.
- For each person in each generation we generate his children using a Poisson distribution with expectation equal to the estimated reproduction rate (that depends on whether the father is of class  $H$  or  $L$ ).
- Each children is of class  $H$  with a probability that depends on the status of his father (we also use the estimated conditional probabilities).
- We simulate three generations.
- 1000 replications.

**Table 11***Simulation results**1st generation data : Electoral census 1898.*

<i>Estimates based on observed data</i>				
$p_H^1$	$p_L^1$	$f_H^1$	$f_L^1$	$OR_p^1$
0.2193	0.1653	2.4194	2.3552	1.4182
(0.0088)	(0.0032)	(0.1704)	(0.0479)	(0.0912)
<i>Simulated data (individual observations)</i>				
$p_H^1$	$p_L^1$	$f_H^1$	$f_L^1$	$OR_p^1$
0.2196	0.1653	2.4200	2.3556	1.4214
(0.0042)	(0.0017)	(0.0400)	(0.0189)	(0.0391)
<i>Simulated data (surname observations)</i>				
$p_H^1$	$p_L^1$	$f_H^1$	$f_L^1$	$OR_p^1$
0.2197	0.1652	2.4227	2.3550	1.4245
(0.0085)	(0.0025)	(0.0857)	(0.0250)	(0.0888)

# Conclusions

- We have developed a novel methodology that allows to study intergenerational mobility using Census data for different years.

Our methodology can be applied regardless of the time elapsed between Census.

To apply our methodology we just need to observe some socioeconomic variables and the surnames of individuals for at least two different generations. This information is available in the Population Census of most developed countries.

- We have applied our methodology to study intergenerational mobility in Cantabria during the 20th Century.
- Our result indicate that for a male born in the mid 18th century, the probability that any of his adult descendants at the end of the 20th century has a "High status", compared to probability of "Low status", is 33% higher if he has a "High status" himself than if he has a "Low status".

# Conclusions

- We have developed a novel methodology that allows to study intergenerational mobility using Census data for different years.

Our methodology can be applied regardless of the time elapsed between Census.

To apply our methodology we just need to observe some socioeconomic variables and the surnames of individuals for at least two different generations. This information is available in the Population Census of most developed countries.

- We have applied our methodology to study intergenerational mobility in Cantabria during the 20th Century.
- Our result indicate that for a male born in the mid 18th century, the probability that any of his adult descendants at the end of the 20th century has a "High status", compared to probability of "Low status", is 33% higher if he has a "High status" himself than if he has a "Low status".

# Conclusions

- We have developed a novel methodology that allows to study intergenerational mobility using Census data for different years.

Our methodology can be applied regardless of the time elapsed between Census.

To apply our methodology we just need to observe some socioeconomic variables and the surnames of individuals for at least two different generations. This information is available in the Population Census of most developed countries.

- We have applied our methodology to study intergenerational mobility in Cantabria during the 20th Century.
- Our result indicate that for a male born in the mid 18th century, the probability that any of his adult descendants at the end of the 20th century has a "High status", compared to probability of "Low status", is 33% higher if he has a "High status" himself than if he has a "Low status".