INVESTIGACIONES ECONÓMICAS. vol. XXXII (3), 2008, 259-287

SURNAMES AND SOCIAL STATUS IN SPAIN

M. DOLORES COLLADO
Universidad de Alicante
IGNACIO ORTUÑO ORTÍN
Universidad Carlos III de Madrid and IVIE
ANDRÉS ROMEU
Universidad de Murcia

We study the information contained in surnames on the socioeconomic status of people in Spain. We find that people bearing uncommon surnames tend to enjoy a higher socioeconomic status than people bearing more common surnames. This bias is statistically very significant and robust to different measures of socioeconomic status, and it holds at the national aggregate level as well as at the regional level. The paper offers an explanation of a significant part of such bias as being generated by a signaling behavior by successful dynasties and a low degree of social mobility.

Keywords: Surnames, socioeconomic status.

(*JEL* J700, J000)

1. Introduction

For the last two hundred years surnames in Spain have been passed from parents to children according to the same general rule. People have two surnames, which are inherited from their parents. The first surname is the father's first surname and the second the mother's first surname. By the end of the 18th century the majority of the population already followed this naming convention (see Salazar-Acha, 1991

We thank Juan Mora and two anonymous referees for helpful comments and suggestions. We also thank seminar participants at the Universidad Pais Vasco, Universidad de Valencia and at the 2007 Malaga Summer School on Political Economy. The first author thanks the Spanish Ministry of Education (grant SEJ2005-02829/ECON) for financial support. Ortuño-Ortín gratefully acknowledges the support of the Spanish Ministry of Education (grant SEJ2004-00968) and Fundación BBVA. Romeu is grateful to Fundación BBVA, Ministry of Education (grant SEJ2006-15172) and Fundación Seneca 03024/PHCS/05

and Fauré *et al.,* 2001 for a History of the Spanish surname evolution), which become legally binding after the introduction of the Civil Registry in 1870[1]. Foreign immigration has been only quantitatively important after the end of the 1990's so that most of the surnames borne by the current generations of Spaniards already existed in the 19th century and even earlier.

Thus, surnames contain information that might link current generations with previous ones. Besides the obvious use of this type of information in Genealogy and Family Studies, surnames have been used to analyze several issues in areas such as Population Genetics and Health Sciences. This is because the distribution of surnames contains relevant information about geographical mobility and the mating structure in a society. Since there are links between surnames and genotypes, scientists working in Population Genetics have incorporated the distribution of surnames into the analysis of population genetic structures (see Lasker, 1985, Jobling, 2001 and Colantonio *et al.,* 2003 for a survey). In Health Science, surnames can be useful in studying the relationship between levels of inbreeding and prevalence of certain types of tumours and other diseases of genetic origin (see for example Holloway and Soafer, 1992). In Economics, the study of surnames has been mostly applied in the analysis of very specific discrimination and social integration problems (see for example Einav and Yariv, 2006, Goldin and Shim, 2004 or Bagüés, 2004), in issues of intergenerational transmission of preferences (Collado, Ortuño-Ortin and Romeu, 2006) and particularly, in the study of intergenerational social mobility (Güell, Rodriguez-Mora and Telmer, 2007).

In this paper we study the information of surnames about the socioeconomic status of people using data on surname frequencies in Spain. We use several different sources to collect this information, the most important being the telephone directory (Yellow and White pages). We also use other sources such as, for example, the list of the names of the Spanish university professors, the lists of political candidates to the Congress and the Senate, and the electoral census of 1890 for the Spanish province of Murcia. The total number of surnames in Spain is quite large, thus we do not analyze the socioeconomic status of people

[1]There have been some changes in naming conventions during the past decades. The law now allows for changes in the order of surnames. This practice, however, is rather unusual and responds to personal motivations, for instance preserving the mother's surname in the subsequent generation.

bearing any specific name. We classify surnames by size, i.e. by the total number of people bearing it. For example, the surname with the biggest size is García (with 3.5% of the population bearing it), and there are over 100,000 surnames of size one, i.e. surnames that are borne by just a single person.

We show that there exists a relationship between socioeconomic status and surname size. People enjoying a better socioeconomic status have a larger (smaller) probability of bearing a small-size (large-size) surname as compared to the whole population. This result is statistically very significant and robust to different alternative definitions of socioeconomic status, and it holds at the national aggregate level as well as at the regional level. Moreover, we find that such *bias* on the size of surnames and the socioeconomic status was already present in the 19th century.

Much of this paper is empirically descriptive and seeks different sources of evidence to support the above claim. However, in the last section of the paper we also try to get a deeper insight on this socioeconomic bias of the Spanish surname distribution. We claim that most of the bias arises as the result of the combination of two forces: a low degree of social mobility and the role of the surname as a signaling device for successful dynasties. Many of the smallest surnames in the 20th century did not exist in the 19th century. They proceed from different sources, namely, foreign surnames, new spelling variants of old surnames and the combination of two previous surnames into a new double-barrelled surname. Regarding the latter, Bagüés (2004) already noticed a high frequency of double-barrelled surnames among certain high-status professions in Spain. We investigate this issue and claim that an important part of the socioeconomic bias arises from such double-barrelled surnames in the following way. A successful person (and its offspring) with rather uncommon first surname passes the surnames in the standard way. However, a successful person (or its offspring) with a very common first surname might combine such surname with its second one to form a double-barrelled first surname (often joined with a hyphen) to distinguish himself or his descendants from other people[2]. Thus, every time a person enjoying a high so-

[2] The practice of changing the name to make it more distinguished is probably very old and widespread across different societies. However, the way it has been done in Spain since the XIX century is different from the one observed in other countries where a new surname is formed when a man's and woman's family names are combined on marriage. Changing the name to make it more distinguished seems to

cioeconomic status creates a new surname by combining his first very common surname with his second one, the number of people with high status level and common surnames decreases and the opposite happens with the number of people with high status and uncommon surnames. It will be argued that this signaling practice explains an important part of the current generation bias in socioeconomic status and size of the surname. Such bias also requires a lack of perfect intergenerational social mobility. If the socioeconomic status of people were independent of the status of their parents the bias would arise only from successful people themselves combining their two surnames. Their descendants would enjoy a high status with the same probability as anybody else in society so that would not contribute to the bias. We will show that, not surprisingly, the data does not support such a perfect social mobility scenario.

Even though the creation of double-barrelled surnames explains an important part of the bias, it is not enough to fully explain it. Thus, we will argue that the bias that still remains to be accounted for after allowing for the double-barrelled practice is the result of migration processes and/or bias inherited from the 19th century. Many foreign migrants bear surnames that are *small* in Spain. If immigrants tend to have a high socioeconomic status, part of the bias could be explained by such foreign surnames. Most recent foreign immigrants have only one surname and have been excluded from our sample. However, native descendants of foreign immigrants have two surnames and could not be excluded from our sample. The effect of this flow of new surnames on the bias is however inconclusive: it may either reinforce or decrease the bias depending on the relative social status of immigrants. Furthermore, the number of adult descendants of immigrants in Spain is quite low and, consequently, the possible contribution of their surnames to the bias is probably unimportant. Thus, we shall argue that most of the current bias that is not generated by double-barrelled surnames is mostly inherited from the 19th century bias. Before the introduction of the Civil Registry in 1870 there were different ways

be a very common practice in Spain already in the XVII century, as the following poem by Calderón de la Barca shows: "Si a un padre un hijo querido/a la guerra se le va,/ para el camino le da/ un Don y un buen apellido./ El que Ponce se ha llamado/se añade luego León,/ el que Guevara, Ladrón/y Mendoza el que es Hurtado./ Yo conocí un tal por cual/ que a cierto Conde servía/ y Sotillo se decía;/ creció un poco su caudal/ salió de mísero y roto,/ hizo una ausencia de un mes,/ conocile yo después/ y ya se llamaba Soto./ Vino a fortuna mejor, eran sus nombres de gonces,/ llegó a ser rico y entonces/ se llamó Sotomayor."

to change one's surname and it was not legally binding to follow the traditional naming convention. Many high-status people in the 19th century gave to their offspring distinguished surnames (that in most cases were very "small") which did not have to be double-barrelled. We investigate this issue analyzing the 1890 census of a Spanish region and find a bias quantitatively similar to the current one. However, we also find that indeed the frequency of double-barrelled surnames was much lower than in the present period. If we are right and part of the current bias comes from such adoption of small surnames during the 19th century it suggests that the degree of social mobility in Spain has probably been quite low.

Güell, Rodríguez-Mora and Telmer (2007) have developed a very interesting method for measuring the degree of intergenerational social mobility using a measure of the information content of surnames in one census. In the theoretical model they characterize the joint distribution of surnames and income and then analyze the link between social mobility and the information content of surnames. Although the main focus of the empirical part of their paper is measuring social mobility, they also show that surnames contain socioeconomic information and, similar to our results, find a negative relationship between size of the surname and socioeconomic status. Namely, in an OLS regression of years of education against surname frequency they obtain a negative coefficient. Our work differs from theirs in a number of ways. First, they use census data, which is clearly an advantage with respect to our data sets. However, the census they use only refers to the region of Catalonia whereas we use several alternative sources of data covering the whole country. Second, we measure the bias using a different methodology which does not restrict the relationship to be linear, and therefore it allows us to quantify the magnitude of the bias across the whole surname distribution. Third, we assess the contribution of the new double-barrelled surnames to the bias, and we argue that such surnames are created as a signaling device for successful dynasties.

The remainder of the paper is organized as follows. Section 2 provides a description of the statistical methodology used in the paper. Section 3 describes the main database used, namely, the telephone directory, and contains the main empirical results. Section 4 provides the empirical results using other databases. Section 5 analyzes data from the 19th century. Section 6 provides an explanation of our main empirical

findings. Section 7 concludes with suggestions for possible extensions and future research.

2. Statistical analysis

Surnames will be informative about the socioeconomic status if the distribution of surnames differs substantially among different socioeconomic groups. We will use data on the distribution of surnames for the whole population and for several samples of people from different socioeconomic groups. Then, our testing strategy will focus on the comparison between the observed distribution of surnames in a given sample and the distribution of surnames in the whole population[3]. Even though we have information on the specific surnames of the whole population, we are not going to directly compare the distribution of all surnames in the sample and in the population. The reason is that we are not interested in the information any particular surname may have, but on the potential relationship between the frequency of the name and the socioeconomic status. In particular, we claim that people bearing very infrequent surnames are more likely to enjoy a high socioeconomic status than people bearing very common names. Then, what we have to do is to group surnames according to their size, where the size is the number of people that bear the surname in the population. Thus, we will have surnames of size one, that are those bear by just one person in the population, surnames of size two, etc., and the random variable we will focus on will be the surname size. This grouping strategy allows us to test whether there is a relationship between the size of the surname and the socioeconomic status.

Let $P = \{p_1, p_2, ..., p_M\}$ be the class partition of all the surnames according to their sizes, where M is the largest size in the population. Thus, surname i belongs to partition element p_j if and only if the size of surname i is equal to j. In the way we define the class P some of its elements might contain no surnames, i.e., there might be j such that $p_j = \{\varnothing\}$. Typically, we will find that the first sets in P contain many surnames and the last ones contain just one surname. For example,

[3] A particular characteristic of the Spanish surnames is that each person has two surnames. Then, we split each actual individual in two fictitious individuals: the first one associated to his/her first surname and the second one associated to his/her second surname. For example, a person with surnames García López yields two persons in our analysis, one person named García and a second person named López.

in Spain there are more than 100,000 different surnames of size 1 (i.e. in set p_1), while López is the only surname in $p_{472,702}$ and the classes $p_{462,761}$ to $p_{472,701}$ are empty. If, we denote by $|p_j|$ the number of people bearing a surname from the set p_j, then, the probability distribution of the variable of interest, that is the surname size, is given by the total share of p_j

$$q_j = \frac{|p_j|}{N},$$

where N is the number of people in the population.

Assume we know the names of people having some socioeconomic characteristics belonging to a set Y. Typical elements of this set are different type of jobs and education levels. We assume that some elements of Y can be classified as denoting a high socioeconomic status. For example, in the case where the characteristic $i \in Y$ indicates *being a Doctor* we say that a person with such trait has a high socioeconomic status. Other traits, however, might not be identified with any socioeconomic level. Thus, the trait *being a public servant* does not contain much information on the social status of people having it.

Take any specific characteristic $i \in Y$. We denote by q_j^i the probability that a person with characteristic i has a surname of size j[4]. For example, if i indicates *being a Doctor*, q_j^i is the percentage of doctors in the population with surnames of size one. If names do not contain any information on the socioeconomic status of agents, these probabilities are independent of the trait, and therefore,

$$q_j^i = q_j \text{ for all } j = 1, 2, ..., M,$$

and this is the null hypothesis that we are going to test.

Given that the surname size is a discrete variable, we could think of using a chi-squared goodness of fit test. The problem that arises at this stage is that some of the population probabilities q_j are too small to use a chi-squared test, even for rather large sample sizes. A solution would be to reduce the curse of dimensionality by considering a more reduced partition of the set of surnames, but then the question would be to determine which criteria to follow in order to choose a given partition since two different partitions may give contradictory results. For this reason we chose not to use a chi-square but a Kolmogorov-Smirnov (KS) test. The KS test was first designed for continuous distributions

[4] i.e. a surname belonging to partition element p_j

but can be adapted to the case in which the distribution under the null has a finite number of *jumps*. The advantage of the KS test is that it is designed for continuous distributions, and therefore, we do not need to worry about negligible probabilities. Then, we will use the whole size partition $P = \{p_1, p_2, ..., p_M\}$. Without considering those sizes for which $p_j = \{\varnothing\}$, this class partitions the whole set of surnames in 2,825 subsets. Let's denote by $F(s) = \sum_{j=1}^{[s]} q_j$ the distribution function of the surname size in the population, i.e. $F(s)$ is the proportion of people in the population with surname size smaller or equal to s. Let's denote by $\hat{F}(s) = \frac{1}{H}\sum_h I_{\left\{s_h^i \leq s\right\}}$ the empirical distribution of the surname sizes in a given sample, where H is the sample size and s_h^i is the size of the surname of individual $h, h = 1, 2, .., H$. I.e. $\hat{F}(s)$ is the proportion of individuals in the sample with surnames of size smaller or equal to s. Then, the (scaled) KS statistic is

$$L_H = \sup_s \sqrt{H}\left|\hat{F}(s) - F(s)\right|$$

The KS test is a non-parametric test that is extensively used in the literature. When the distribution $F(s)$ is continuous, the asymptotic distribution of L_H under the null is characterized in Shorack and Wellner (1986) among others. Thus, the 5% critical value of the asymptotic test is 1.359 approximately. When the null distribution has a finite number of *jumps* or discontinuity points, as it is the case in our function $F(\cdot)$, asymptotic critical values for L_H can be obtained by simulation, provided that $F(\cdot)$ is known at each jump. Appendix A3 shows how to obtain these critical values in our application.

The KS test will allow us to detect any deviation from the null hypothesis that surnames are uninformative about socioeconomic status. However, if the null is rejected, the KS test cannot help us in assessing the direction of the bias. For example, if we reject the null we will not be able to know whether is due to having more small size surnames in the sample than expected under the null, or more large size surnames, or any other reason. Then, we propose a complementary test procedure, which requires further grouping, but has the advantage of detecting the direction of the potential deviation from the population distribution. Let's first consider the random variable f_j^i that denotes, for any random sample of H people of status i, the proportion of people in the sample with surname of size j. This random variable has mean

$$E\left[f_j^i\right] = q_j^i$$

and variance

$$V\left[f_j^i\right] = \frac{q_j^i\left(1-q_j^i\right)}{H}.$$

As we explain above, the hypothesis that surnames are uninformative about socioeconomic status implies that for any given trait i the probability of having a surname of size $j(q_j^i)$ is independent of the trait and coincides with the probability for the whole population (q_j). Then, for a random sample of H people with trait i, we could test the null hypothesis

$$H_0 : q_j^i = q_j,$$

for any $j = 1, 2, .., M$, using the test statistic

$$\frac{f_j^i - q_j}{\sqrt{\frac{q_j(1-q_j)}{H}}}.$$

This test statistic has a limiting normal distribution under the null. The problem we face is again that for most surname sizes (except for the most frequent ones), the population probabilities q_j are very small and therefore the test proposed above does not perform very well even for large sample sizes. Thus, we divide surnames in a small number of classes m. These classes are defined using a size criterion in the following way: Class 1 will contain the surnames in the population with smallest size that are those included in $P_1 = \{p_1, p_2, ..., p_h\}$, so that $r_1 = q_1 + q_2 + ... + q_h$ is approximately equal to $1/m$. Class 2 will contain the surnames in $P_2 = \{p_{j_1+1}, p_{j_1+2}, ..., p_{j_2}\}$, so that $r_2 = q_{j_1+1} + q_{j_1+2} + ... + q_{j_2}$ is again approximately equal to $1/m$ and so on. The last class m, will then contain the largest size surnames in the population. Our test consists in testing whether for any particular trait i the probability of having a surname in class k is equal to r_k. If we denote by s_k^i the probability of a person of status i has a surname in class k, for a random sample of H people with trait i, we can test the null hypothesis

$$H_0 : s_k^i = r_k$$

for any $k = 1, 2, .., m$, using the test statistic

$$\frac{g_k^i - r_k}{\sqrt{\frac{r_k(1-r_k)}{H}}}$$

which also has a limiting normal distribution under the null, where g_k^i is the proportion of individuals in the sample with a surname in class

k. The problem with this procedure is that the results may depend on the choice of m. Then, we will use different values of m to check the robustness of our results.

In the next section we will consider several characteristic denoting a high socioeconomic status. In all the cases, the difference $g_k^i - r_k$ shows a similar pattern: it is positive for uncommon surnames (i.e. when k is small) and negative for common ones (i.e. when k is large). Furthermore, $g_k^i - r_k$ is decreasing in the size of the surname. Our tests confirm that this departure of g_k^i from the *non-information* value r_k is statistically significant. Moreover, the magnitude of such bias is quite substantial among the smallest and the largest sized surnames.

3. Data from the Telephone Book

Our main data source is the 2004 Spanish telephone directory. The directory is available on a commercial CD-ROM (INFOBEL, http://www.infobel.com) which contains 11.5 million domestic users (White Pages) and provides information on the full name and address of the subscriber, including the province and the zip code. The sample size is considerable since the total population in Spain in 2004 was 43.2 million, and the total number of main family residences was around 14.4 million. For this reason we consider the distribution of surnames in the white pages of the telephone directory as the distribution of surnames in the Spanish population. As we explain in Section 2, we use the two surnames of each person, and therefore, the population size is around 23 million. There are over 220,000 different surnames in the telephone directory, and there are several reasons to believe that most of them are of Spanish ancestry. First, there have been no significant foreign migration inflows in the modern history of Spain. Second, although in 2004 the number of foreign residents was about 6% of the total population (see INE) most of them arrived very recently and probably a large majority of them are not listed in the directory. Finally, we do not consider individuals bearing single surnames, which in Spain are unmistakably recognized as people with a foreign nationality. In any case, to check that our results were not affected by recent foreign migration flows we repeated all our calculations using the 1999 telephone directory[5], obtaining very similar results to the ones reported here.

[5] In 1999 the number of foreign residents was as low as 1.4% of the total population (see OECD 2002).

The telephone directory also contains a business section (Yellow Pages) with 800,000 numbers registered under a person's name[6]. This business section provides information on the name and address of the subscriber and the type of business or professional activity. The number of different professions is 1,169. Since subscribing to the Yellow Pages is voluntary, we were concerned on the issue of geographical sample bias since the geographical distribution of surnames in Spain is not homogeneous across territories. First, we checked whether the distribution of surnames appearing in the Yellow Pages represents a geographically unbiased sample of the distribution in the White Pages, and we found that this hypothesis should be rejected as some of the provinces are over-represented. Second, there might be professions that are more popular in certain provinces than in some others. Then, for each profession in the Yellow Pages, and in order to account for the potential bias, we correct the observed population frequencies at the national level in the following way: We take as the national frequency of any given surname a weighted average of the provincial frequencies in the White Pages, where the weight of each province is the proportion of observations in the sample of the Yellow Pages that correspond to that province.

We will now present our results for a selection of professions from the Yellow Pages. In the first exercise we consider all the doctors, lawyers and pharmacists listed in the Yellow pages (around 63,000 people). We view these three professions as indicating a high socioeconomic status. The KS test statistic for this sample is 26.4 with a p-value smaller than 0.0001[7]. Therefore, we strongly reject that the surnames of doctors, lawyers and pharmacists are a representative sample of the distribution of surnames in the Spanish population. Next, in order to see the direction of the bias, we group the surnames according to their size in m=10 classes[8]. In Figure 1(a) we present, for each class, the percentage differences in the proportion of people in the sample and in the population, relative to the proportion in the population[9]. For instance, in the first class that corresponds to the smallest surnames, this

[6] Notice that all sample sizes in this paper are two times the number of people because we use both the first and the second surname of each person.

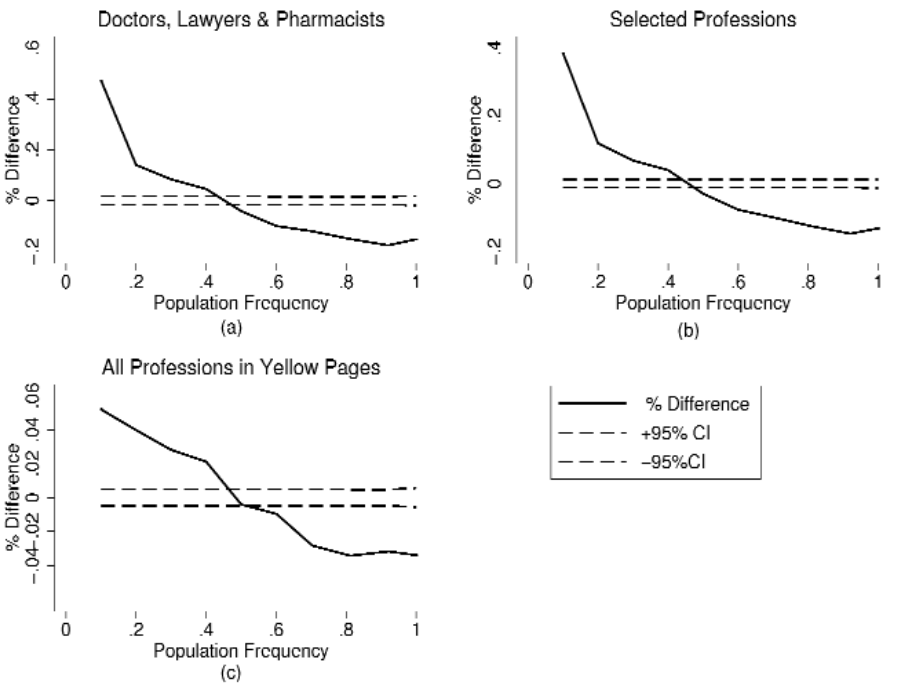
[7] This and the subsequent results on the KS test are presented in Appendix 1.

[8] To check the robustness of our result to aggregation level, we have repeated the tests using 20 classes and 100 classes and the qualitative results presented below do not change.

[9] Notice that the percentage differences are very closely related to the second test statistic we present in Section 2. The test statistic is $(g_k^i - r_k)/\sqrt{(r_k(1-r_k)/n)}$,

figure is 47%. This means that, in class 1, there is 47% more people than there should be if the surnames of doctors, lawyers and pharmacist were a representative sample of the distribution of surnames in the whole population. It is also interesting to see what happens when we concentrate in the very *small* surnames. In particular, defining *very small* surnames as those with size smaller or equal to 20 the bias is very large. We find that this group represents a 2.5% of the people in the population, whereas this share is 4.1% among doctors, lawyers and pharmacists. Regarding the top 20 most frequent surnames we find that *large* surnames represent 27% of the whole population, whereas this figures is only 22.6% in the sample.

FIGURE 1



Thus, among people in these professions, and compared to the society average, there is an *excess* of small sized surnames and a *shortage* of large sized surnames. Moreover, it seems that such bias changes in a monotone way with the surname sizes. I.e., the smaller the size of the surname is the higher the share of people in these professions as compared to the whole population. We have repeated the same test

whereas the % *Difference* refers to $(g_k^i - r_k)/r_k$. The confidence lines in the figures are adapted from the confidence limits of the test.

for each of the three professions independently obtaining in each of them very similar results to the one in Figure 1(a).

Our second exercise considers all the professions and business listed in the Yellow Pages which require that the person in charge of it should hold at least a Bachelor Degree diploma (around 125,000 people)[10]. We view each of these professions as representing a high socioeconomic status of the person. One might claim, however, that some of these professions should not be classified as indicative of a high socioeconomic status, or at least should not be seen as enjoying the same status as, for example, doctors. Nevertheless, we believe that the majority of them can be seen as not belonging to low socioeconomic groups, even though many of them are not considered as prestigious as doctors. Thus, we think that is safe to assume that all such professions are relatively high-status.

The KS global test indicates again that the sample of these professions is not representative of the total population of surnames in Spain. The statistic yields a value of 30.8 with a p-value smaller than 0.0001. By analogy to Figure 1(a), Figure 1(b) shows the percentage differences for the sample of Selected Professions. It is reassuring to see that the same bias found for doctors/lawyers/pharmacists in Figure 1(a) appears in this more comprehensive case. Notice that among people with these more frequent professions 3.8% have *very small* surnames (2.4% in the whole population), and only 23% have *large* surnames (27% in the whole population).

However, we cannot regard all the professions outside the previous group of professions as belonging to a low socioeconomic status. Many economically very successful and/or highly educated persons might run businesses that are listed in the Yellow Pages and do not require a Bachelor Degree. In fact, one might think that being listed in the Yellow Pages is a signal of certain success. Thus we next analyze the distribution of surnames for the whole set of professions in the Yellow Pages. Figure 1(c) shows the results of our test in this case. The same type of bias found above is still present although it is less pronounced. The KS test rejects with a test statistic of 16.9 and a p-value smaller than 0.0001. It is important to notice that there are around 700,000

[10]The list of such professions was independently elaborated in a subjective way for each of the three authors. It turned out that the number of discrepancies was extremely small and an agreement was easily reached. The list is available upon request.

professionals listed in the Yellow Pages, i.e., around a 1.6% of the whole Spanish population. Therefore, the bias cannot be associated with a much reduced number of professions without any impact on the overall population distribution.

One might claim that the bias found above is due to a geographical effect. The distribution of surnames and the level of social and economic development vary among regions. It might happen that within each region there is no bias in the size of the surnames and the socioeconomic status and still such bias does appear at the national level. Thus, we have repeated the previous tests first for all the provinces belonging to the former Kingdom of Aragon and second for the remaining provinces. To further check for a possible regional effect we have also carried out our tests for Catalonia[11]. All our basic results remain unchanged for all these regional level cases. Appendix A2 contains the results of these tests and the percentage differences for the sample of Selected Professions for these three regional cases.

4. Other Sources

To check the robustness of our previous result we have gathered data from several other sources and repeated our tests. In principle, any trait analyzed here could be more common in certain provinces than in some others. Then, for each trait, and in order to account for such potential geographical bias, we correct the observed population frequencies at the national level as described in Section 3.

In all the cases analyzed here the results are similar to the ones presented in the previous section. For professions or traits associated with high socioeconomic status there is a bias towards small sized surnames and against large sized surnames.

4.1 University professors

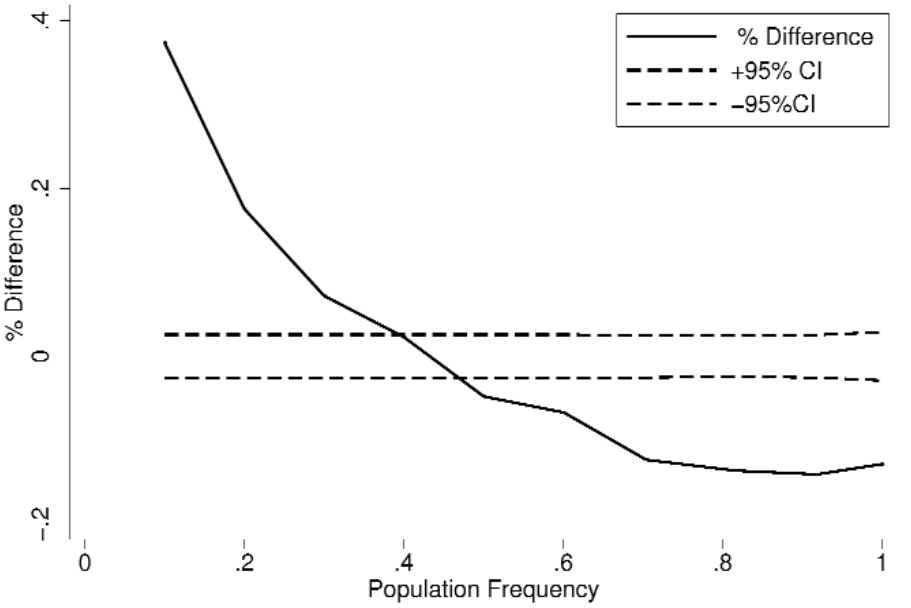
We analyze a list of 26.000 professors at the Spanish public universities. This list basically contains the whole population of full and

[11] Most of the regions in Spain belong to the formers kingdoms of Aragon and Castile. The kingdom of Aragon comprised the current regions of Aragon, Catalonia, Valencia and Baleares. Catalonia has about 6 million inhabitants and is one of the wealthiest regions in Spain. Catalonia has its own language and differentiated cultural and social traditions.

[11] This list is available at http://www.mec.es/educa/ccuniv/html/habilitacion/documentos/conv_21_09_05/Lista_Sorteables_Definitiva_2005.zip or from the authors upon request.

associate university professors in Spain in 2005. The KS test statistic is 14.8 and the p-value is smaller than 0.0001. Thus we reject the hypothesis that the surname distribution of the university professors is an unbiased draw from the whole population distribution. Figure 2 shows percentage differences in this case. Notice that this figure is very similar to Figure 1.

FIGURE 2
University Professors



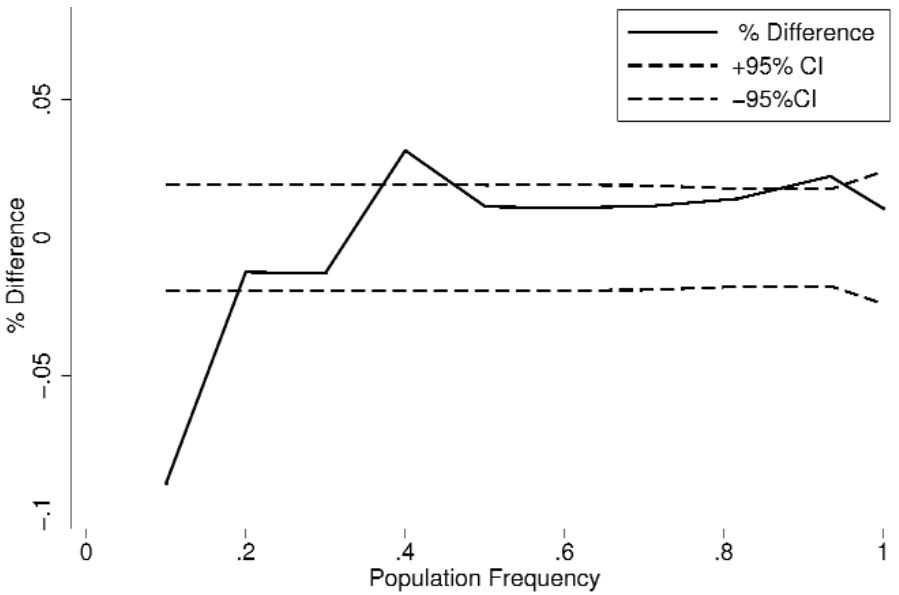
4.2 Job candidates to civil servant administrative assistant positions

The list contains the names of around 48,000 applicants to these jobs in year 2004[12]. These jobs do not require a high qualification. However, they are permanent and offer excellent benefits and job conditions so that some highly qualified people apply to them. Thus, a priori, one should expect no clear bias in any direction. The KS test statistic is 3.69 with a p-value smaller than 0.0001, thus rejecting the null that the sample is a good representation of the population. Our test then shows that these applicants are not a random sample of the population, however, the test statistic is much smaller than in our previous examples. Figure 3 shows the percentage differences for this sample.

[12]The list was published by the Spanish Ministry of Public Administration in April 2004, and it is available from the authors upon request.

The shape of the graph is very different from the other figures. The value of $g_k^i - q_k$ is always within the confident limits except at population frequencies 0.1, 0.4 and 0.9, and the direction of the bias is clearly not decreasing with the surname size.

FIGURE 3
Candidates to Administrative Assistant Possitions

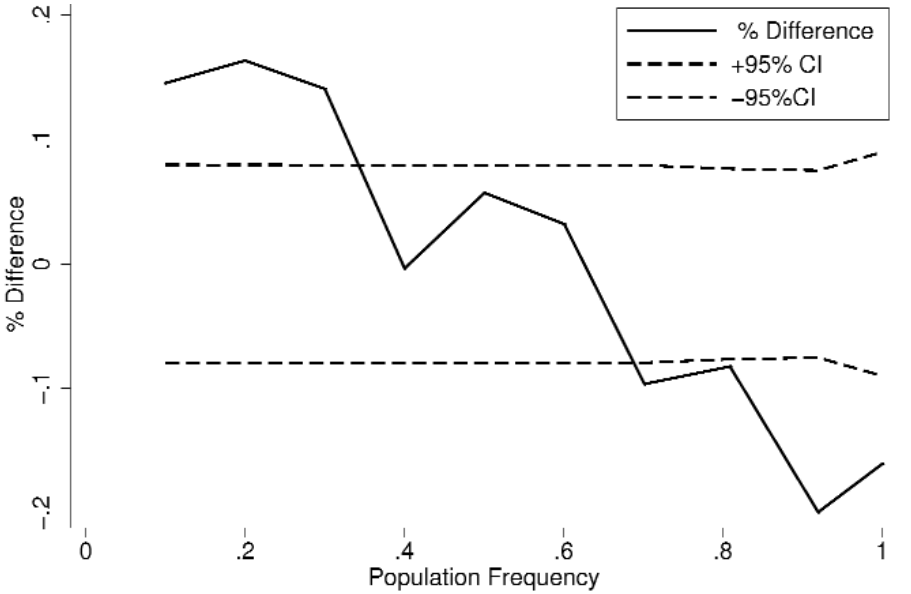


4.3 Political candidates

We examine the surnames of all the candidates from the main political parties running in the 2004 Election to the National Congress and National Senate. It contains 5,872 surnames (i.e., 2,936 candidates) from eight parties[13]. The KS test statistic is 4.04 with a p-value smaller than 0.0001 so that we reject the null that these politicians are a representative sample of the Spanish surnames. Figure 4 shows again that there is overrepresentation of small sized surnames and under representation of large surnames. Thus, we find again that *being a political candidate* presents the typical surname bias of a high socioeconomic trait.

[13] The parties considered here are: Partido Popular (PP), Partido Socialista (PSOE-PSC), Izquierda Unida (IU), Partido Nacionalista Vasco (PNV), Esquerra Republicana (ER), Convergencia i Unio (CiU), Bloque Nacionalista Galego (BNG), Eusko Alkartasuna (EA) . All of them achieved parliamentary representation and represent 97% of seats in Congress and 98% of the elected seats in Senate.

FIGURE 4
PSOE+PP+IU+Nationalists



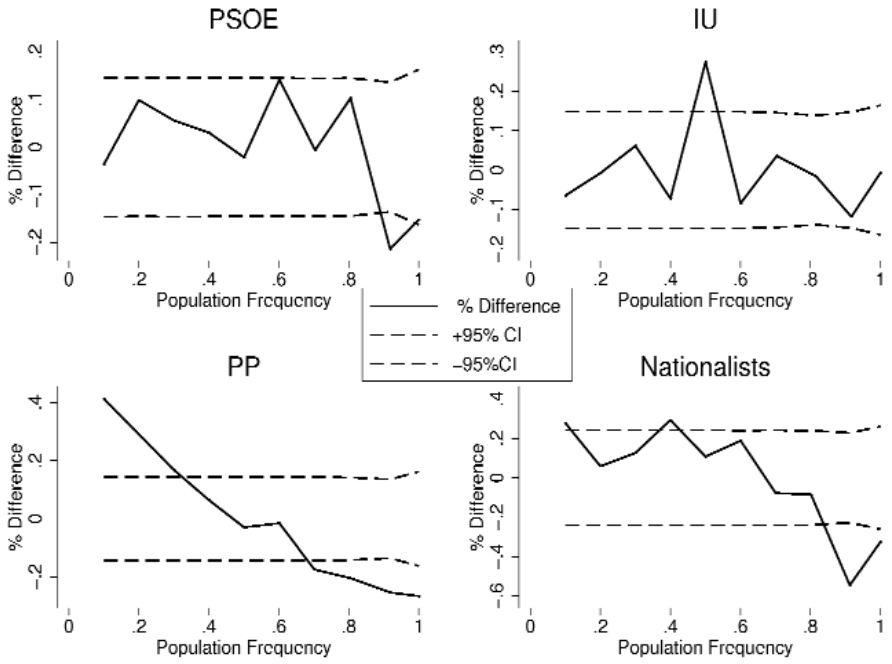
It might be interesting to study the possible bias for each party separately. Thus, we carry out the tests for each of the three main political parties at the national level, PSOE, PP and IU. The remaining parties (ERC, CiU, BNG, PNV and EA) are regionalist and only run for the provinces of their respective regions (ERC and CiU only run in Catalonia, PNV and EA run in the Basque Country and BNG in Galicia), and the sample size for each of them is very small. To solve this problem we have aggregated these five parties under the nationalistic label and test for their possible overall bias[14]. The results are shown in Figure 5.

It turns out that the bias is very clear among members of the conservative party (PP). The figure corresponding to the socialist party (PSOE) shows a more ambiguous result. It is almost representative of the population, with values of $g_k^i - q_k$ most of the time within the confident limits but still showing a significant bias for the largest size

[14] Obviously, we compare the distribution of the surnames of the candidates of these nationalistic parties with the population distribution of surnames in the regions they run, and not with the distribution in the whole country.

surnames. The KS test also confirms this situation, with a clear rejection in the case of the conservative party (KS test: 3.94, p-val<0.0001) and in the limit for PSOE (KS test: 1.54, p-val=0.018). The nationalistic parties show a bias, which seems not so strong as in the case of the conservative party but still with a value of the KS test of 2.70 and p-value<0.0001. It seems that the only clearly unbiased party is IU, (a coalition of communists, social-liberals and ecologists) which shows a KS test of 0.85 with a p-value of 0.4455.

FIGURE 5



5. Historical data

In this section we investigate whether the bias found in the current generation is also present in data from previous generations or not. This is an interesting question in itself, and could also be very useful to provide an explanation of how such bias has arisen. Unfortunately, there are not many large sample data bases in electronic format with information on names and social status of people in previous generations. So far we only have the electoral census of 1890 for the province of Murcia[15]. The data from this source suggest that the same type of bias found in the late 20th century already appears in the 19th century.

The national electoral census of 1890 contains the full name, age, address, occupation, and whether the person is illiterate or not, for all the Spanish male population over 25 year old. The census is only available in electronic format for the province of Murcia[16] and this is the only information we analyze here. Thus, our results should be taken with caution for two reasons. First, it is hard to tell how representative of the whole country this provincial sample is[17]. Second, some *small* surnames in Murcia might be medium size surnames in other Spanish regions. This would complicate the interpretation of the bias. Murcia is one of the poorest provinces in Spain and during the 19th century was a net exporter of migrants to other regions. However, it probably received highly qualified people such as doctors, judges and teachers from other provinces. If most of those immigrants bore surnames that were medium sized in the whole country but small in Murcia, we could find a bias due just to the variation of surnames frequencies across provinces. In that case, were the whole census available, one might obtain no bias nationwide. Even though we suspect that this possible problem due to highly qualified immigrants is quantitatively very small, in principle, we cannot discard it.

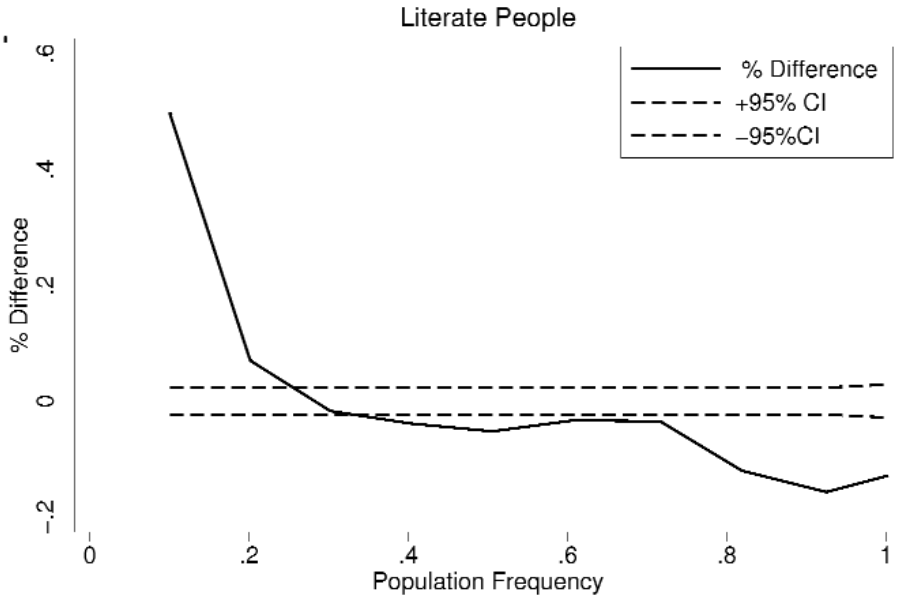
[15] We also have the 1788 *Marqués Ensenada* Census for the city of Malaga. Our tests show a certain bias on economic status and surname size here. However, the sample size is very small so that we cannot be fully confident of its true existence. These tests are available from the authors upon request.

[16] The electoral census for the province of Murcia is available online from the Spanish Ministry of Culture in PDF format at http://www.mcu.es/prensahistorica/. Using Optical Character Recognition software we have processed the files and converted into spreadsheets.

[17] There are 50 provinces in Spain. Provinces are an administrative division and they are rather homogeneous in extension and size. In 1887, the population of Murcia was 451,611 which represented a 3.1% of the 17,534,416 Spaniards (Spanish demographic census, 1887).

The number of people in the Murcia census is around 106,000 and we classify any person who is able to read and write as belonging to the high socioeconomic group (a 27.9% of the whole population in the Census). Figure 6 shows the same type of bias in this case as the one found for current high socioeconomic traits in the previous sections. The KS test indicates rejection with a test statistic of 14.0 and p-value smaller than 0.0001.

FIGURE 6
Electoral Census 1890



6. Why do surnames contain information about the socioeconomic status of people?

In the previous sections we have shown that, at least for the 20th century, there is strong statistical evidence that surnames contain information on the socioeconomic status of people. Moreover, the nature of such information is different from the one observed in other countries. In U.S., for example, it is clear that many surnames contain important information, mostly if they are associated to certain ethnic groups. For example, citizens with Hispanic surnames or distinctively black names (Fryer and Levitt, 2004) tend to belong to low socioeconomic groups. In Spain, the number of foreign immigrants, and descendants from foreign immigrants, in 2000 was extremely low so

that the explanation cannot be based on straightforward arguments as the different social status of immigrants.

How then could the bias described in this paper arise? Since the number of small surnames is very large (there are 130,414 surnames of size one, 29,516 of size two and 13,454 of size three), it is unfeasible to study directly how all of them were created. However, there is a possible explanation that is consistent with our empirical findings. We should first notice that until the late mid 19th century there was not a legally binding rule regarding the adoption of surnames in Spain. It is true that the majority of the population followed the convention at that time. However, some people bore surnames different from their parent's surnames. It is only after the introduction of the Civil Registry in 1870 that the convention becomes legally binding. Thus, for the last 130 years the legal rule has been that the first surname of a person is the first surname of the father and the second surname is the first surname of the mother. There is, however, one exception to this general rule. The law allows the combination of the two surnames of a person into a new one to be passed on to their offspring. For example, a man with surnames García López marries a woman with surnames Martínez Pérez. Under the general rule their children would have surnames García Martínez. The exception consists in the possibility of combining the father's surnames (it could also be the mother's surnames) so that the new double-barrelled surname García-López is created and, in this case, the children would bear Garcia-Lopez as the first surname and Martínez as second. This possible merger of surnames is not immediate and has to be approved by the legal authority[18]. We observe that such type of double-barrelled surname count for a significant share of the small size surnames. For instance, 10.3% of the surnames of size 1 are double-barrelled, whereas this proportion is only 0.7% in the whole population. Thus, one might suspect that at least a significant part of the bias found in the previous sections is related to this type of surname.

We propose the following explanation of the way such a bias on the socioeconomic status and the size of surnames arises: Successful people, or their offspring, might want to be unambiguously identified by other people. If they already have a very uncommon surname they can be easily recognized. However, this is not the case when they bear

[18] In most cases, the person applying for the merger is the son or daughter who must prove that the father (or mother) was known by his two surnames.

a very common surname. Thus, successful families with very common surnames would like to adopt a new more *distinguished* surname. As explained before, since 1870 the way to create a new surname consists in combining the original surnames. An example might clarify our point. Suppose that a person with surnames García Sal marries a woman with surnames Martínez Pérez. García is a very common surname, Sal, however, is a very unusual one. Say that such person becomes very successful, for example, he becomes a very famous doctor in the country or in his region and is known as *doctor Garcia Sa.l* His offspring might want to be recognized as the descendant of such a distinguished person (they might want to do that thinking that it will help in their social relations or just for vanity or ostentation). Under the general rule their surnames would be García Martinez, two very common surnames. Thus, they might want to combine the father's surnames so that they would be named García-Sal Martínez and be easily recognized as descendants of the renowned doctor Garcia Sal[19]. One can also suppose that in the case the two surnames of our *famed doctor* were very unusual the incentives to combine them would be much lower.

Why don't unsuccessful people combine the surnames? Because is costly in several aspects. First, as mentioned before, combining the surnames is not automatically granted and requires time and certain resources. Second, we suspect that social approval of their peers plays an important role here and merging the surnames might be seen as an unfounded pretentiousness or snobbishness on the part of the person doing it. Thus, our claim is that *mostly successful people with common surnames create new surnames.*

Assume next that, in the short run, there is no perfect intergenerational social mobility so that children and grandchildren of a successful person are more likely to enjoy a high socioeconomic status than children of people of low socioeconomic status. Thus, if our hypothesis is correct, most new double-barrelled surnames appear among successful families with very common (first) surnames, which implies that people of high socioeconomic status are shifted from the set of large sized surnames to the set of small sized surnames. This would explain the relatively large proportion of small sized surname among high so-

[19] Many people specify, for brevity, only one of the two surnames. Thus, in this example, it might happen that the descendants mostly use the first surname, García-Sal, so that they are even more easily recognized as related to the famous doctor.

cioeconomic status people, and the relatively small proportion of large sized names.

The following empirical observations support our hypothesis:

a) Among people with high socioeconomic status there is an excess of double-barrelled surnames[20]. For example, among the selected professions in the Yellow Pages analyzed in Section 2 there is 2.3% of people with double-barrelled surnames, as compared to 0.9% for the whole population in the White Pages.

b) Most of the double-barrelled surnames are of very small size. Thus, more than half of them are of size one and 90% of them are of size smaller than 16. This implies that most of them are of recent creation.

c) When the double-barrelled names are truncated and only the first part of the name is considered (for the example García-Sal is truncated and considered as García), the bias becomes much smaller. Figure 7 shows our test for the selected professions from the Yellow Pages when all the surnames are included (the same case as in Figure 1.b) and in the case double-barrelled surnames are truncated[21]. For the smallest size class, the proportion of people from the sample of selected profession is 13.9%. When double-barrelled surnames are truncated such proportion becomes 12.6%. Thus, double-barrelled names account for 1/3 of the bias. A test of proportions shows that these differences in the bias with and without double-barrelled surnames are statistically significant with a p-value$<$0.0001.

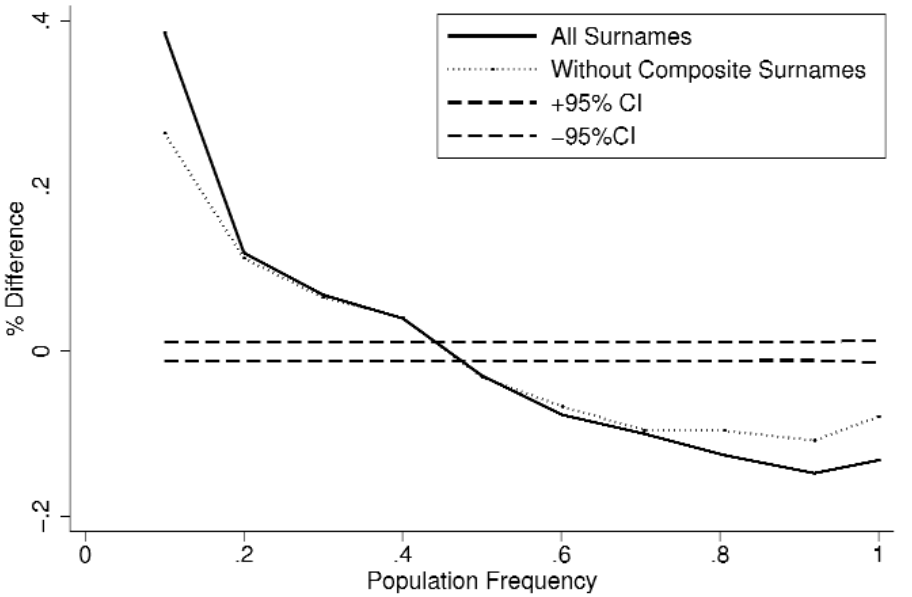
Nevertheless, Figure 7 shows that there is still a bias when double-barrelled surnames are truncated. How can we explain it? We believe that such bias is mainly inherited from the one present in the 19th century (see Figure 6). As mentioned before, until 1870 most people followed the traditional naming convention, but there was a great flexibility in the way surnames could be adopted. Thus, successful families did not need to combine the parental surnames to create a distinguished surname. They could, for example, change the order of the surnames or adopt the surname of a famous relative different from the parents' surnames. This possibility is consistent with our finding

[20] We consider as double-barrelled any composite name as García-Atienza or García de Atienza. Names as Del Río, De la Torre, etc are not considered as double-barrelled.

[21] A very similar result is obtained for the case of Doctors, Lawyers and Pharmacists as well as in the case of all the professions in the Yellow Pages.

that in 1890 the bias look similar to the one found a century later and yet the proportion of double-barrelled surnames is only 0.6%, which is much lower than the percentage today. Of course, for the actual bias without considering double-barrelled surnames to be a consequence of the existing bias in the 19th century, the degree of social mobility in Spain during the last century must have been relatively low.

FIGURE 7
Selected Professions



Summing up, our explanation of the bias is based on the lack of perfect social mobility and the signalling behaviour of successful dynasties. Until 1870 those dynasties had much flexibility in the way they could choose a *distinguished* surname. From 1870 on, the almost unique way to do it consists in combining the surnames. Thus, the share of the actual bias that is not inherited from the 19th century bias has been created by new successful dynasties with common surnames that, by combining them, generate a new differentiated surname.

7. Further research

We have shown that surnames contain information on the socioeconomic status of people in Spain. Thus, there is a negative relationship between the size of the surname and the probability that the person bearing it belongs to a high socioeconomic status. This result holds at the aggregate national level as well as the regional level. Moreover, such relationship is not exclusive to a reduced number of very specific surnames. During the last 130 years surnames have been adopted following a very well defined rule. This fact, together with the socioeconomic information contained in surnames, can be used to study the degree of social mobility during the last generations (see Güell *et al.* 2007 for an interesting method to analyze social mobility using the information content of surnames in one census). We leave it for further research.

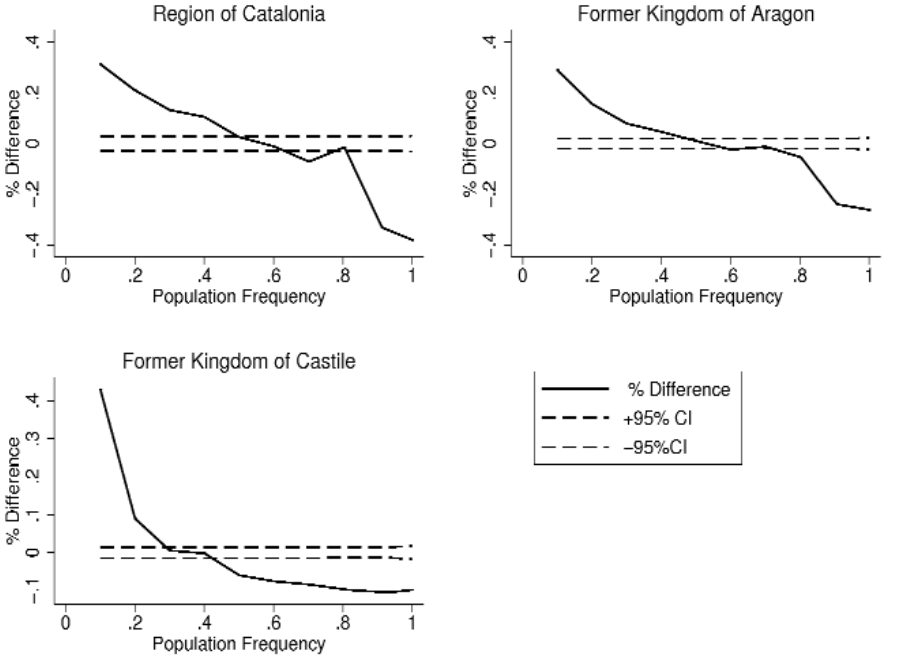
Appendix A1.

CUADRO A1
Results of the KS Tests (p-values in parenthesis)

SAMPLE	KS TEST STATISTIC AND P-VALUES				
	By region				
	All	*Catalonia*	*Kingdom of.Aragon*	*All except K. Aragon*	
Doctors, Lawyers, Pharmacists	26.4036	14.4336	14.8550	18.9381	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	
High educated	30.7650	16.8673	17.1026	21.9246	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	
All professions	16.9312	14.4103	13.2559	7.1528	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	
University Professors	14.7942				
	(0.0000)				
Applicants to Assistant Position in the Administration.	3.6916				
	(0.0000)				
	Political Parties				
	All	*PSOE*	*PP*	*IU*	*NAT*
Politicians	4.0407	1.5374	3.9438	0.8497	2.7004
	(0.0000)	(0.0179)	(0.0000)	(0.4455)	(0.0000)
Literate people in Murcia in 19th. Century	14.0153				
	(0.0000)				

Appendix A2. Graphical Tests for Selected Professions in different regions

FIGURE A2
Selected Professions



Appendix A3. A note on the KS Test

We first normalize the variable "size" so that its support is the [0,1] interval. I.e. we define the variable $\varsigma = s/S$, where s is the original size and S is the maximum size in the population. The distribution function of ς then is

$$F_\varsigma(\varsigma) = F(S_\varsigma) = \sum_{j=1}^{[S_\varsigma]} q_j \text{ and } \hat{F}_\varsigma(\varsigma) = \frac{1}{H}\sum_h I_{\left\{\frac{s_h^i}{S} \leq_\varsigma\right\}}.$$

Then, we can rewrite the KS test statistic L_H in terms of the ς as:

$$L_H = \sup_{\varsigma\in[0,1]} \sqrt{H}\left|\hat{F}_\varsigma(\varsigma) - F_\varsigma(\varsigma)\right|.$$

As shown in Billingsley (1999), the empirical process L_H weakly converges to the process $\sup_{\varsigma\in[0,1]}\left|W^0_{F(\varsigma)}\right|$ where $W^0(\cdot)$ is the standard Brownian bridge. Now, say $\varsigma_1, ..., \varsigma_M$ the (normalized) discontinuity points of $F(\cdot)$. It is immediate to see that finding the supremum of $\left|W^0_{F(\varsigma)}\right|$

on $\varsigma \in [0,1]$ is equivalent to search for a maximum of the sequence $\left|W^0_{F(\varsigma 1)}\right|, ..., \left|W^0_{F(\varsigma_M)}\right|$.

Thus, in order to find the asymptotic critical values of L_H we carry out the following steps:

1. Get a draw from a standard Brownian motion process $W(\cdot)$ by simulating a random walk along a grid with random normal increments of very small variance.
2. Linearly transform the Brownian motion with $W^0(\varsigma) = W(\varsigma) - \varsigma W(1)$ to get a Brownian bridge.
3. Compute the absolute value of the Brownian bridge at the points $\varsigma_1, ..., \varsigma_M$ and take the supremum.
4. Keep the supremum and go back to 1 until the desired number of iterations, which in our case was constant and equal to 10,000.

Finally, it is important to note the following:

a) The discretization of the limiting process detailed above would not be necessary in practice if the jumps of the distribution were sufficiently small. Given the big number of surnames in the population, the jumps of the distribution are indeed rather small for most of them, except for the upper tail of the distribution i.e., the most common surnames. For example, the surname Garcia has a frequency of around 3.5% of the population. Nevertheless, the discrepancy between the critical values of the continuous and discretized limiting distributions of the empirical process were found to be very small in practice.

b) Since $W^0_{F(\varsigma)}$ is a discretization of the standard Brownian bridge at the points $\varsigma_1, ..., \varsigma_M$, it turns out that the supremum of its absolute value must always be equal or lie below the supremum of $\left|W^0_\varsigma\right|$ for any realization. Therefore, the critical values from the standard KS test tables are conservative in the sense that one can safely reject using these tables.

References

Bagüés, M. (2004): "¿Qué determina el éxito en unas oposiciones?", mimeo.

Billingsley, P. (1999): "Convergence of probability measures", in Willey Series in *Probability and Mathematical Statistics,* John Willey and Sons, New York.

Colantonio, S.E., Lasker, G.W., Kaplan, B.A. and V. Fuster (2003): "Use of surname models in human population biology: A review of recent developments", *Human Biology,* 75, 6, pp. 785-807.

Collado, M.D., Ortuño-Ortín, I. and A. Romeu, (2006): "Vertical transmission of consumption behavior and the distribution of surnames", IVIE WP-AD 2006-09.

Einav, L. and L. Yariv (2006): "What's in a surname? The effects of surname initials on academic success", *Journal of Economic Perspectives* 20, pp. 175-188.

Fauré, R., Ribes, M. and A. García (2001), *Diccionario de apellidos españoles,* Espasa, Madrid.

Fryer, R.G., Jr. and S. D. Levitt (2004): "The causes and consequences of distinctively Black Names", *Quarterly Journal of Economics* 119, pp. 767-805.

Güell, M., Rodríguez Mora, J.V. and C. Telmer (2007): "Intergenerational mobility and the informative content of surnames", CEPR Discussion Paper 6316, June 2007. Available at www.cepr.org/pubs/dps/DP6316.asp.

Goldin, C. and M. Shim (2004): "Making a name: Women's surnames at marriage and beyond", *Journal of Economic Perspectives* 18, pp. 143-160.

Holloway, S.M. and J.A. Sofaer (1992): "Coefficients of relationship by isonomy among registrations for five common cancers in scottish males", *Journal of Epidemiology and Community Health* 46, pp. 368-372.

Jobling, M.A. (2001): "In the name of the father: surnames and genetics", *Trends in genetics* 17, pp. 353-57.

Lasker, G.W. (1985), *Surnames and Genetic Structure,* Cambridge University Press, New York.

OECD (2002), *Trends in International Migration,* Publications Service, Paris.

Salazar-Acha, J. (1991), *Génesis y evolución histórica del apellido en España,* Real Academia Matritense de Heráldica y Genealogía, Madrid.

Shorack, G.R. and J.A. Wellner (1986): "Empirical processes with applications to statistics", in Willey Series *in Probability and Mathematical Statistics,* John Willey and Sons, New York.

Resumen

En este trabajo se estudia la información contenida en los apellidos sobre la categoría socioeconómica de la población en España. Encontramos que las personas con apellidos poco frecuentes tienden a tener un nivel socioeconómico mayor que las personas con apellidos más comunes. Este sesgo es estadísticamente muy significativo y robusto ante cambios en la manera de medir la categoría socioeconómica, y se observa tanto a nivel nacional como a nivel regional. Se ofrece una explicación de una parte importante de ese sesgo basada en un comportamiento señalizador de las dinastías con éxito junto a un bajo nivel de movilidad social.

Palabras clave: Apellidos, nivel socioeconómico.

Recepción del original, marzo de 2007
Versión final, febrero de 2008