

Vertical transmission of consumption behavior and the distribution of surnames*

M. Dolores Collado[†]
(Universidad de Alicante)

Ignacio Ortuño-Ortín
(Universidad de Alicante)

Andrés Romeu
(Universidad de Murcia)

August 2006

Abstract

This paper attempts to detect the existence of links in consumption preferences between generations. Preferences for consumption goods may be determined by the preferences of parents (vertical transmission) and/or by preferences arising from the environment (horizontal transmission). We propose an indirect methodology to overcome the lack of data on consumption choices of dynasties, i.e., parents and their adult offspring. This new approach is based on the analysis of the correlation between the geographical distributions of surnames and consumption choices. Our results show that there is horizontal transmission of preferences regarding non-food items and possibly vertical transmission for food items.

JEL classification: **D12, R23.**

Keywords: Preference formation, surnames, vertical and horizontal transmission.

*We specially thank Luis Ubeda for his very useful suggestions and comments. We also thank Jaime Kahhat, Javier Ruíz-Castillo, Klaus Desmet, Christian Schultz and participants at The Econometric Society World Conference, London 2005, The European Economic Association Meeting, Amsterdam 2005, and seminars' participants at Universidad Carlos III, Universitat Pompeu Fabra, UAB and CEMFI. We are grateful for financial support from the IVIE, the Spanish Ministerio de Educación y Ciencia SEJ2005-02829/ECON and Fundación BBVA 3-04X.

[†]Corresponding author: Departament of Economics, Universidad de Alicante, 03080-San Vicente (Alicante), Spain. collado@merlin.fae.ua.es

1 Introduction

A central theme in the study of preference formation is the relative influence of the family versus the influence of the social environment in shaping a person's preferences. Thus, preference formation can be understood as emerging from a vertical transmission from parent to offspring, and a horizontal transmission between any two individuals. Vertical transmission of preferences can take place through children's imitation of their parents' tastes or through the parents' teaching of certain habits and values (and perhaps through some genetic inheritance). Horizontal transmission occurs when children adopt the habits of other agents outside the household, as for example when they learn values or preferences taught in school or when they imitate friends¹.

This paper provides a novel empirical approach to assess the importance of the parental and the environmental influence on the formation of children's preferences over *consumption goods*. Namely, we study the transmission of the rates of substitution for consumption goods as when a child adopts his parents' strong taste for eating meat rather than being vegetarian. However, we neither seek to describe the precise working of those channels of preference transmission nor attempt to provide any new theoretical explanation on the issue of preference formation. The paper just provides an empirical framework for assessing the relative impact of vertical and horizontal transmission of preferences.

Since preferences are not observable we will have to use data on consumption behavior instead. It will be argued that, after controlling for income and other variables that may affect consumption decisions, a correlation on the preferences of parents and their offspring implies a correlation on the consumption bundles chosen by parents and the consumption bundles chosen by their (adult) offspring. If there were data available on those consumption choices, assessing the significance of vertical transmission would be an easy task. Unfortunately, unlike other types of intergenerational transmission, such as earnings or abilities, we are not aware of the existence of any survey that provides such information on a large number of consumption goods.

Thus, we propose an indirect strategy to overcome the data problem: we will compare the spatial distribution of consumption behavior and the spatial distribution of surnames. By studying whether regions with similar surname distributions also have similar expenditure patterns, we will be able to draw some conclusions on the effect of vertical versus horizontal transmission of

¹This terminology is based on Cavalli-Sforza&Feldman [5] (see also Bisin&Verdier [2]). These authors also distinguish between horizontal and oblique transmission. We do not make such a distinction and we use the term horizontal transmission for both types.

preferences.

The intuition behind this approach is simple: if vertical transmission is very strong, preferences are transmitted from parents to children in a similar way to surnames. Consider an economy composed by geographical regions that originally were quite different in both the distribution of surnames and the distribution of consumption patterns. Suppose that there have been important recent migration flows among the regions. If vertical transmission is perfect, so that surnames and preferences are transmitted in the same way, regions that have experienced large migration flows should be close both in the distribution of surnames and in the distribution of expenditure patterns. In contrast, if horizontal transmission of preferences is the main force in the preference formation process, surnames and preferences evolve in different manner and regions that have experienced more migration flows will have more similar surnames distributions but no more similar consumption patterns. Thus, we will be able to detect vertical transmission by studying the correlation between the distributions of surnames and the distributions of consumption patterns among regions. More specifically, we will construct two matrixes of distances between regions. The first matrix will be computed by calculating the distances in consumption patterns between regions and the second matrix by computing the distances in the distribution of surnames between regions. It will be argued that if vertical transmission plays a significant role, after controlling for other relevant variables, those two matrixes should be positively correlated. The case of no correlation between the two matrixes will indicate a strong horizontal transmission effect. Moreover, the stronger the linkage between parent's preferences and children's preferences the slower immigrants adapt to their host societies. Thus, the analysis of the correlation of such matrixes will also enable us to draw some new insights on the issue of immigrant integration.

The main result obtained in the paper indicates that there exists a positive and significant correlation between parents' preferences and those of their offspring, i.e. vertical transmission might play a significant role in the formation of consumption preferences. However, when consumption goods are divided into food and non-food items our results show *little or no vertical transmission for non-food consumption goods and vertical transmission for food items*.

The statistical tool used to asses the possible correlation between distances in the distribution of surnames and distances in the distribution of consumption patterns is the multivariate Mantel Test, which is a method for testing (linear) correlation between distance matrices. The Mantel test has been applied to problems of Spatial Autocorrelation in Ecology and in Population Genetics (see Mantel [23], Sokal&Rohlf [31] and Legendre&Legendre

[22]). However, to the best of our knowledge, this is the first time that this test has been applied in Economics (see Desmet, LeBreton, Ortuno-Ortin and Weber [11] for another recent application in Economics).

In this paper we use data from Spain, although the approach is general and the data needed is also available for many other countries. There are several reasons why Spain is an excellent case study. First, there were large internal migration flows in the last century.² Second, the number of foreign immigrants in Spain until the late 90's was very low compared to other countries. This is important because having a very high number of foreign immigrants would substantially complicate the analysis, as it would be necessary to have information on consumption patterns from the immigrants' countries of origin. Third, information on surnames is easily available in electronic format from the telephone directory. Fourth, the Spanish surveys on household consumption are of high quality as documented in Browning&Collado [4]. Finally, provinces are relatively small areas and data are available at provincial level.

There is an extensive literature dealing with intergenerational correlations in, for example, earnings (Solon [32]), wealth (Charles&Hurst [7]) abilities and aggregated consumption (Mulligan [25]), IQ (Daniels et al [8], Feldman et al [13]), political orientation (Jennings et al [18]), intertemporal preferences (Becker&Mulligan [1]) and altruistic preferences (Mulligan [24]). However, Waldkirch, Ng and Cox [34] is, to the best of our knowledge, the only work dealing with the type of "intratemporal" preferences considered in this paper. They study the intergenerational transmission of consumption preferences using data from the PSID, which contain information on the *total* food expenditure of parents and of their adult offspring. They are the first to investigate potential intergenerational correlation in consumption beyond that induced through an intergenerational transmission of permanent income. After controlling for income and other relevant variables, they find a significant intergenerational transmission on tastes for food. Unfortunately their data set does not provide more disaggregated information on consumption choices, which is the sort of information required to analyze intergenerational transmission of preferences over consumption bundles.

The distribution of surnames in the population has been used to analyze several issues in areas such as Population Genetics and Health Sciences. This is because it 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

²At the beginning of the seventies, over 20% of the Spanish citizens were living in a province different from the one they were born.

the distribution of surnames into the analysis of population genetic diversity (Lasker [21] and Jobbling [20]). In Health Science, surnames can be useful in studying the relationship between levels of inbreeding and prevalence of certain types of tumor and other diseases of genetic origin (see for example Holloway&Soafer [17])

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&Yariv [12], Fryer&Levitt [14] and Goldin&Shim [15]) but not in issues of intergenerational transmission of preferences. One possible reason why there have not been more studies on surnames is that large data sets were not available in electronic format until very recently. However, things have changed dramatically in the last few years, and in most developed countries telephone directories on CD-ROM are now easily available and contain information about basically all households.³

The structure of the rest of the paper is as follows. In Section 2 we provide a simple theoretical model of vertical and horizontal transmission of preferences and explain our approach relating surnames and consumption patterns. Section 3 presents the data on surnames and consumption. Section 4 presents the statistical tests and our main results. Section 5 concludes with some final comments and suggestions for further work.

2 Theoretical Framework

2.1 A Simple Model on Consumption Preference Formation.

The empirical analysis in the subsequent sections is based on the following simple model. Suppose that each agent lives for two periods: as a child and as an adult. Each adult agent has a child. An agent only consumes in the second period. The same n consumption goods are available to each generation.⁴ The preferences of an (adult) agent are represented by the utility function $U(c; \alpha)$ where $c = \{c_1, c_2, \dots, c_n\}$ is the consumption vector and $\alpha \in R^m$ is the vector of parameters of the utility function. Thus, we identify the preferences of an agent with the vector of parameters α of his

³In some countries, such as the UK, there are other comprehensive sources available, e.g. national census and electoral registers.

⁴Thus, we assume that the set of consumption goods remains the same across generations. In reality such set changes dramatically in the long run. However, the composite goods that we consider in the empirical part can be regarded as quite stable in recent decades.

utility function. Given the vector of prices $p = \{p_1, p_2, \dots, p_n\}$ and the income level w , an agent with preferences α demands the consumption vector $c = X(p, w; \alpha)$, where $X()$ is the demand function.⁵ We assume that preferences are formed during childhood. Thus, in the first period of life an agent adopts certain preferences, which will remain constant through his adult life. In the following sections we will try to assess the possible correlations between the preferences of parents and children. Our model, however, does not look at the specific channel of preference transmission from parents to children, and it is compatible with the possibility that children mimic particular consumption behavior of parents and/or that consumption behavior is attributed to genetic inheritance (see Rowe [29] and Harris [16] for the view that parental influence on child outcomes is limited). Let α_t denote the preferences of an adult agent at period t , α_{t-1} the preferences of his father at the previous period and $\hat{\alpha}_{t-1}$ the vector of preferences of the remaining adult agents at period $t - 1$. We take the view that preferences follow the law of motion

$$\alpha_t = F(\alpha_{t-1}, \hat{\alpha}_{t-1}, \varepsilon) \quad (1)$$

The effect of the first variable, α_{t-1} , on α_t can be seen as the *vertical transmission* of preferences, and the effect of the second variable, $\hat{\alpha}_{t-1}$, as the *horizontal transmission* of preferences. The variable ε represents personal characteristics that are uncorrelated with parental and environmental preferences. Thus, the influence of parents on a child's preferences is not a deterministic one-to-one. Therefore, even if there is no horizontal transmission, children will not perfectly inherit their parents' preferences.

A more general approach would assume that the law of motion is

$$\alpha_t = F(c_{t-1}, \alpha_{t-1}, \hat{\alpha}_{t-1}, \varepsilon) \quad (2)$$

so that the preferences of an agent are given by a function that depends on his father's preferences, on society's preferences and on c_{t-1} , the vector of goods consumed by his father. In this case, vertical transmission would work through two channels: a "direct" transmission of preferences given by $F(\cdot, \alpha_{t-1}, \cdot)$ and an "indirect" transmission⁶ given by $F(c_{t-1}, \cdot, \cdot)$. Notice that the vector c_{t-1} depends on the income of the father so that this indirect transmission could be rewritten as a function of income. In the empirical part, we control for income so that our approach could also fit into this more general framework.

⁵In reality the demand function also depends on other variables such as household structure, the age of the consumer and so on. In the empirical part we will control for several such variables.

⁶This indirect transmission might be modelled following the literature on habit formation. See, for example, Pollak [27]

Since we do not observe α_t and α_{t-1} the correlation between them cannot be estimated directly. One possible approach to overcome this problem is to analyze the correlation between the consumption vectors c_{t-1} and c_t . A correlation between the vectors of preference parameters should, controlling for prices and income, be associated with a correlation on consumption vectors, i.e., vertical transmission of preferences should imply vertical transmission of consumption behavior.

2.2 Vertical Transmission, Surnames and Consumption Preferences.

Unfortunately, and contrary to the cases of transmission of earnings or abilities, there are no good data available on the consumption vector of parents and their (adult) offspring to check for such possible intergenerational link on consumption. Thus, apart from some very limited surveys on consumption of specific goods, there is no information on the vectors c_{t-1} and c_t . Therefore, we follow an indirect approach in order to detect the existence of vertical transmission. This approach is based on assessing how surnames and consumption patterns are distributed across different geographical regions of the country using a measure of geographical dissimilarity. The key idea is that surnames remain unaltered when transmitted vertically across generations, while consumption patterns may be determined by horizontal and/or vertical transmission.

An example might help to clarify the central idea of this paper. Consider two provinces, A and B , and let c_I be the mean consumption vector and y_I the vector of surname frequencies of agents in province I ($I = A, B$). Let $d(\cdot)$ be a distance measure in the space of consumption patterns and surname frequencies. Assume that at a certain period in time provinces A and B do not share any surname in common, so that if the j^{th} element of y_A is zero the same element of y_B is not zero and vice versa.⁷ Suppose that individual preferences are very different across regions, and therefore, consumption patterns are also very different.⁸ These two assumptions imply that both $d(c_A, c_B)$ and $d(y_A, y_B)$ are large. Suppose that a representative agent moves from A to B and has a child born in B . In the next period the old agent dies and his child becomes an adult and stays in B . What are the implications of this migration in terms of $d(c_A, c_B)$ and $d(y_A, y_B)$? The distance in surnames, $d(y_A, y_B)$

⁷This is an extreme assumption, it is only required that in each province there is a set of surnames that is distinctively more frequent than in any of the other provinces.

⁸We are implicitly assuming that prices are the same in both provinces and all agents have the same income.

will decrease since the child bears a surname that did not exist in province B before. However, the distance in consumption $d(c_A, c_B)$ will decrease or remain constant depending on the importance of vertical transmission. Under full vertical transmission the child of the immigrant consumes the same vector as his father and this implies that $d(c_A, c_B)$ will also decline. This reasoning shows that, under full vertical transmission, migration implies a decrease in surname and consumption distances. Therefore, in the second period, the correlation between the distance in surnames and the distance in consumption will be positive. In the event that vertical transmission plays no role and the child of the immigrant acquires the preferences of province B , the distance $d(c_A, c_B)$ would remain constant. However, the children of the natives might also acquire the preferences of the immigrants and, in this case, the distance $d(c_A, c_B)$ would decrease. Thus, a positive correlation between the two distances in the second period is a *necessary but not sufficient* condition for vertical transmission.

We will construct a matrix of surname distances and a matrix of consumption distances between the provinces of mainland Spain. Based on the idea above, we claim that vertical transmission should be reflected in a positive and significant correlation between these two matrices.

3 Data on Surnames and Consumption

3.1 Surname Distribution in Spain

We consider the geographical surname distribution in Spain in 1999 extracted from the telephone directory. The directory is available on a commercial CD-ROM (INFOBEL, <http://www.infobel.com>) which contains 11.5 million domestic users and provides information on the full name and address of the subscriber, including the province and the zip code.⁹ The total population of Spain in 1999 was around 40 million and the total number of main family residences was around 14 million. As mentioned in the introduction, the number of foreign immigrants in Spain until the late 90's was very low compared to other countries. Even in 1999, the number of foreign residents was 1.4% of the total population compared to 6.1% in France and 8.8% in Germany (see OECD [26]). Furthermore, because there have been no significant foreign migration inflows in the modern history of Spain the surnames analyzed here are largely of Spanish ancestry.

⁹To the best of our knowledge this is the first paper using the information on the surnames of all the telephone users in Spain and not just a sample of them.

The naming convention in Spain is different from most Western countries, and similar to that prevailing in some Latin-American countries. The main distinctive feature is that the family name is formed by two surnames, the first being the father's first surname and the second the mother's first surname.¹⁰ A second important feature is that married women keep their original surnames and do not adopt the husband's surname. This convention was legally established at the beginning of the nineteenth century but it had been followed by a large part of the population since much earlier.

Even though the telephone directory is possibly the best source available it has some problems: first, there may be duplicity of some surnames as an individual may have several telephone lines. Second, there is a potential bias towards people living in urban areas. Finally, it is difficult to decide which surnames are variant spellings and which are different surnames. Most variant spellings diverged many generations ago and thus we treat different spellings as different surname lines (for instance, Gimenez and Jimenez). Furthermore, the computer file on the CD-ROM records the paternal and maternal surname in a single field and in some cases it is difficult to disentangle the first from the second, particularly when compound names are involved. To overcome this problem, we have programmed an algorithm that separates the first (paternal) and the second (maternal) surname as accurately as possible¹¹ and we use only the first surname in our analysis.

Spain is divided into 52 districts called *provinces*.¹² In this paper we only consider the 47 provinces in the Iberian Peninsula, leaving out the territories of the Canary archipelago, the Balearic Islands, and the two autonomous cities in the North of Africa, Ceuta and Melilla. Even though the telephone directory provides the zip code of each user, which is a finer division than the province, we will take the province as the smallest geographical unit since the additional data needed in our analysis are available only at provincial level.

Provinces are of similar geographical extension (see Figure 1), though the population varies widely, ranging from Madrid with around 5.5 million inhabitants to Soria with just 95,000. Provinces are quite heterogeneous

¹⁰There have been some changes in naming conventions during the past decades. The law now allows for changes in the order of the surnames. This practice, however, is rather unusual and responds to personal motivations, for instance preserving the mother's surname in the next-to-present generation.

¹¹Our algorithm drops 3.1 percent of the total population. This includes foreign residents who have just one surname.

¹²This geographical division was formally set in 1833 and in most cases followed the dominions of Medieval kingdoms, principalities and bishoprics, thus grouping people with a historically common institutional linkage.

in population density and urban concentration, which made us concerned that the telephone directory might be over sampling predominantly urban provinces compared to those with dispersed rural populations. However, for all provinces, the number of individuals included in the telephone directory corresponds to approximately 30% of the population¹³ so we observe no particular bias towards under-sampling in rural provinces.

We found 132,882 different surnames in the whole country. Table 1 shows the fifty most common surnames and their frequencies in the Spanish population.¹⁴ García is the most common surname in Spain (3.57%). Surprisingly, it is also the most common in each of the provinces. In fact, the first 10 names in Table 1 are always among the 20 most frequent names in each province. This indicates a great uniformity among the most frequent surnames across the Spanish provinces in contrast with other European countries, a fact already noted by Rodriguez-Larralde et al. [28].

3.2 Surname Distances

We denote by x_i the relative (national) frequency of surname i and by x the vector of such (national) frequencies. Similarly, we denote by x_i^j the relative frequency of surname i in province j and by x^j the vector of such frequencies in province j . We first construct a *matrix of surname distances* between the provinces, which is denoted by D .¹⁵ The (j, k) element of the D matrix represents the distance between province j and province k and is given by

$$d_{jk} = \left(\sum_{i=1}^N |x_i^j - x_i^k|^p \right)^{\frac{1}{p}} \quad (3)$$

where N is the total number of different surnames in Spain (i.e., $N = 132,882$) and $p \geq 1$. We choose the Manhattan distance,¹⁶ i.e. p equal to 1, because is the only one among the distances defined in (3) satisfying the following desirable "anonymity" property: Consider two provinces A and

¹³The R^2 of the OLS regression is 0.98.

¹⁴We also found that the number of family names of the same size (the size of a family name is number of people bearing this surname) follows a power law distribution with parameter -1.7, which is in accordance with the empirical findings for other countries and the theoretical predictions reported in Derrida, Manrubia and Zanette [9] and [10]

¹⁵This matrix and all the other matrices calculated in the paper, as well as the software programs used for the computation of correlation tests, are available from the authors upon request.

¹⁶We have checked that all the results in this paper are robust to choice of p . Namely, we have repeated all our calculations for $p = 2, 5, 10, 100$ and the limiting case when $p \rightarrow \infty$, obtaining very similar results.

B and assume that one person moves from A to B. If this movement contributes to reduce (increase) the surname distance between the two provinces, the reduction (increase) in the distance is the same no matter the surname borne by the migrant.

Spain has experienced low immigration rates until very recently, therefore, matrix D can be thought to contain aggregate information on the amount of interior migration flows (between provinces) over the last few centuries. Using this matrix we find that the average distance between provinces is 1.22. Notice the maximum value the Manhattan distance can take is 2, therefore, this result indicates that there are substantial differences in the surname distribution across provinces. The largest distance is 1.67 between Lugo and Huesca, two provinces far apart that have had very different histories, and the shortest distance is 0.52 between Seville and Cádiz, which are next to each other and are considered very similar from a sociological point of view. We calculate the “center of gravity”, i.e., the province that minimizes the distance to all the other provinces weighted by the population. We find that the primary center of gravity lies with Madrid, followed by Barcelona, i.e., the two major provinces in terms of population. However, the closest province to Madrid is Toledo and the closest to Barcelona is Tarragona. It is also interesting to calculate the distance between the vector of province frequencies, x^j , and the vector of national frequencies, x . This distance can be interpreted as the distance to the “average” province. Madrid is again the province that shows the shortest distance to the mean and Lugo the longest. Figure 1 shows the mapping of the distances to the mean for the Spanish provinces. Darker provinces are farther from the mean than the lighter ones. Madrid and Barcelona are the closest to the average province while the north-west of the country concentrates those provinces whose surname frequencies are more distant from the country as a whole.

3.3 Consumption Expenditure

In this section, we construct an aggregated consumption vector for each province that will be used to compute a matrix of “preference distances” between provinces. Preferences are unobservable but consumption is not. Individuals choose their consumption profiles by maximizing their preferences subject to their corresponding budget restrictions. One option would be to use standard econometric regression methods to estimate the unknown parameters of the Engel curves, which correspond to the unknown parameters of the individual preferences. However, we think that to pick up vertical transmission of preferences it is very important to use a highly disaggregated classification of goods. Therefore, it would be unfeasible to estimate an Engel

curve for each good considered in our analysis. For this reason, we use the raw vector of consumption shares to define a distance matrix. We are aware that variability in these shares responds not only to differences in preferences across provinces but also to heterogeneity in real prices, income and other factors. We control for these factors in the empirical analysis.

We calculate budget shares for each province using the 1990/91 Consumer Expenditure Survey (Encuesta de Presupuestos Familiares, EPF). The EPFs are large surveys conducted every ten years by the Spanish Statistics Office. These surveys use a representative sample of the Spanish population by provinces, providing very detailed information on household expenditure. They are used by officials to calculate consumption weights in the Consumer Price Index (IPC). In 1997 the EPF was replaced by the Continuous Consumer Expenditure Survey (Encuesta Continua de Presupuestos Familiares, ECPF) which uses a smaller sampling design. Thus, the EPF 1990/91 was the last large survey that was carried out in Spain which included information at provincial level, with a sample size of 21,155 households. We consider one hundred and ten composite goods that correspond to the subclasses defined by the Spanish Statistics Office. The description of the subclasses is listed in the Appendix.

We calculate the *budget share* of good i in province j as

$$w_i^j = \frac{\sum_{h \in \text{province } j} c_{ih}}{\sum_{h \in \text{province } j} c_h} \quad (4)$$

where c_{ih} is the amount expended on good i by household h , and c_h is total expenditure by household h . Let w^j denote the vector of budget shares for province j .

As we did in the case of surname frequencies, we now define the distance in consumption shares between province j and province k as

$$s_{jk} = \sum_{i=1}^{110} |w_i^j - w_i^k| \quad (5)$$

The matrix of preference distances between provinces is denoted by S and contains the distance s_{jk} as the (j, k) element.

Before proceeding, one might claim that the matrix S is not capturing the differences in preferences between populations in the provinces for at least two reasons:

1. Our vector of budget shares, w_i^j , is an aggregated measure and some relevant information may be shaded by the way we aggregate.

2. As already claimed, consumption shares do not depend only on preferences but also on prices, income and other types of difference, such as weather, or the proportion of urban/rural population.

Regarding the aggregation problem, there is not much we can do since aggregating always implies a loss of information. Still, one could claim that the specific way used to compute the vector w_i^j is not satisfactory and averaging over individual budget shares might be more desirable. This alternative, however, was not adopted due to the problem of infrequency of purchases of durable goods. In any case, we compute the average budget shares for each province using only food expenditures¹⁷ to avoid the durable goods problem, and the correlation between the corresponding distance matrix and the S matrix calculated with food goods is 0.966. Hence, both matrices contain essentially the same information and using one or the other would not affect our results.

In principle, the second problem raised could be solved by incorporating the necessary information on those variables. However, this is not an easy task since some of the required information, particularly about prices, is not available so that one has to look for feasible strategies to reduce the severity of the problem. In our case, the following strategy is adopted:

- To reduce the problem of consumers facing different prices in different provinces we drop expenditures on housing from our analysis. Real estate prices vary widely across provinces and represent a very large component of the budget. Thus, we eliminate seven subclasses¹⁸ from the calculations, leaving 103 consumption items.
- We control for differences in income across provinces as well as other factors such as the proportion of urban/rural population and differences in climate and household composition.

Thus, our matrix S contains information on the distances between provinces on the aggregated vector of relative expenditures for 103 different consumption goods.¹⁹ We interpret those distances as representing the consumption preference distances between provinces.

¹⁷Food expenditures are divided into thirty-three subclasses ranging from subclass 110-A to 120-A (see Table 3).

¹⁸Subclasses 310-A to 320-B (see Table 3). Our main results do not depend on the exclusion of these subclasses.

¹⁹Similar to the analysis of surname distances, we compute the center of gravity here. For budget share distances, the center of gravity is Valencia. We also calculate the distance to the average vector of budget shares. Valencia is also the province closest to the mean, whereas Lugo and Jaen are farthest from it.

4 Main Results

4.1 A First Test

Our hypothesis can be formally stated in the following question: Is there a positive correlation between the matrix D and the matrix S ? We will show that the answer is yes, i.e.:

Closer provinces in terms of surnames also tend to be closer in terms of preferences.

Since the elements of a distance matrix are not independent²⁰ we cannot use standard methods of least square estimation. To overcome this problem we use the Mantel test that is specially designed for testing linear correlation between distance matrices. The Mantel test is a non-parametric randomization procedure that can be used to test any linear relationship but is especially useful in the case of non independent data points. The Mantel's test statistic is the correlation coefficient, r , of the distance matrices D and S , and its value range is $[-1, 1]$. The significance of the correlation is evaluated via random permutation of the rows and corresponding columns of D and S . For each random permutation, the correlation r is computed. After a sufficient number of iterations²¹ the distribution of values of r is generated and the critical value of the test at the chosen level of significance is found from this distribution.

We want to test whether there is a statistically significant positive correlation between distance in terms of consumption shares and in terms of surname frequencies. The correlation coefficient between matrices D and S is 0.4198 and the hypothesis of non-positive correlation is strongly rejected on the basis of a Mantel test with 10,000 replications (p-value zero). Table 2 in the Appendix contains this and subsequent results.²²

One might say that this result is not surprising at all and it only detects that provinces that have had much mixing one with another have similar preferences and similar surnames. Though this effect is clear for surnames, it is not obvious that population mixing should lead to more similar preferences as this depends on the attitudes of the newcomers (or locals) with respect to their new environment and their willingness to assimilate the values of the host region. Furthermore, it is even less obvious that provinces with similar preferences should have experienced intense population mixing. It might well be the case that two populations have similar preferences as a

²⁰This is due to the triangle inequality property.

²¹The number of iterations for all the runs in this article was 10,000.

²²All our results are robust to the exclusion of the most common surnames or the very rare ones.

consequence of the spread of certain cultural and social views that do not require population mixing.²³ Before delving into these questions, we should first control for other factors that may have some explanatory power in the correlation found between matrix D and matrix S .

4.2 Controlling for Other Variables

So far, we have considered a simple correlation between surname distances and preference distances. As we mention above, it is clear that this correlation might depend on a number of different facts, which include:

- Geographical distance between provinces
- How urban/rural provinces are
- Income differences
- Climate
- Household composition

The geographical distance between provinces could be very closely correlated with other variables, for example the type of local agricultural produce available, that might explain some of the differences in consumption. Thus, we will define the matrix of geographical distances between provinces, G , where the element g_{jk} indicates the distance, in kilometers, between the capital of province j and the capital of province k . Since provinces are relatively small in area and in most cases the majority of the population is concentrated around the provincial capital, this distance is a good index of the geographical distance between the whole populations in the provinces.

It is natural to assume that consumption shares in urban environments might differ from consumption shares in rural areas, even if household preferences were identical, as urban and rural households face different prices and the set of available goods may be different. In order to control for this, we classify municipalities into eight groups²⁴ and assign each household in the EPF sample to the corresponding group. We denote by u_i^j the percentage of

²³This is somewhat similar to an old question in Population Genetics about the demic versus the cultural transmission of technological changes. See Cavalli-Sforza et al. [6] and Jobling et al [19]

²⁴Groups are defined in terms of population intervals: 2,000 or less, from 2,000 to 5,000, from 5,000 to 10,000, from 10,000 to 20,000, from 20,000 to 50,000, from 50,000 to 100,000, from 100,000 to 500,000 and more than 500,000.

households in group-size i in province j . Then the “urban” distance between province j and province k is given by

$$u_{jk} = \sum_{i=1}^8 |u_i^j - u_i^k|$$

and the corresponding matrix of distances is denoted by U .

The EPF contains information on the total income reported by household members. This information could be used to control for the income effect. However, we suspect that the income variable could be under-reported. For this reason we decided to use total expenditure as it is highly correlated with household income and it does not suffer from significant measurement error. Thus, denoting by m^j the mean household total expenditures in province j , the “income” distance between province j and province k is given by $m_{jk} = |m^j - m^k|$ and the corresponding matrix is denoted by M .

The climate in Spain varies greatly across regions. In general, the North is cold and rainy and the South warm and dry. Because of these different climates, people with the same preferences might need to consume different goods depending on their province of residence. To control for this a matrix of “climate distances”, T , is defined in the following way: for each province j we compute the vector $t^j = (t_1^j, \dots, t_{12}^j)$, where the t_i^j element indicates the average temperature ²⁵ during month i . The climate distance between province j and k is given by

$$t_{jk} = \sum_{i=1}^{12} |t_i^j - t_i^k|$$

and the matrix T is formed with these elements.

Finally, household composition may be a major explanatory factor in determining differences in consumption patterns. Rural provinces contain a higher proportion of elderly, retired people, whose consumption pattern differs substantially from that of a middle-aged family with children, for instance. The EPF also contains information on household composition in the form of a categorical variable for the fourteen different types of households. In Table 4 we describe the different type of household defined in the EPF. We compute the vector h^j for province j containing the proportion of households of each type. Thus, matrix H denotes the distances in terms of household composition.

²⁵These temperatures are averages for 1997-2002. These data are available from the website of the Spanish Statistics Office, <http://www.ine.es/>.

Summing up, we have the matrices S , D , G , U , M , T and H , the distance matrices of, respectively, consumption shares, surnames, geographical distances, proportion of urban/rural population, income per capita, climate and household composition. The question now is how to extend the bivariate Mantel test to our context of multiple control variables. Smouse et al. [30] propose the following three-step technique:

- OLS estimation of D on G , U , M , T and H
- OLS estimation of S on G , U , M , T and H
- Bivariate Mantel test using the residuals of the previous two regressions.

Therefore, we perform a multivariate Mantel test to determine the significance of the correlation coefficient of the D and S matrices, controlling for G , U , M , T and H . The correlation is now 0.2277 and is significantly greater than zero as the p-value is 0.0042. Thus, after controlling for how close provinces are in income, urban/rural environment, geographical distance, climate and household composition, we still find that provinces that are similar in the frequencies of surnames tend to be similar in their consumption preferences.

As explained above, this result might be seen as evidence of a significant vertical transmission effect on consumption behavior. Individuals migrate from one place to another and when doing so they bring their surnames with them, which will be bequeathed unaltered to their descendants. They also bring their own tastes and preferences, which are inherited by their offspring. Contrary to the case of surnames, the (vertical) transmission of preferences to descendants is far from exact, as it is affected by learning (cultural assimilation or horizontal transmission) and mixed marriages, and these factors are probably also affected in very complex ways by the environment, genes and individual specific random effects.

4.3 Different Groups of Consumption Goods.

We now look at differences in the correlation depending on the type of consumption good considered. Vertical transmission might play a more important role in, for example, the formation of preferences over food than in the formation of other types of preference. Or, equivalently, for some types of consumption good such as food, the offspring of immigrants can be less prompted to adopt the habits of the host province than for other consumption goods. This new exercise will also help to analyze whether the positive correlation between surname and consumption distances is due to “ghetto

grouping". If immigrants tend to group in closed ghettos, newcomers will be living in exactly the same environment as in their provinces of origin. In this case, the preferences of the parent would coincide with those of the environment, and therefore, we should find a large correlation coefficient for all consumption goods both under vertical and horizontal transmission.

We create two new distance matrices in consumption shares. The first matrix, S_f , includes exclusively all food goods, which are the first thirty-three items on the list used to compute S .²⁶ The second matrix, S_{nf} , contains the remaining seventy items. Then we repeat our previous multivariate Mantel test twice, firstly replacing matrix S by matrix S_f and secondly replacing S by S_{nf} . The results of these two tests indicate a striking difference between food and non-food cases. The test for food items (matrix S_f) shows a significant correlation coefficient of 0.3735 (p-value 0). The correlation coefficient when the matrix with non-food items, S_{nf} , is used is 0.0465, which is non-significant (p-value 0.3190).

The non-significance of this coefficient indicates that newcomers assimilate the preferences of locals in non-food consumption. This fact is difficult to interpret under the assumption of "ghetto grouping" and no vertical transmission. Such behavior is more consistent with the following proposition:

*There is horizontal transmission on preferences over the non-food consumption goods, and therefore perfect integration of immigrants to their host environment. Regarding preferences over food items the result suggests, but cannot unambiguously conclude, the existence of a strong vertical transmission.*²⁷

This possible vertical transmission for food preferences is related to the results provided in Waldkirch et al [34]. It is important to notice that these authors provide a result on possible vertical transmission for the total amount spent on food. Our analysis, however, is much more disaggregated and focuses on the expenditure shares of thirty-three different food items. Moreover, after controlling for the State of residence of parents and their offspring they obtain no correlation in the total amount spent on food. Interestingly, we have checked and found that a similar result holds in our case, i.e. after controlling for income and geographical distance the total amount spent on food is not related to surname distances. Thus, our findings also suggest that vertical transmission of preferences affects the shares of the different food products rather than the total amount spent on food.

²⁶Subclasses 110-A to 120-A in Table 3.

²⁷This ambiguity derives from the discussion in Section 2.2 where we have shown that a positive correlation between surnames and consumption distances is a necessary but not sufficient condition for vertical transmission.

Three additional comments are called for here. First, it may be that our test does not detect vertical transmission for non-food goods because the classification of these goods is less disaggregated than the classification of food goods. For example, it might be the case that children have the same preferences as parents for going to the theater instead of to the cinema. However, both theater and cinema expenditure belong to the same category and therefore the transmission of this sort of preference cannot be detected with our data. However, we have repeated all the previous tests using the maximum level of disaggregation available in the EPF and the results remain unchanged.²⁸ Second, our result could be due to the fact that preferences for non-food goods are basically the same across provinces. Our data, however, do not support this view since the non-food consumption patterns are different across the Spanish provinces even after controlling for income, geographical distance, household composition and climate (the unexplained variability of the regression is 79.2%). Finally, following the approach in Stigler and Becker[33] an alternative explanation of our result for food goods is that we are not detecting transmission of preferences but *transmission of skills* in the household production of food related goods. Cooking is a production process that is learnt within the family environment and children and young adults might easily acquire their parents’ cooking skills or “human capital”. Since parents and their offspring share similar cooking skills it is natural to expect them to buy similar “cooking inputs” as well, i.e. similar food, and this fact is reflected in a similar composition of their food consumption shares. We cannot discriminate between this “skill transmission” hypothesis and our “preference transmission” hypothesis but in our view this fact does not detract any merit from our finding.

4.4 Controlling for Recent Migration

One might claim that immigrants’ adaptation to their host environment takes place during the second generation (see Borja [3]). In this case, since there were some significant migration flows in Spain in the 60’s and 70’s, our previous conclusion relating a significant correlation coefficient with a vertical transmission effect might be misleading. To see this, suppose that the first generation of adult immigrants keep their original preferences. Their offspring, however, adopt the preferences of the host province, i.e. horizontal transmission is the only factor in the preference formation process. However, the matrix of surname distances and the matrix of consumption shares could

²⁸The total number of goods at this level of disaggregation is 769. The results using these data are available from the authors upon request.

be correlated because of the recent migration, leading us wrongly to deduce that there is vertical transmission.

To control for this possibility we should disentangle the contribution of recent migration to the consumption and surname vectors and this requires information on the net migration flows between Spanish provinces during the last generation. Unfortunately, this information is not available. However, the Spanish Labor Force Survey (EPA) conducted by the Spanish Statistics Office contains information on the current province of residence and the birthplace of individuals. The EPA is a quarterly survey and the sample size is around 190,000 individuals. Households are interviewed for six consecutive quarters. In this study we merge two waves of the EPA: First quarter 1999 and third quarter 2000.²⁹ Thus, using the data from the EPA, we construct a matrix, E , of “migration” distances in the following way: we associate the vector $b^j = (b_1^j, \dots, b_{47}^j)$ to province j such that the b_i^j element contains the percentage of people living in province j who were born in province i . We construct the matrix E in a similar way to the previous matrices. The element e_{jk} is the distance between provinces j and k , i.e. $e_{jk} = \sum_{i=1}^{47} |b_i^j - b_i^k|$. Since geographical mobility over the last 15 years in Spain has been very low we are confident that a large majority of the people born in province i and currently living in province j did not live in other provinces before migrating to j . Thus, our matrix E may be seen as a good approximation to the matrix of recent migration flows between Spanish provinces.

We repeat our previous multivariate Mantel test now including the additional matrix E . More precisely, we test the significance of the correlation coefficient between matrix S and matrix D , controlling for G , U , M , T , H and E . We obtain that the correlation coefficient is 0.1669, lower than before but still large and significant (p-value 0.0359). We also carry out the test when consumption goods are divided into food (matrix S_f) and non-food groups (matrix S_{nf}) as in the previous subsection. It is reassuring to observe that the correlation coefficient when only food is considered is still high (0.3154) and significant (p-value 0.0002), and the coefficient for the non-food case remains low (0.0180) and clearly non-significant (p-value 0.4191), indicating that for non-food goods vertical transmission plays no role in the preference formation process.³⁰

²⁹We use these two waves to avoid duplicity of households.

³⁰One might wonder whether our result is driven by “outlier provinces”. We have checked the robustness of our test by dropping three provinces. Thus, we run our test for all possible subsets of provinces (16215) leaving out one triplet in each run. When we consider all goods we find a non-significant correlation in about one third of the runs. In the case of food consumption the correlation coefficient in all the runs is positive and significant at the 5% level. For non-food items the correlation coefficient is not significantly

5 Final Comments and Further Work

We have developed a novel indirect approach to analyze the existence of inter-generational transmission of preferences on consumption. The main finding indicates the existence of horizontal transmission of preferences for non-food goods and a possible vertical transmission for food items. A number of issues are left for further research:

i) Our approach is not able to distinguish paternal from maternal transmission and it is important to find a way to do this. Moreover, since our matrix of surname distances is calculated using the first surname, our results may be seen as applying uniquely to paternal vertical transmission. A possible way to overcome this problem may be the use of genetic distances between provinces instead of the surname distances used here. Scientists working on Human Evolutionary Genetics have developed matrices of genetic distances for many countries and world regions (see Cavalli-Sforza et al [6] for a path-breaking work on this field, and Jobling et al [19] for a recent survey). Unlike surnames, genes come from both the father and the mother, and by using genetic distances we could better incorporate maternal influence on the transmission of preferences. Moreover, the genes on the non-recombinant part of the Y-chromosome are exclusively transmitted through the paternal line³¹ and mitochondrial DNA through the maternal line. In principle, the information on their geographical distributions may be used to tell apart maternal and paternal vertical transmission. However, genetic information at provincial level in Spain is still very scarce and, except for the genes associated with blood type, the sample sizes are clearly insufficient for our task. Since new data is being obtained quickly, the situation is rapidly changing and we hope to be able to undertake this task in the near future.

ii) Information on geographical distribution of surnames is becoming easily available for many other countries where there are also good surveys of consumption behavior. Applying our methodology to a second country would be an important exercise in investigating the robustness of our results.

References

- [1] Becker, G.S. and C.B. Mulligan (1997). “The Endogenous Determination of Time Preference,” *Quarterly Journal of Economics*, **112**, 729-

greater than zero at the 5% level in any of the runs. Hence, we are confident that our results are not just due to a small number of special provinces.

³¹Provided certain assumptions are met, surnames and genes in the Y-chromosome should be correlated, see Jobling[20]

758.

- [2] Bisin, A. and T. Verdier (2001) "The Economics of Cultural Transmission and the Dynamics of Preferences" *Journal of Economic Theory*, **97**, 298-319.
- [3] Borjas, G. (1994) "Long-run Convergence of Ethnic Skill Differentials: the Children and Grandchildren of the Great Migrations" , *Industrial & Labor Relations Review*, **47**, 553-573.
- [4] Browning, Martin and M. Dolores Collado (2001), "The response of expenditures to anticipated income changes: panel data estimates", *American Economic Review*, **91**, 681-92.
- [5] Cavalli-Sforza, L.L. and M.W. Feldman (1981), *Cultural Transmission and Evolution: A Quantitative Approach*, Princeton: Princeton University Press.
- [6] Cavalli-Sforza, L.L., Menozzi P. and A. Piazza (1994) *The History and Geography of Humans Genes*. Princeton: Princeton University Press.
- [7] Charles, K.K. and E. Hurst (2003). "The Correlation of Wealth across Generations." *Journal of Political Economy*, **111**(6):1155-82.
- [8] Daniels. M., Devlin, B. and K. Roeder (1977). "Of Genes and IQ" in B. Devlin, Fienberg, S., Resnick, D. and K. Roeder (eds.) *Intelligence, Genes, and Success*. New York: Springer-Verlag, 45-70.
- [9] Derrida, B., Manrubia, S.C. and D.H. Zanette (2000) "On the Genealogy of a Population of Biparental Individuals" *Journal of Theoretical Biology*, **203**, 303-315
- [10] Derrida, B., Manrubia, S.C. and D.H. Zanette (1999) "Statistical Properties of Genealogical Trees" *Physical Review Letters*, **82**, 1987-1990.
- [11] Desmet, K., Le Breton, M., Ortuno-Ortin, I. and S. Weber (2006) "Nation Formation and Genetic Diversity", manuscript, Universidad de Alicante.
- [12] Einav, L. and L. Yariv (2006) "What's in a Surname? The Effects of Surname Initials on Academic Success", *Journal of Economic Perspectives*, **20**, 175-188.

- [13] Feldman, M.W., Otto, S. P. and F.B. Christiansen (2000) “Genes, Culture, and Inequality.” in K. Arrow, S. Bowles and S. Durlauf (eds.) *Meritocracy and Economic Inequality*. New Jersey: Princeton University Press. 61-85.
- [14] Fryer, R.G., Jr. and S. D. Levitt (2004) “The Causes and Consequences of Distinctively Black Names.” *Quarterly Journal of Economics*, **119**(3), 767-805.
- [15] Goldin, C. and M. Shim (2004), “Making a name: Women’s Surnames at Marriage and Beyond.” *Journal of Economic Perspectives*, **18**(2), 143:160.
- [16] Harris, J.R. (1998) *The Nurture Assumption: Why Children Turn Out the Way They Do*. The Free Press.
- [17] 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**, 368-372.
- [18] Jennings, M.K., L. Stoker and J. Bowers (2001). *Politics Across Generations: Family Transmission Reexamined*. Institute of Governmental Studies, University of California, Berkeley.
- [19] Jobling, M.A., Hurles, M.E. and C. Tyler-Smith (2004). *Human Evolutionary Genetics. Origins, People & Disease*. New York: Garland Science.
- [20] Jobling, M.A. (2001) “In the Name of the Father: Surnames and Genetics.” *Trends in Genetics*, **17**(6): 353-57.
- [21] Lasker, G.W. (1985) *Surnames and Genetic Structure*, Cambridge University Press.
- [22] Legendre P. and L. Legendre (1998). *Numerical Ecology*, Elsevier, New York.
- [23] Mantel N. (1967) “The Detection of Disease Clustering and a Generalized Approach,” *Cancer Research*, **27**, 209-220.
- [24] Mulligan, C. (1997). *Parental Priorities and Economic Inequality*. Chicago: University of Chicago Press.
- [25] Mulligan, C. (1999) “Galton vs. Human Capital Approaches to inheritance” *Journal of Political Economy*, **107**(6), 184-224.

- [26] OECD (2002), *Trends in International Migration*, Publications Service, Paris.
- [27] Pollak, R.A. (1970) "Habit Formation and Dynamic Demand Function" *Journal of Political Economy*, **78**, 745-63.
- [28] Rodriguez-Larralde, A., A. Gonzales-Marin, C. Scapoli, and I. Barrai (2003) "The Names of Spain: A Study of the Isonymy Structure of Spain," *American Journal of Physical Anthropology*, **121**, 280-292.
- [29] Rowe, D. (1994) *The Limits of Family Influence: Genes, Experience and Behavior*. New York: The Guilford Press.
- [30] Smouse P.E., Long J.C. and R.R. Sokal (1986) "Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence." *Syst. Zool.*, **35**, 627-632.
- [31] Sokal, R.R. and F.J. Rohlf (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*. W.H. Freeman and Company, New York.
- [32] Solon, Gary (1992) "Intergenerational Income Mobility in the United States." *American Economic Review*, **82**(3),393-408.
- [33] Stigler, G.J. and G.S. Becker (1977) "De Gustibus non Est Disputandum" *American Economic Review*, **67**(2), 76-90.
- [34] Waldkirch, A., Ng, S. and D. Cox (2004), "Intergenerational Linkages in Consumption Behavior," *The Journal of Human Resources*, **39**(2), 355-381.

Appendix: Tables

Table 1. The 50 most common surnames in Spain

Rank	Name	Frequency	Rank	Name	Frequency
1	García	3.57	26	Serrano	0.32
2	Fernández	2.20	27	Ramos	0.32
3	González	2.18	28	Blanco	0.31
4	Rodríguez	2.09	29	Sanz	0.28
5	López	2.08	30	Ortega	0.27
6	Martínez	2.03	31	Suárez	0.26
7	Sánchez	1.90	32	Molina	0.26
8	Pérez	1.85	33	Rubio	0.26
9	Martín	1.20	34	Ramírez	0.26
10	Gómez	1.15	35	Delgado	0.25
11	Ruíz	0.86	36	Morales	0.25
12	Hernández	0.79	37	Castro	0.25
13	Jiménez	0.79	38	Ortiz	0.25
14	Díaz	0.77	39	Marín	0.24
15	Álvarez	0.70	40	Iglesias	0.23
16	Moreno	0.69	41	Garrido	0.21
17	Muñoz	0.64	42	Núñez	0.20
18	Alonso	0.52	43	Santos	0.19
19	Romero	0.46	44	Calvo	0.19
20	Gutiérrez	0.45	45	Prieto	0.19
21	Navarro	0.42	46	Vidal	0.18
22	Torres	0.36	47	Lozano	0.18
23	Gil	0.35	48	Díez	0.18
24	Domínguez	0.35	49	Cano	0.18
25	Vázquez	0.35	50	Castillo	0.17

Table 2. Mantel Test Statistics.

Control vars.	All EPF	Food	Non Food
None	0.4198 (0.0000)	0.5391 (0.0000)	0.1991 (0.0089)
Urban, Income, Climate, Km, Household Composition	0.2277 (0.0042)	0.3735 (0.0000)	0.0465 (0.3190)
Urban, Income, Climate, Kms, Household Composition, Recent immigration	0.1669 (0.0359)	0.3154 (0.0002)	0.0180 (0.4191)

Right tail p-values in parenthesis

Table 3. Subclasses of consumption goods in the EPF survey.

Class	Subclass
110	A. Rice B. Flour and lightly processed cereals C. Bread D. Pastry-cooked products E. Pasta products and other cereal based products
111	A. Beef B. Veal C. Pork D. Sheep meat E. Poultry F. Cooked pork G. Canned and processed meat H. Other meats and meat offal
112	A. Fresh and frozen fish B. Dried, smoked, canned and processed fish C. Fresh and frozen crustaceans and molluscs
113	A. Liquid milk B. Preserved milk C. Cheese and other dairy products D. Eggs
114	A. Butter and margarine B. Edible oils
115	A. Fresh fruit B. Nuts and raisins, olives, canned fruit and fruit juices C. Fresh vegetables D. Dried vegetables E. Frozen, preserved and canned vegetables
116	A. Potatoes and their by-products
117	A. Sugar
118	A. Coffee, cocoa, infusions and substitutes
119	A. Chocolate and chocolate substitutes B. Other food products
120	A. Non-alcoholic beverages
130	A. Spirits B. Wine C. Beer D. Other alcoholic beverages
140	A. Tobacco
210	A. Men's clothes B. Men's underwear C. Women's clothes D. Women's underwear E. Children's clothes F. Children's and babies' underwear G. Clothing complements and repairs
220	A. Men's footwear B. Women's footwear C. Children's and babies' footwear
221	A. Footwear repair
310	A. Housing rentals B. Expenses related to property C. Repair and maintenance of rented housing D. Repair and maintenance of property
311	A. Water supply
320	A. Electricity and gas B. Heating fuels

Table 3. *(Continued)*

Class	Subclass
410	A. Furniture for kitchen and bathroom B. Other furniture and decorative ornaments for the household C. Floor coverings and repairs
420	A. Household textiles B. Other furniture goods and repairs
430	A. Refrigerators, washing machines, dishwashers and irons B. Cookers C. Heating appliances D. Other electrical appliances and repairs
440	A. Glassware, tableware, cutlery and their repairs B. Other kitchen and household equipment and their repairs
450	A. Goods for cleaning and routine household maintenance B. Other non-durable household goods
451	A. Household services, except domestic service
460	A. Domestic service
510	A. Medical products B. Other pharmaceutical products
520	A. Therapeutic appliances and equipment and repairs
530	A. Medical, nursing and other out-patient services
540	A. Hospital services and similar
550	A. Health insurance
610	A. Motor cars B. Other vehicles
620	A. Tyres, spare parts and repairs
621	A. Fuels and lubricants
622	A. Other goods related to personal transport
630	A. Local transport
631	A. Long-distance transport
640	A. Postal services and communications
710	A. Radio and television equipment and repairs B. Other audio-visual equipment
711	A. Photographic equipment, computers and others
712	A. Equipment for sport B. Games and toys C. Other recreational goods
720	A. Cinema, theatre, football and others performances
721	A. Recreational services
730	A. Books newspapers and magazines
740	A. Pre-primary and primary education B. Secondary education C. Expenses related to education D. Secondary education E. Education not definable by level

Table 3. *(Continued)*

Class	Subclass
810	A. Personal care services
811	A. Durable personal care goods B. Non-durable personal care goods
820	A. Jewellery, imitation jewellery and their repairs
821	A. Other personal effects
822	A. Stationery materials
830	A. Restaurants, bars and cafes
831	A. Hotels and other accommodations
840	A. Tourist services
850	A. Financial services
860	A. Other services

Table 4. Type of household considered

Column	Type of household
1	One adult aged 65 or older without children
2	One adult younger than 65 without children
3	One adult with one or more children
4	Couple without children (head aged 65 or older)
5	Couple without children (head younger than 65)
6	Couple with one child
7	Couple with two children
8	Other households with two adults (without children)
9	Other households with two adults
10	Three adults (without children)
11	Three adults (with children)
12	Four adults or more (without children)
13	Four adults or more (with children)

Figure 1: Map of the provinces of mainland Spain and distances to the mean distribution of surnames. The lighter the province the closer to the national mean.

