Using Spanish Surname Ratios to Estimate Proportion Hispanic in California Cities via Bayes Theorem*

Bernard Grofman, *University of California, Irvine*
Jennifer Garcia, *University of California, Irvine*

Objectives. To generate, via application of Bayes Theorem, accurate estimates about the size of Hispanic populations in California cities from very limited data on the surnames of those living in the cities. *Methods.* We make use here of the ratio of those with the name "GARCIA" to those with the name "ANDERSON" in those cities, one of which is far more likely to be Hispanic and one of which is far more likely to be non-Hispanic. *Results.* For four cities that vary dramatically in their Hispanic populations, using only two common names we are able to estimate the Hispanic population in the cities. *Conclusions.* We lay the background for our surprising results by underscoring common fallacies in using surnames for purposes of ethnic identification, such as the belief that the proportion of bearers of a given name who are Hispanic can be specified as a unique percentage. We show that how "Hispanic" any given name will turn out to be is a function of the overall demography of the subpopulation being analyzed, which will also affect the distribution of names within that subpopulation.

There are a number of situations where we do not have reliable information about a group's proportion of a given population, but wish to estimate that proportion. Let us assume that surnames held by members of the group, say Hispanics,[1] are relatively distinctive. Let us further imagine that we have the names of those in the population whose group composition we wish to estimate, for example, hospital patients, or registered voters, or purchasers of some particular commodity. In principle, by matching surname with estimated ethnicity, we may derive estimates of group population shares in the given list of names.

The main theoretical result in this article is the exposition of a methodology that uses very limited data on surnames—indeed, data from only two surnames, "ANDERSON"

*Direct correspondence to Bernard Grofman, Center for the Study of Democracy, University of California, Irvine, CA 92697 ⟨bgtravel@uci.edu⟩. Bernard Grofman shall share all data and coding for replication purposes. We are indebted to Charles Hammond of the U.S. Bureau of the Census for making available to us in EXCEL format the Census-based list of common surnames showing the proportion of self-identified Hispanics for each name. This research was supported by the Jack W. Peltason Endowed Chair at the University of California, Irvine and by the UCI Center for the Study of Democracy. Earlier work on surname matching by the first-named author was done under contract from the U.S. Department of Justice, Civil Rights Division, Voting Rights Section, in the case of *Garza v. County of Los Angeles Board of Supervisors*, 918 F. 2d 763 (9th Cir. 1990), in conjunction with the demographer William O'Hare and with the assistance of Robert Kengle of the DOJ; and, under contract from the Government Accountability Board of Wisconsin, in *Baldus et al. v. Government Accountability Board of Wisconsin*, Federal District Court, Case No. 11-CV-562 JPS-DPW-RMD, decided March 22, 2012. Opinions and analysis reflected in this article are the authors' own and do not reflect the views of either the U.S. Department of Justice or the Government Accountability Board of Wisconsin.

[1]We will use the term "Hispanic" interchangeably with "Latino," and interchangeably with "Spanish origin," in accord with the question currently asked on the Census form.

SOCIAL SCIENCE QUARTERLY, Volume 96, Number 5, November 2015

DOI: 10.1111/ssqu.12214

and "GARCIA"—to derive estimates of Hispanic population. The main empirical result in this article is an application of this methodology to California cities that vary dramatically in their actual Hispanic population, from under 10 percent to well over 90 percent. Our findings demonstrate the rather remarkable accuracy of this simple methodology.

In order to lay the foundation for our surprising empirical results, we first illustrate common pitfalls in using names that are drawn from a list of most common Hispanic surnames to estimate the proportion of Hispanics in some subpopulation. In the next section, using data taken from the 2010 Census that involves matching surnames to whether or not individuals with those names self-identify as of Spanish heritage, we show the power of Bayes Theorem and analyses based on it.[2] In a mathematical appendix that is available online at the website of the senior author,[3] we provide a full derivation of the mathematical intuitions we summarize.

While we only examine issues related to Spanish surname matching in the United States,[4] essentially identical issues arise with respect to name matching for other ethnicities, for example, Asian Americans (Abrahamse, Morrison, and Bolton, 1994). Moreover, the methods (and cautionary notes) we provide are general ones that are applicable to any type of name matching—and not just in the United States.[5] Indeed, they are applicable to many other types of situations distinct from surname matching where there is the need to balance Type I errors (false positives) and Type II errors (false negatives) by taking into account baseline probabilities (see below).

Surname List

The U.S. Census has provided a way to estimate the link between name and Hispanic identity by matching surnames to the proportion of those who self-identified as Spanish origin on the Census. Based on the 2010 Census, the Census Bureau has created a U.S.-Census-based list of common surnames (names with greater than 300 instances) that also

[2]The present article builds on earlier work by the authors (Grofman and Garcia, 2014) introducing a Bayesian approach to surname matching in the context of voting rights litigation. However, that essay does not show the power of the ratio approach that is the heart of our estimates of the Hispanic population of California cities. Bayesian ideas are briefly mentioned in some Census publications without a specified model or empirical analyses, and some expert witnesses in voting rights litigation in the 1990s considered a Bayesian approach to Spanish surname analysis, but dropped it after an appellate court decision in *Garza v. Los Angeles County Board of Supervisors* in 1990 because it appeared that federal courts had accepted the validity of simply using the 12,000+ Census Spanish surname list (personal communication, Kenneth McCue, October, 2012). Elliot et al. (2008, 2009) offer work that is closely related to our own. They consider a number of different ways of pooling information across cases to improve estimates of racial/ethnic populations, including data from more than one point in time, information on surnames, and information on demographic attributes of neighborhoods. However, none of the statistical models they consider take full advantage of a Bayesian framework.

[3]See ⟨http://www.socsci.uci.edu/~bgrofman/⟩.

[4]Word and Perkins (1996:3–4) identify a number of different areas where Spanish surname matching methods of one type or another have been used, including studies of births and deaths, hospitalization studies, retrospective estimates of the Hispanic population among Social Security recipients, analysis of immigration data, customer data for firms of various types, the creation of customized mailing lists for marketing to the Hispanic community, and methods for imputations of Hispanic identity where data are missing from Census forms. Although Word and Perkins (1996) do not mention this application, one arena in which Spanish surname matching is important in the United States is in voting rights litigation involving issues of vote dilution (see Grofman and Garcia, 2014, for a review of relevant litigation).

[5]For example, Bhavnani (2012) has used official records of election commissions in India to examine the effect of name and caste on voting behavior. Similarly, Harris (2012) uses data on the surnames common in various ethnic groups to identify the changing ethnic distribution of political appointments in Kenya from 1963 to 2010. Indeed, Harris (2012:1) identifies works from numerous fields, including economics, history, marketing, population biology, and public health, where names have been taken to be markers of ethnicity.

provides the proportion of self-identified Hispanics for each name for the country as a whole. This list is a public document and can be downloaded as a data file. Similar lists were generated by the Census Bureau for earlier periods (Word and Perkins, 1996:1), and other surname match-up lists have also been created by various entities,[6] though virtually all surname matching done in situations where legal issues are involved draw in some fashion on the list prepared by the Census.

In practice, a surname list is usually used to generate a much smaller (and more manageable) list of only the surnames that are found to have high/highest proportions (or, in some applications, numbers) of Hispanics, treating anyone with a name on the list as Hispanic and anyone whose name does not fall on the list as non-Hispanic. The 2010 Census Bureau list, which is far and away the most comprehensive to date, includes over 50,000 common names and covers over 220 million people who have answered whether or not they are of Spanish origin. It is this data set that we will draw upon in our empirical work.

However, even if one uses a Census surname list, such as those created after the 1990 and 2000 Censuses, individual investigators have varied greatly in how they generated the list of names they would classify as "Hispanic." We have identified applications of surname matching using fewer than 700 names to ones with over 12,000 names being used.[7] While the Census provides information on surname ethnicity characteristics, and although Census staff have identified smaller subsets of various sizes of "heavily Hispanic" names, the Census does not offer "best practices" advice on how to make use of these data. Indeed, there is no developed theory to justify using any particular cut-off point as to how many names should be treated as Hispanic in some given application of surname matching technology.[8]

It might seem obviously preferable to simply take the estimated proportion Hispanic of each name as input and calculate a weighted average of the Hispanic proportions of all the names in a database, weighting by name frequency. The reason this is not done is because of the difficulties of doing the matching when there are tens of thousands of names to be compared against the names in the data set. In the redistricting arena, for example, from the 1980 redistricting round to the 2010 redistricting round, *every* application of Spanish surname matching of which the faculty co-author of this article is aware involved treating one set of names as if they were 100 percent Hispanic and all other names as if they were 0 percent Hispanic.

Issues in Using Surname Lists

There are four fundamental and closely related problems in using a specified list of heavily Hispanic surnames to identify the Hispanic proportion in a population. Here, we introduce these errors in intuitive terms and with hypothetical examples. In the next section, we illustrate them with actual 2010 Census data.

[6]See, for example, ⟨http://www.family-crests.com/family-crest-coat-of-arms/surnames-7-7/common-spanish-surnames.html⟩. This is a list of 660 names.

[7]For example, Barreto, Segura, and Woods (2004) draw from Word and Perkins (1996) a list with over 8,000 names, while an expert witness for the plaintiffs, in his testimony in 2012, in *Baldus v. Wisconsin Government Accountability Board*, used that same source, but made use of only 639 names.

[8]For example, Word and Perkins (1996:14) observe: "In theory, we are not providing a Spanish surname 'list'. Rather, we provide auxiliary data for each surname that can be sorted into a continuum allowing the prospective user to determine his or her own criteria as to what is or is not a Spanish surname." This note of caution is simply not very helpful unless we appreciate how the link between surname and ethnicity depends upon demographic context, as is done below.

To describe those errors we need some simple notation. Let $p(H)$ be the probability of being Hispanic, that is, the (expected) proportion of Hispanics in some given population. Let $p(N)$ be the probability of having a given name, that is, the (expected) proportion of those with that name in the same population. When there are two events or conditions or outcomes, such as H and N, then we can look at the probability of each separately, that is, $p(H)$ and $p(N)$. Or, we can look at the conditional probabilities, $p(H|N)$ and $p(N|H)$, that is, the likelihood that someone is Hispanic, given that his or her name is N, on the one hand, and the likelihood that we will observe someone has the name N given that s/he is Hispanic, on the other.[9]

ERROR 1: Treating $p(H|N)$ as if it were a constant.

It seems obvious that the probability of being Hispanic is not the same in different populations. It seems equally obvious that the probability of having some particular name is not the same in different populations. However, what is less obvious is that the probability that someone with a given name is Hispanic will vary with the population under investigation. In fact, however, there is no such thing as *the* proportion of bearers of a given name who are Hispanic. How Hispanic any given name is a function of the overall *Hispanicity* (i.e., Hispanic proportion) of the population, which affects both the conditional probability that the possessor of any given name will be Hispanic and also the distribution of names. A simple illustration can make these points clear.

Imagine that, in the population as a whole, 90 percent of those with name "GARCIA" are Hispanic. What is the proportion of "GARCIAs" who are Hispanic in a subpopulation that is 100 percent Hispanic? Well, a moment's reflection reveals that the answer has to be 100 percent, not 90 percent. Similarly, if we ask what is the proportion of "GARCIAs" who are Hispanic in a subpopulation that is 0 percent Hispanic, it is obvious that the answer has to be 0 percent. So clearly $p(H|\text{GARCIA})$ is not a constant; it varies with demographic context. In neighborhoods whose Hispanic population is between 0 percent and 100 percent, $p(H|\text{GARCIA})$ will take on intermediate values.[10] But it is easy to forget that reality and act as if the fact that, overall, in some population, 90 percent of the "GARCIAs" are Hispanic means that in *any* subpopulation 90 percent of the "GARCIAs" are Hispanic, and thus "GARCIA" is always to be classified as a Hispanic name. Of course, even in a neighborhood with few Hispanics, a higher proportion of those with the name "GARCIA" will be Hispanic than, say, those with the name "ANDERSON."

Now imagine further that of those who are Hispanic, 20 percent have the name "GARCIA," while of those who are not Hispanic, only 1 percent have the name "GARCIA." Note that these are different percentages than those reported in the paragraph above because now we are looking at $p(N|H)$ rather than $p(H|N)$. Note also that, while $p(H|N) = 1 - p(\text{non-}H|N)$; $p(N|H) \neq 1 - p(N|\text{non-}H)$, that is, in this example 20 percent is not the same as 99 percent (= 100 percent − 1 percent). How common a name is "GARCIA" in different subpopulations? Well, obviously it depends upon the demographic composition of the subpopulation. In an all-Hispanic neighborhood, with the parameters as given above, we might expect that 20 percent of the population is named "GARCIA." In a completely non-Hispanic neighborhood, we might expect that only 1 percent of the population will have that name. In a mixed neighborhood that is 50 percent Hispanic and 50 percent

[9]It should be obvious that, in general, $p(H|N) \neq p(N|H)$, but it is easy to confuse these two conditional probabilities. If H and N are mutually exclusive, however, then $p(H|N) = 0 = p(N|H)$; if the two conditions are statistically independent of one another, than we have $p(H|N) = p(H)$ and $p(N|H) = p(N)$.

[10]For the actual 2010 Census data we review in the next section, even when the population is only 10 percent Hispanic, more than 80 percent of all "GARCIAs" will be Hispanic. In a population that is around two-thirds Hispanic, more than 99 percent of all "GARCIAs" will be Hispanic.

non-Hispanic, we might expect that 10.5 percent will have the name "GARCIA." And so on.

ERROR 2: Focusing on Type I error and neglecting Type II error.

In the context of estimating a Hispanic population using surnames, Type I errors are *false positives*, that is, judgments that someone with a given name is Hispanic when that person is in fact not Hispanic; while Type II errors are *false negatives*, that is, judgments that someone with a given name is not Hispanic when that person is in fact Hispanic. The standard way to think about surname matching is in terms of finding a list of the names that are most heavily Hispanic. In so doing, we are looking to minimize Type I error. But, as we show in the next section, to maximize the accuracy of our dichotomous classifications of names as either Hispanic or not Hispanic, what we actually need to do is to set the number of Type I errors (false positives) equal to the number of Type II errors (false negatives). Looking at Type I errors is not enough. Moreover, as we will illustrate in the next section, the optimization rule may lead us to a classification scheme with a large number of false positives!

ERROR 3: Confusing a decision rule for matching names to status as a Hispanic or non-Hispanic that maximizes the likelihood that we correctly classify given *individuals* in terms of their *Hispanicity* or lack thereof with a rule that maximizes the accuracy of our overall estimate of the proportion of the *subpopulation* that is Hispanic.

To minimize errors of individual classification, it would appear that we should simply predict that anyone with a name that had more than a 50 percent probability of being Hispanic should be labeled Hispanic, while anyone with a name that had more than a 50 percent probability of being non-Hispanic should be labeled non-Hispanic. But paying attention to error number 1 tells us that this is too simplistic. We know that $p(H|N)$ is not a constant. In some subpopulations, a given name may have a greater than 50 percent chance of being held by someone who is Hispanic; in other subpopulations that will not be true. But even if we could develop an optimal surname-based rule for classifying individuals, rather counterintuitively, the rule that maximizes the number of individuals who are correctly classified as Hispanic or non-Hispanic (i.e., the rule that minimizes the sum of *false positives* and *false negatives*) is not, in general, the rule that maximizes the accuracy of our prediction of the *overall Hispanic proportion in the subpopulation*. As noted above, that rule requires us to set the number of *false positives* equal to the number of *false negatives*.

ERROR 4(a): When we treat anyone with one of the names on a surname list as Hispanic and anyone whose name is not the list as non-Hispanic, thinking that there exists a *single* list of (heavily Hispanic) surnames that is optimal (i.e., error minimizing in terms of predicting the Hispanic proportion in that subpopulation) in *all subpopulations*.

Just as $p(H|N)$ is not constant, but rather varies with the demographic context, similarly, if we sort names in terms of the proportion of those with that name who are Hispanic, the number of names we want to code as "Hispanic" will vary with the demographic context. Unfortunately, determining the demographic context is normally exactly what we want to use the surname list to establish, so that we have a very real "chicken and egg" problem. The ratio method introduced in the third section of this article is intended to deal with exactly that problem.

ERROR 4(b): When we treat anyone with one of the names on a surname list as Hispanic and anyone whose name is not on the list as non-Hispanic, thinking that the more names we have on the list the more accurate will be our prediction of the Hispanic proportion in the subpopulation.

Accuracy in predicting Hispanic proportion in the subpopulation comes in equating Type I and Type II errors. For any given set of names coded as "Hispanic," the magnitude of each of these types will differ with demographic context. We can go wrong in two different ways: by having more Type I errors than Type II errors, on the one hand, or by having more Type II errors than Type I errors, on the other hand. In general, the smaller the list of names coded as "Hispanic," the higher the proportion of Type II errors and the lower the proportion of Type I errors. The trick is to find the point (the number of names on the list) at which the cumulative sum of each of these two types of errors becomes identical to one another.

Now we turn to an exposition of Bayes Theorem in the context of surname matching. In this next section, we also illustrate each of the four types of errors, and the correct form of analysis, using 2010 Census data.

Bayes Theorem: The Basic Principles and Applications to 2010 Census Data

At the heart of our analyses is Bayes Theorem. Before we state that theorem we need some additional notation. Let $p(H \text{ and } N)$ be the proportion of those in the subpopulation who are both Hispanic and have the given name. This joint probability is linked to the individual probabilities and the conditional probabilities by the *Law of Conditional Probability*.

$$p(H \text{ and } N) = p(H|N) \times p(N) = p(N|H) \times p(H) = p(N \text{ and } H) \tag{1}$$

The *Law of Conditional Probability* is deceptively simple and yet it is key to understanding the potential pitfalls in seeking to use surnames to determine the proportion of Hispanics in a given population.[11] It is also the cornerstone for understanding many other fundamental concepts, such as Type I and Type II errors and Bayes Theorem, that we make use of in this article.[12]

In the context of surname analysis, *Bayes Theorem*, which follows from the *Law of Conditional Probabilities* and some other basic features of probabilities,[13] states that

$$p(H|N) = [p(N|H) \times p(H)] \,/\, [p(N|H) \times p(H) + p(N|\text{non} - H) \times (1 - p(H))]\,. \tag{2}$$

Thus, if we know the proportion of Hispanics who have a given name in the overall population and we, similarly, know the proportion of non-Hispanics who have that same name, then we can use that information to specify the conditional probability that someone with a given name is Hispanic *as a function of the proportion of Hispanics in the subpopulation*. This observation lead us to realize that $p(H|N)$ should be viewed as a joint function of $p(H)$ and other parameters. In other words, as noted earlier, $p(H|N)$ varies with demographic

[11] Note that if H and N are mutually exclusive, then $p(H \text{ and } N) = 0$; if H and N are statistically independent, then $p(H \text{ and } N) = p(H) \times p(N)$.

[12] Note, however, that while Bayes Theorem is fundamental to all the analyses in this article, we are not really engaged in what is commonly called *Bayesian analysis* or *Bayesian inference*. Rather than using subjective judgments about background conditions and a priori likelihoods to adjust probability estimates (see, e.g., McGrayne, 2011), we are merely making use of conditional probabilities in a mathematically straightforward way.

[13] We can derive Bayes Theorem from the identities below: $p(H|N) = [p(N|H) \times p(H)] \,/\, [p(N \text{ and } H) + p(N \text{ and non} - H)] = [p(N|H) \times p(H)] \,/\, [p(N|H) \times p(H) + p(N|\text{ non} - H) \times p(\text{non} - H)]$.

context.[14] The false belief that the proportion Hispanic among bearers of a given surname is a fixed value can be thought of as a variant of the "Blue Cab, Green Cab" probability misassessment made famous by Tversky and Kahneman (1982).[15]

ILLUSTRATING ERROR 1 WITH 2010 CENSUS DATA: Treating $p(H|N)$ as if it were a constant.

Table 1 presents an illustrative set of four surnames chosen to reflect a range of situations along two dimensions: from heavily Hispanic names to names with a low percentage of Hispanics, and from common surnames to less common surnames.[16] For each surname, we show its count in the data set, the proportion of people with that surname found to be Hispanic, the surname's proportion of all the surnames in the Census national data set, its proportion of all Hispanics in that data set, and its proportion of all the non-Hispanics in the data set. That is to say, for each surname, we provide both raw counts and percentage data, along with conditional probabilities both of the conditional probability that, in this data set, a given name is Hispanic (non-Hispanic) and of the conditional probability that a Hispanic (non-Hispanic) has a given name. Additionally, we provide some data on where a given surname ranks with respect to various characteristics of the national data.

In the national data set, the relationship between how numerous is a surname and how likely it is to be Hispanic is complicated by two factors that go in opposite directions. On the one hand, the Hispanic population is more concentrated into a limited number of names than is the non-Hispanic population. For example, half of all Hispanics are captured by only 1,500 surnames. In contrast, it takes nearly 17,000 surnames to capture half of all non-Hispanics. On the other hand, in the national data set there are many fewer Hispanics than non-Hispanics (13.43 percent Hispanic in the sample we are using), which makes it much harder for a highly Hispanic surname to be among the most common. The latter effect is the stronger.[17]

What we can immediately see from these illustrative examples is the need to distinguish proportion from raw count. For example, because "ANDERSON" is such a common surname, even though its percentage of Hispanics is low in the national sample, there

[14]Of course, we need to be careful about the realism of the implicit assumption that prob(name *i*|Hispanic) and prob(name *i*|non-Hispanic) are the same in every subpopulation except for sampling error. But, that assumption is still more plausible than assuming that prob(Hispanic|name *i*) and prob (non-Hispanic|name *i*) are constant, since we know that to be wrong. When we later use a double application of Bayes Theorem to estimate the Hispanic population proportion for cities in California, we check that this assumption is plausible by looking at the magnitude of possible confounds such as Filipino populations who are Hispanic but with a different surname distribution.

[15]We may characterize the Tversky and Kahneman (1982) example as follows. A subject is told that in a given city 85 percent of the taxis are Green Cabs (painted green) and the remaining 15 percent are Blue Cabs (painted blue), and that all witnesses who saw someone being run over (and fatally injured) agree that it was a taxi that fled the accident scene. Moreover, the sole (noncolorblind) witness identified the car involved in the accident as a Blue Cab. The subject is also told that the trial court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80 percent of the time and made an erroneous classification only 20 percent of the time. The subject is then asked: "What is the probability that the cab involved in the accident was blue rather than green?" Most subjects answer with an estimate that is close to 80 percent. The correct answer, using Bayes Theorem, is that the probability equals $0.8 \times 0.15/(0.8 \times 0.15 + 0.2 \times 0.85) = 0.41$. What subjects fail account for is the baseline proportions (0.15 and 0.85) in doing their probability assessments. It is clear that (most) subjects do not really understand the concept of conditional probability.

[16]Recall, however, that all the names in the Census data set we use have at least 300 instances in the national population.

[17]In the 2010 Census data set, when we look at the correlation between surname count and surname proportion Hispanic, we find it to be −0.284. In our analyses, we have arrayed names by *proportion* Hispanic. If we were to eliminate names that were highly Hispanic, but also rare, we could cut dramatically the number of names we would need. For example, to capture 50 percent of the Hispanic population in the United States as a whole, we would go down from 1,500 names to just 113. These names would, on average, be 90.4 percent Hispanic in the national data set.

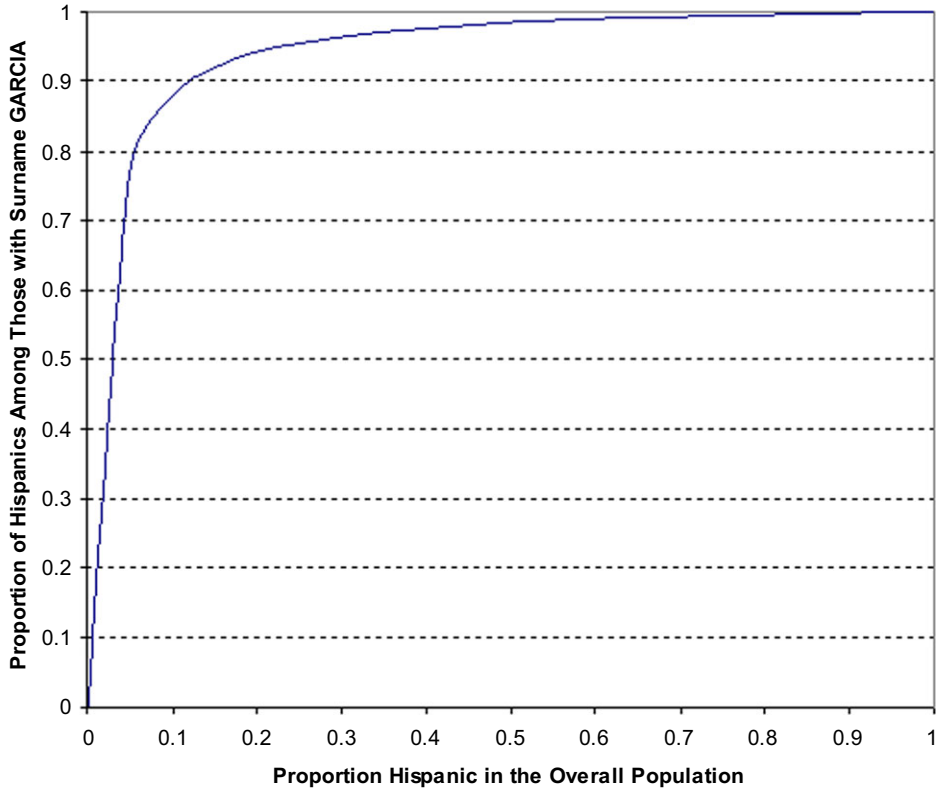
TABLE 1

Illustrative Surname List for Relationships Between Unconditional and Conditional Probabilities Linking Surname with Spanish Origin

Surname	Count of Population ($n = 222,316,554$)	Count of Hispanics	Count of Non-Hispanics	Proportion Hispanic	Proportion of All Population	Rank on Overall Surname Frequency ($n = 53,286$)	Proportion of All Hispanics	Rank on Proportion Hispanic	Proportion of All Non-Hispanics
ANDERSON	762,394	12,046	750,348	0.0158	0.0034293	12	0.00040	31,872	0.00389887
GARCIA	858,289	779,412	78,877	0.9081	0.0038607	8	0.02610	1,533	0.00040985
SAGRERO	433	430	3	0.9931	0.0000019	41,730	0.00001	5	0.00000002
WIST	398	7	391	0.0176	0.0000018	43,380	0.000000235	27,422	0.00000203

FIGURE 1

Hispanic Proportion Among Those with Surname GARCIA as a Function of Overall Hispanic Proportion



are still far more Hispanic "ANDERSONs" than there are Hispanic "SAGREROs," even though those named "SAGRERO" are about 60 times more likely to be Hispanic than are those named "ANDERSON."

To see how the proportion of those with a given surname, say "GARCIA," who are Hispanic varies with the proportion Hispanic in the population or sample, we solve for prob(Hispanic|GARCIA) by substituting the values for prob(GARCIA|Hispanic) and prob(GARCIA|non-Hispanic) from Table 1 into Equation (1) to obtain:

$$\text{prob(Hispanic|GARCIA)} = \frac{0.02610 \times \text{prob(Hispanic)}}{0.02610 \times \text{prob(Hispanic)} + 0.00040985 \times (1 - \text{prob(Hispanic)})}.$$

Figure 1 plots this function as we vary the proportion Hispanic in the population (or sample). It is visually apparent from this graph how the value of (Hispanic|GARCIA) can

vary dramatically depending upon the demographic context. Note, however, that once we have a 10 percent or higher Hispanic population, those with surname "GARCIA" have a 90 percent or more probability of self-identifying as Hispanic.[18]

ILLUSTRATING ERROR 2 WITH 2010 CENSUS DATA: Focusing on Type I error and neglecting Type II error.

As noted earlier, what is usually done with a list of names and their expected Hispanic proportion is to sort them according to the likelihood that a random draw from those with that surname will be Hispanic. From that, a much smaller (and more manageable) list of only the surnames found to have high proportions of Hispanics is generated. Then, anyone with a name on the list is treated as Hispanic, and anyone whose name does not fall on the list is treated as non-Hispanic. Here, the focus is on Type I error, though the presumed justification for doing this is that if we set the threshold appropriately as to what names to include, then the mistakes (Type I errors) we make by including non-Hispanics in the set of names we assign to the category "Hispanic" will (roughly) equal the mistakes (Type II errors) we make by including Hispanics in the set of names we assign to the category "non-Hispanic." Also, for practical reasons of manageability, we wish to use a matching procedure that does not require us to check for tens of thousands of names.

Of course, the equalizing of Type I and Type II errors only occurs at some optimizing cut-off point. If we use too many names, we overcount Hispanics; if we use too few, we undercount them. To find the optimizing point for a known distribution of surnames and a given proportion Hispanic, such as the 13.4 percent Hispanic in the 2010 national data set, we can find the optimal threshold by looking at the intersection of the curve giving the cumulative distribution of non-Hispanic names and the curve giving the reverse cumulative distribution of Hispanic names. When these two curves intersect then the number of non-Hispanics to the left side of the intersection point (Type I errors, false positives) equals the number of Hispanics to the right side of the intersection point (Type II errors, false negatives). The point where the two lines intersect is the point where Type I error equals Type II error, and thus where the two types of errors "cancel out." If we set our surname threshold at this point, then we will be correctly identifying the "true" Hispanic population proportion, that is, in this not quite random national sample, involving only those for whom we have full information about Hispanic status and only names that have at least 300 instances, we will obtain a value of 13.43 percent.

We find that, for the national data set, this intersection occurs at the name that is located at the 8.1 percentage point on the cumulative distribution of names arrayed from most to least Hispanic in percentage in the national sample, as shown in Figure 2. Alternatively, Figure 3 shows these cumulative frequency distributions in terms of Hispanic proportions among names. Here, the intersection point occurs at a name that is roughly 34 percent Hispanic. Of course, the same name, here "VARON," must be the name corresponding to the intersection point in both figures—and it is. Those who hold one of the first 4,310 most (in percentage terms) Hispanic names sum up to comprise exactly 13.43 percent of the people in the data set, that is, the same fraction as the proportion of Hispanics in the data.[19]

[18]Because "GARCIA" is a surname that has a higher proportion of all Hispanics (0.0261) than for non-Hispanics (0.0004) among its members, the curve shown in Figure 1 is convex.

[19]A total of 4,310 is the number of most heavily Hispanic names needed to optimally estimate the Hispanic proportion in the national population when we classify names dichotomously. These most heavily Hispanic names in the United States contain a very high proportion of all Hispanics; 87.9 percent of all Hispanics have one of the 4,310 most Hispanic names. Using Bayes Theorem we can also show that Hispanics make up 87.9 percent of the set of people with one of the 4,310 names on the optimizing list. This equivalence of Hispanic proportion in the name set and proportion Hispanic in the data only holds for the name set that

FIGURE 2

Equalizing Type I and Type II Errors

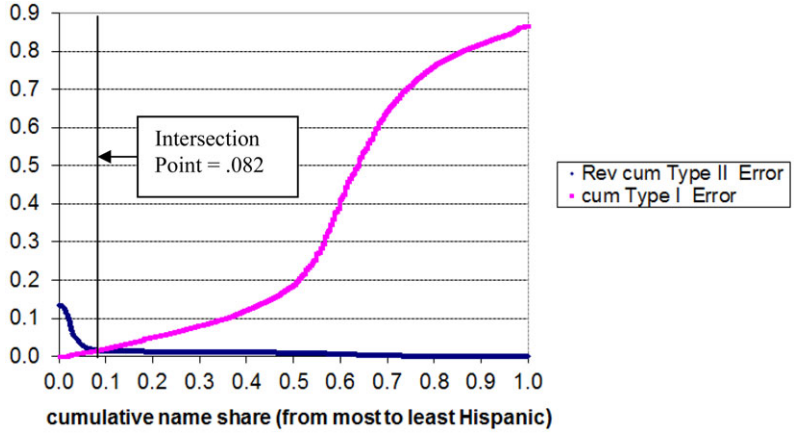
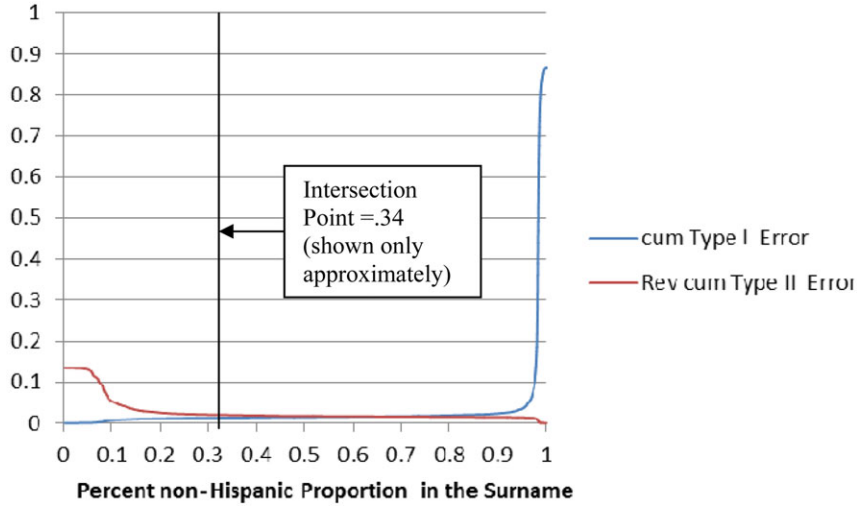


FIGURE 3

Equalizing Type I and Type II Errors by Hispanic Proportion in the Surname (From Most to Least Hispanic)



ILLUSTRATING ERROR 3 WITH 2010 CENSUS DATA: Confusing a decision rule for matching names to status as a Hispanic or non-Hispanic that maximizes the likelihood that we correctly classify given *individuals* in terms of their *Hispanicity* or lack thereof with a rule that maximizes the accuracy of our overall estimate of the proportion of the *subpopulation* that is Hispanic.

equates Type I and Type II errors. A proof of this result is given in the Mathematical Appendix to this article that is available online at the senior author's website ⟨http://www.socsci.uci.edu/~bgrofman/⟩.

TABLE 2

Optimal Number of Most Hispanic Surnames to Treat as 100 Percent Hispanic as a Function of Hispanic Population Proportion (Based on Parameters in 2010 National Census Data for the Subset with Data on Hispanics)

Hispanic Fraction	Optimal Number of Names
0.05	2,620
0.1	3,685
0.2	5,724
0.3	8,525
0.4	11,530
0.5	15,486
0.6	20,011
0.7	24,526
0.8	28,198
0.9	31,596
0.95	34,440

When we put the cutoff at the 4,310th name (VARON), we overcount non-Hispanics by 3,606,488 (in the 4,310 names that we count as 100 percent Hispanic that are not 100 percent Hispanic), and we undercount Hispanics by 3,606,581 (in the 48,077 names that we count as 100 percent non-Hispanic that are not 100 percent non-Hispanic). So we are making many errors of both Type I and Type II; but these errors are canceling out. Moreover, we find, rather counterintuitively, that it is "optimal" to treat names that are 34 percent or more Hispanic as if they were 100 percent Hispanic, while treating names less than 34 percent Hispanic as non-Hispanic. In other words, we are counting names that are not even majority Hispanic as Hispanic—and that is exactly the right thing to do in these circumstances.

As we have emphasized earlier, optimizing predictive accuracy of the mean proportion Hispanic in the sample is not the same thing as minimizing the number of Type I errors, minimizing the number of Type II errors, or minimizing the sum (or some weighed average) of Type I and Type II errors. Moreover, for aggregate optimization purposes, how many (what proportion of) individuals we wrongly classify is essentially irrelevant. It can be perfectly okay, for aggregate predictive purposes, to misclassify many individuals in both directions (false positives and false negatives), *if*, in so doing, the misclassifications in each direction exactly cancel out.

ILLUSTRATING ERROR 4(a) WITH 2010 CENSUS DATA: When we treat anyone with one of the names on a surname list as Hispanic and anyone whose name is not the list as non-Hispanic, thinking that there exists a *single* list of (heavily Hispanic) surnames that is optimal (i.e., error minimizing in terms of predicting the Hispanic proportion in that subpopulation) in *all subpopulations.*

We show in Table 2 the optimal size of Spanish surname lists for various proportions of Hispanic in the overall population, ranging from 5 percent to 95 percent for a sample that has the same conditional probabilities for each surname's fraction of the Hispanic and of the non-Hispanic populations as is true in the 2010 national data set with Hispanic information we have been making use of.

ILLUSTRATING ERROR 4 (b) WITH 2010 CENSUS DATA: When we treat anyone with one of the names on a surname list as Hispanic and anyone whose name is not on

the list as non-Hispanic, thinking that the more names we have on the list, the more accurate will be our prediction of the Hispanic proportion in the subpopulation.

We can look at the question of an optimal cutoff for the surname list to be treated as 100 percent Hispanic from the reverse perspective. We have shown that, for our national data subset, with a 13.43 percent Hispanic population share, the optimizing cut-off point is 4,310. That is, if we take the 4,310 names that are most Hispanic, in Hispanic population percentage, and treat them as 100 percent Hispanic, those 4,310 surnames are held by a set of individuals who together constitute 13.43 percent of the national population, that is, the actual proportion. But, what happens if we use a smaller number of surnames to estimate the national Hispanic population via surname matching (or a larger one)? If we were to use, say, only the top 639 Hispanic percentage surnames in our data set, we would estimate the national Hispanic population to be only 8.39 percent, that is, we would miss more than a third of all Hispanics. If there are very few Hispanics in a population, it is easier (requires fewer surnames) to accurately estimate the proportion Hispanic in the population by counting as Hispanic all those with a given relatively small set of surnames; while we need many names to accurately assess the Hispanic population proportion when the Hispanic population proportion is high. However, we can still use too many names.

If we were to use the top 8,000 names in Hispanic percentage surnames, we would estimate the national Hispanic population to be 15.56 percent, that is, we would be estimating the Hispanic population to be about 115 percent of its actual size. If we were to use 12,497 names, which is the number most often used in the studies done in the 1980s, for 2010 data, we would estimate the national Hispanic population to be 18.19 percent, about 135 percent of its actual size.

Using Bayes Theorem to Estimate the Hispanic Proportion of California Cities

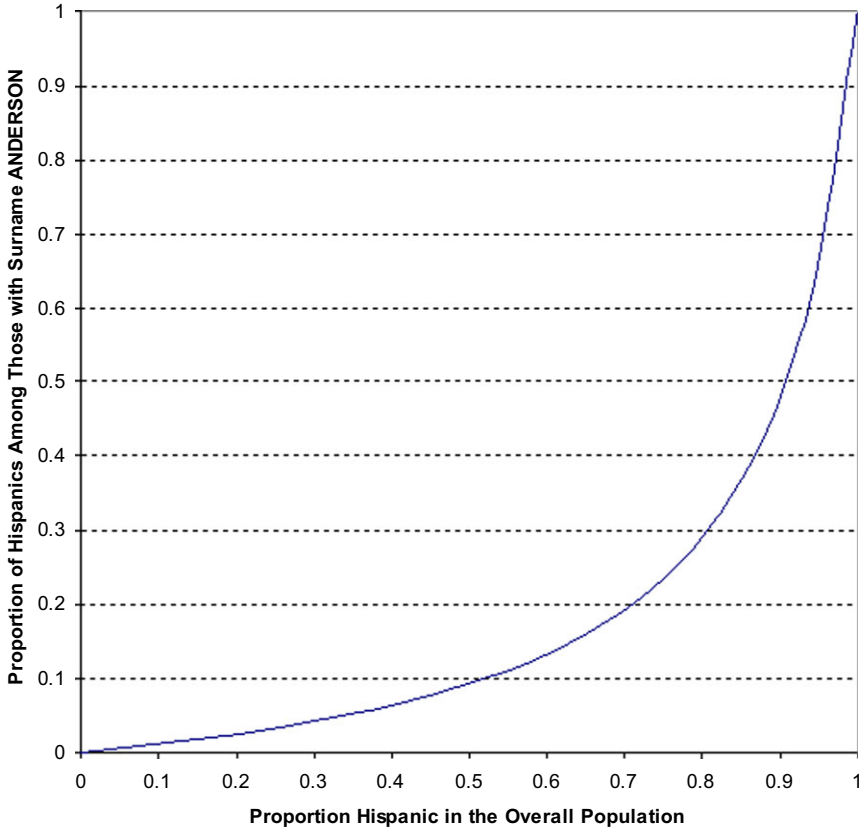
Figure 4 shows how the occurrence of the surname "ANDERSON" varies with the proportion Hispanic in the population. This graph parallels the earlier graph shown for the surname "GARCIA" (Figure 1).

In the national population, which is 13.43 percent Hispanic, there are somewhat more "GARCIAs" (858,289) than there are "ANDERSONs" (762,394), for a ratio of 1.13. What happens to the relative proportion of these two names in the population as we change the overall proportion Hispanic? To answer this question, we can integrate into a single graph the information from Figures 1 and 4 about the names "ANDERSON" and "GARCIA." As shown in Figure 5, what we see is that, in a population that is 0 percent Hispanic, there are 1/10th as many "GARCIAs" as there are "ANDERSONs." In a population that is 100 percent Hispanic, there are estimated to be 65 times as many "GARCIAs" as "ANDERSONs." The ratio hits 1 at about 12 percent Hispanic in the population.

Figure 5 shows how the *ratio* of the occurrence of given pairs of surnames, "GARCIA" and "ANDERSON," can be expected to vary with the proportion Hispanic in the population. But, we can also do the analysis in the opposite direction. If, say, we observe a given ratio of "GARCIAs" to "ANDERSONs" in a population, we can read from the graph in Figure 5 what proportion Hispanic in the population would have been expected to give rise to that ratio.

FIGURE 4

Hispanic Proportion Among Those with Surname ANDERSON as a Function of Overall Hispanic Proportion

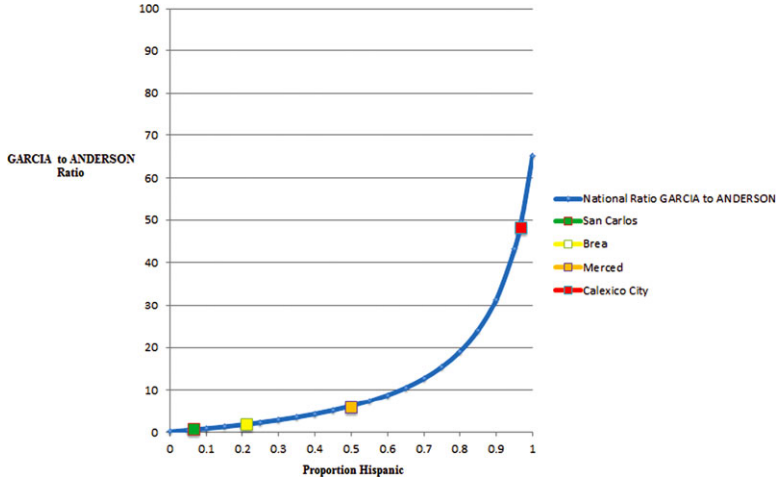


In Figure 5, we also show estimated proportion of Hispanic in the city based on the "GARCIA" to "ANDERSON" ratio for four California cities—San Carlos, with a low Hispanic population of 10.1 percent (ca. 2010), Brea, with a moderately low Hispanic population of 25 percent (ca. 2010), Merced, with a moderately high Hispanic population of 49.8 percent (ca. 2010), and Calexico, with a very high Hispanic population of 98.6 percent (ca. 2010). Names are counted by using ReferenceUSA, which provides a database of directory information compiled from White Pages nationwide. ReferenceUSA has a total of 18,655 names for San Carlos; 16,835 names for Brea; 25,824 names for Merced; and 10,037 names for Calexico. According to the 2010 Census, San Carlos has a population of 28,406 and 11, 332 households; Brea has a population of 39,282 and 14,386 households; Merced has a population of 78,958 and 23,753 households; and Calexico has a population of 38,572 and 9,561 households.

By looking at the phonebook data, we get a ratio of "GARCIAs" to "ANDERSONs" for each of the four Californian cities. A ratio of 0.59 is found for San Carlos (24/41); a ratio of 1.780 is found for Brea (73/41); a ratio of 6.08 (237/39) is found for Merced; and a ratio of 48.2 is found for Calexico (241/5). Translating backward from the ratio, we can get to the estimated Hispanic population proportion that corresponds to that ratio in

FIGURE 5

Proportion of Hispanic in San Carlos (Actual $H = 0.10$), Brea (Actual $H = 0.25$), Merced (Actual $H = 0.50$), and Calexico (Actual $H = 0.99$): Estimated from GARCIA to ANDERSON Ratio in City Phonebooks



the national data. For San Carlos, the estimated Hispanic proportion is around 0.067; for Brea, the estimated Hispanic proportion is 0.21; for Merced it is 0.48; and for Calexico it is 0.964. Considering (a) that we are projecting values derived from national data into particular cities in California, (b) that the phonebook data suffer from an unknown bias in terms of the relative proportions of Hispanics and non-Hispanics who are too poor to have land lines, (c) that the phonebook data suffer from an unknown bias in terms of the relative proportions of Hispanics and non-Hispanics who can afford to have land lines but who choose to rely on a cell phone or Skype, and (d) that the phonebook data suffer from an unknown bias in terms of the relative proportions of Hispanics and non-Hispanics who have landlines but choose not to be listed in the online phone book, it is truly remarkable to have the kind of fit shown here: 0.067 versus 0.10, 0.21 versus 0.25, 0.483 versus 0.498, and 0.963 versus 0.986.

The use of the names "ANDERSON" and "GARCIA" is not accidental, though it is also true that each of these names has a personal significance to one of the authors. They were chosen because we expect the pair-wise ratios to be most predictive of the true Hispanic proportions if (a) one name is heavily Hispanic and one name is heavily non-Hispanic, (b) both are common names, and (c) each has a nontrivial occurrence rate among the opposite ethnicity. "GARCIA" and "ANDERSON" satisfy these properties. Looking at ratios involving names that are uncommon will not be useful when we are looking at data subsets smaller than the full national data set, since such names may be nonexistent in our data, or have so few instances that ratio estimates will be misleading because of sampling error.

Even though there are some non-Hispanic groups in California, for example, Portuguese and Filipinos, who have a high incidence of "Hispanic" names, the problem this nonuniformity of name structure will cause for ratio analyses of the sort we have performed here is a matter that can only be studied empirically. For California cities in general, and the four cities for which we have reported analyses using the pair-wise ratio method in

particular, we have checked to see if the presence of Filipinos or Portuguese is large enough to cause any kind of real problem, and it clearly is not. On the other hand, while we certainly recognize that there are various Hispanic populations that have a different surname structure (e.g., Mexican American, Cubans, Central Americans, etc.), and the distribution of these different Hispanic groups varies geographically, in areas where there are substantial concentrations, say of Cuban Americans, it should be possible to develop tailor-made surname distribution statistics for such groups.

Discussion

As a necessary prequel to our subsequent results for California cities, we began this article by showing how to use Bayes Theorem to demonstrate that the optimal number of (most heavily Hispanic) names to count as "Hispanic" varies with the demographic context in a way that can be specified precisely in terms of balancing off Type I errors (false positives) and Type II errors (false negatives). As a population grows more (less) Hispanic, the proportion of Hispanics among those with any given surname will grow (decline). However, while these analyses show that we size the list of names we use for identifying Hispanic names (with names not on the list being counted as non-Hispanic) to vary with demographic context, they do not show us how to solve the "chicken and egg" problem of finding an optimal size for the list to be used for any particular subpopulation.[20] If we knew the Hispanic subpopulation proportion, we would not need to be doing surname matching. The real contribution of the article is the new pair-wise ratio method we provide to estimate the Hispanic proportion via surname data that allows us to bypass this "chicken and egg" problem.

Using a phone directory, we tested this method with data from four California cities that vary greatly in their proportion Hispanic (from 10 percent to over 98 percent). Despite all the obvious limitations of a phone-directory-based list of names, for all four cities, we find a remarkable fit to the estimates derived from only two names, "ANDERSON" and "GARCIA." That simple ratio approach, which requires us to count only two names, was never off by more than 4 percentage points and, in some cities, came within 2 percentage points of the true value. It is also far easier to apply than the usual surname matching, with lists of names that may number in the thousands.

We believe this pair-wise approach is one very much worthy of further investigation. By conducting this type of analysis for selected pairs of surnames, we can, we believe, set plausible bounds on the likely proportion Hispanic in the population whose ethnic characteristics we are seeking to estimate, at least if that population is large enough that sampling error will not make it impossible to derive reliable results from the ratio method. Moreover, if we need more precision, we can use this method as only a first approximation for the Hispanic proportion of the electorate. Then we would use the estimate derived

[20]No Census publications about surname matching of which we are aware lay out in a clear fashion exactly how prob(Hispanic|name *i*) can be expected to vary with prob(name *i*|Hispanic) and prob(Hispanic) in the data set. Furthermore, there does not appear to be an academic article that does so clearly either. Passel and Word (1980) suggest that the Spanish surname list they compile, one with over 12,000 names, should be used only in areas of high Hispanicity. But they also acknowledge that even using 12,000+ names will tend to underestimate Hispanic population in areas of very high Hispanic concentration, although they indicate that the magnitude of error in this instance, which they assert to be around 5 percentage points, is tolerable. Word and Perkins (1996) simply caution those doing Spanish surname matching that the accuracy of any list varies with geography.

from a surname list that was appropriate for approximately that proportion Hispanic in the name list to develop a more accurate final estimate.

The idea of surname ratios may also be applicable to developing other ways to improve surname matching. For example, in dealing with Asian surnames, names like Lee tend to be highly context dependent. By using ratios such as Kim/Lee or Fong/Lee, we may be able to improve our ability to differentiate among Asian populations (say Korean vs. Chinese) and, in a similar fashion, to distinguish Lees who are of Asian descent from those who may be African American or Caucasian.

REFERENCES

Abrahamse, Allan F., Peter A. Morrison, and Nancy Minter Bolton. 1994. "Surname Analysis for Estimating Local Concentrations of Hispanics and Asians." *Population Research and Policy Review* 13:383–98.

Barreto, Matt, Gary Segura, and Nathan D. Woods. 2004. "The Mobilizing Effect of Majority Minority Districts on Latino Turnout." *American Political Science Review* 98(1):65–75.

Bhavnani, Rikhil R. 2012. A Primer on Voter Discrimination Against India's Lower Caste Politicians: Evidence from Observational Data and Survey Experiments. Unpublished manuscript, University of Wisconsin, Madison.

Elliot, Marc N., Allen Fremont, Peter A. Morrison, Philip Pantoja, and Nicole Lurie. 2008. "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Record Lack Self-Reported Race/Ethnicity." *Health Services Research Journal* 43(5):1722–35.

Elliot, Marc N., Daniel F. McCaffrey, Brian K. Finch, David J. Klein, Nate Orr, Megan K. Beckett, and Nicole Lurie. 2009. "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities." *Health Services and Outcomes Research Methodology* 9(2):69–83.

Grofman, Bernard, and Jennifer Garcia. 2014. "Using Spanish Surname to Estimate Hispanic Voting Population in Voting Rights Litigation: A Model of Context Effects Using Bayes Theorem." *Election Law Journal* 13(3):375-93.

Harris, J. Andrew. 2012. A Method for Extracting Information about Ethnicity from Proper Names. Unpublished manuscript, Nuffield College, Oxford, November 22.

Mayer, Kenneth R. 2011. Rule 26 Expert Witness Report in *Voces de La Frontera et al. v. Members of the Wisconsin Government Accountability Board*. Case No 11-CV-1011 JPS-DPW-RMD (Consolidated with *Baldus et al. v. Government Accountability Board of Wisconsin*, Federal District Court, Case No. 11-CV-562 JPS-DPW-RMD), decided March 22, 2012.

McGrayne, Sharon Bertsch. 2011. *Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. New Haven: Yale University Press.

Passel, Jeffrey S., and David L. Word. 1980. "Constructing the List of Spanish Surnames for the 1980 Census. An Application of Bayes Theorem." Paper presented at the Annual Meeting of the Population Association of America, Denver, CO, April 1012.

Tversky, A., and D. Kahneman. 1982 "Evidential Impact of Base Rates." Pp. 153–62 in D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Word, David L., and Colby Perkins Jr. 1996. "Building a Spanish Surname List for the 1990s—A New Approach to an Old Problem." U.S. Bureau of the Census, Population Division. Technical Working Paper 13, March.

APPENDIX:

Modeling the Relationship Between Surname and Hispanicity in Different Demographic Contexts

Bernard Grofman
School of Social Sciences, University of California Irvine

BGTravel@uci.edu

April 12, 2015

Five propositions about surname matching

Let

h_i = the number of Hispanics among those with the ith name

$nonh_i$ = the number of non-Hispanics among those with the ith name

p_i = the number of people with the ith name

N = total number of distinct names in the data set

H = total number of Hispanics in the data set

nonH = total number of non-Hispanics in the data set

prob(Hispanic|name i) = The proportion of individuals with a given name who self-identify as Hispanic/of Spanish heritage

prob(name i |Hispanic) = the proportion of Hispanics who have a given name (in the national sample of non-Hispanics)

prob(non-Hispanic|name i) = The proportion of individuals with a given name who self-identify as non-Hispanic

prob(name i |non-Hispanic) = the proportion of non-Hispanics who have a given name (in the national sample of non-Hispanics)

prob(Hispanic) = 1 – prob (non-Hispanic) = the proportion of Hispanic /those of Spanish heritage in the sample

$\bar{H}_n =$ the cumulative mean proportion Hispanic among the names arrayed from most to least Hispanic, for the range from the first to the nth name.

$$\bar{H}_n = \sum_{i=1}^{i=n} h_i \Big/ \sum_{i=1}^{i=n} p_i$$

$non\bar{H}_n =$ the cumulative mean proportion non-Hispanic among the names arrayed from most to least Hispanic, for the range from the first to the nth name.

$$non\bar{H}_n = \sum_{i=1}^{i=n} nonh_i \Big/ \sum_{i=1}^{i=n} p_i$$

Proposition 1: If, for each surname, in any sample, the surname's share of total Hispanic population, prob (name i |Hispanic), and its share of total non-Hispanic population, prob (name i |non-Hispanic), is treated as a random sample from the corresponding national name distributions within each of the two groups, then the proportion of individuals with a given surname who self-identify as Hispanic, prob(Hispanic|name i), is not a constant, but is a function of the Hispanic proportion (and thus also of the non-Hispanic proportion) of the sample we are looking at. In particular,

$$\text{prob(Hispanic|name i)} = \frac{\text{prob(name i |Hispanic)} * \text{prob(Hispanic)}}{\text{prob(name i |Hispanic)} * \text{prob(Hispanic)} + \text{prob(name i |non-Hispanic)} * \text{prob(non-Hispanic)}}$$

Proof: The result is simply a restatement of Bayes Theorem. The basis of Bayes Theorem is the identity

prob(Hispanic|name i) * prob(name i) = prob(name i |Hispanic) * prob(Hispanic)

From this identity we derive the equation

prob(Hispanic|name i) = (prob(name i |Hispanic) * prob(Hispanic)) / prob(name i)

Now, we can use a further identity, namely

$$p(A) = p(A|B)p(B) + p(A|\text{not } B)p(\text{not } B)$$

to show, after some straightforward algebra, that

$$\text{prob(Hispanic|name i)} = \frac{\text{prob(name i |Hispanic)} * \text{prob(Hispanic)}}{\text{prob(name i |Hispanic)} * \text{prob(Hispanic)} + \text{prob(name i |non-Hispanic)} * \text{prob(non-Hispanic)}} \quad (1)$$

But, from Eq. (1), we can see that prob(Hispanic|name i) depends both on the underlying conditional probabilities, prob (name i |Hispanic) and prob (name i |non-Hispanic), which under the given assumptions, for a large enough sample, we may take to be essentially fixed, while the further parameter,

prob(Hispanic) = 1- prob(non-Hispanic),

is context dependent. Thus, prob(Hispanic|name i) varies with the Hispanic proportion in the sample. *q.e.d.*

Proposition 2: If we array names from most Hispanic to least Hispanic and we treat the first s names as 100% Hispanic and the remaining names (from the (s+1)th to the Nth) as

non-Hispanic, then the value of s such that the names classified as Hispanic yield the true Hispanic population is given by s such that

$$\sum_{i=1}^{i=s} nonh_i = \sum_{i=s+1}^{i=N} h_i$$

Proof: If we array names from most Hispanic to least Hispanic and we treat the first s names as 100% Hispanic and the remaining names (from the (s+1)th to the Nth) as non-Hispanic, then we are positing that the total Hispanic population is given by $\sum_{i=1}^{i=s} p_i$, but

$$\sum_{i=1}^{i=s} p_i = \sum_{i=1}^{i=s} h_i + \sum_{i=1}^{i=s} nonh_i = \sum_{i=1}^{i=s} h_i + \sum_{i=s+1}^{i=N} h_i = \mathrm{H}.$$

q.e.d.

In other words, to maximize the accuracy of our [0,1] classifications of names as either Hispanic or not Hispanic we wish to set the number of Type I errors (false positives) equal to the number of Type II errors (false negatives).

Proposition 3: If, for each surname, its share of total Hispanic population and its share of total non-Hispanic population is treated as fixed, then the number of (most Hispanic) names we would need to use to equalize the number of Type I and Type II errors increases with the proportion Hispanic in the total population.

Proof: The proof of this proposition is quite straightforward. For any given cutoff point, as we increase the proportion Hispanic in the sample, the number of Type I errors (false positives) above that cutoff declines, since we are reducing the share of non-Hispanics in the population. Thus, the number of non-Hispanics in each surname will also go down since we are assuming that the proportion of non-Hispanics coming from any given surname is fixed.

Similarly, for that same cutoff point, as we increase the proportion Hispanic in the sample, the number of Type II errors (false negatives) below that cutoff increases, since we are increasing the share of Hispanics in the population. Thus, the number of Hispanics in each surname will also go up since we are assuming that the proportion of Hispanics coming from any given surname is fixed. But, if we have reduced Type I error to the right of the former cutoff and increased Type II error in the other direction, then to equalize the two now requires us to increase the number of names we count as 100% Hispanic, i.e., lower the threshold.[1] *q.e.d.*

In the next proposition we offer an alternative way to think about what needs to be equalized to maximize the predictive success of our choice of name threshold.

Proposition 4: If we array names from most Hispanic to least Hispanic and we treat the first s names as 100% Hispanic and the remaining names (from the (s+1)th to the Nth) as non-Hispanic, then the value of s such that the names classified as Hispanic yield the true Hispanic population is given by s such that the (cumulative) average Hispanic share of the population among the names from the first to the sth name equals the proportion of the total Hispanic population found among those names, i.e.,

$$\bar{H}_s = \sum_{i=1}^{i=s} h_i / \sum_{i=1}^{i=s} p_i = (\sum_{i=1}^{i=s} h_i) / H$$

Proof: Once we set up this proposition in mathematical notation, the result become obvious, since we have the same numerator on both sides and the denominators are equal by assumption of our choice of s. *q.e.d.*

The intuitive meaning of this proposition is less clear than that of either of our other three propositions, but in the later empirical section we will be able to give Proposition 4 an enlightening (and perhaps surprising) empirical content.

---

[1] Note that this result does not necessarily go through were we to array names not according to their percentage Hispanic but according to what proportion of all Hispanics are found with that name. The latter takes into account how common the name is, while the former does not.

Proposition 5: Consider any two surnames, say A and B, that have the property that they differ from one another both in the proportion of all those who claim Hispanic heritage who have each of the two surnames and in the proportion of all those who do not claim Hispanic heritage who have each of the two surnames. If we assume that any given population of Hispanics is a close to random sample from the national population of Hispanics in terms of surnames and any population of non-Hispanics is a close to random sample from the national population of non-Hispanics in terms of surnames, and we know the shares of the national Hispanic and national non-Hispanic population, respectively, that each surname constitutes, by finding the ratio of those in a given population who have surname A to those who have surname B, we can directly infer the Hispanic proportion of that population.

Proof: This proposition follows directly from the law of conditional probability and from Bayes Theorem. We start with

prob(Hispanic|name A) * prob(name A) = prob(name A |Hispanic) * prob(Hispanic)

From this identity we derive the equation

prob(name A) = (prob(name A |Hispanic)* prob(Hispanic)/ prob(Hispanic|name A)

Similarly,

prob(name B) = (prob(name B |Hispanic)* prob(Hispanic)/ prob(Hispanic|name B)

Dividing these two equations we obtain the ratio

prob(name A)/ prob(name B) =

$$\frac{\text{prob(name A |Hispanic)* prob(Hispanic) / prob(Hispanic|name A)}}{\text{prob(name B |Hispanic)* prob(Hispanic)/ prob(Hispanic|name B)}} \qquad (2)$$

Therefore,

$$\text{prob(name A)/ prob(name B)} =$$

$$\frac{\text{prob(name A |Hispanic) / prob(Hispanic|name A)}}{\text{prob(name B |Hispanic)/ prob(Hispanic|name B)}} \qquad (2)'$$

since one of the terms in Eq. (2) is found in both numerator and denominator and may be cancelled out.

Moving terms from numerator to denominator, we may write Eq. (2)' as Eq. (2)'' below.

$$\text{prob(name A)/ prob(name B)} =$$

$$\frac{\text{prob(name A |Hispanic) * prob(Hispanic|name B)}}{\text{prob(name B |Hispanic)*prob(Hispanic|name A)}} \qquad (2)''$$

Now, we can twice substitute the identity of Bayes Theorem, Eq. (1) into Eq. (2), to eliminate two of the conditional probabilities in that equation. We obtain, after some algebra, Eq. (3).

prob(name A)/ prob(name B) =

$$\frac{\text{prob(name B |Hispanic) * prob(Hispanic) * prob(name A |Hispanic) /(prob(name A |Hispanic) * prob(Hispanic) + prob(name A |non-Hispanic) * prob(non-Hispanic))}}{\text{prob(name A |Hispanic) * prob(Hispanic) * prob (name B |Hispanic) /(prob(name B |Hispanic) * prob(Hispanic) + prob(name B |non-Hispanic) * prob(non-Hispanic))}} \quad (3)$$

Which, in turn, after cancellation, simplifies to

prob(name A)/ prob(name B) =

$$\frac{\text{prob(name A |Hispanic) * prob(Hispanic) + prob(name A |non-Hispanic) * prob(non-Hispanic)}}{\text{prob(name B |Hispanic) * prob(Hispanic) + prob(name B |non-Hispanic) * prob(non-Hispanic)}} \quad (3)'$$

But, since we may take prob(name A |Hispanic), prob(name A |non-Hispanic), prob(name B |Hispanic), and prob(name B |non-Hispanic) to be essentially known parameters (from the national sample), and since

prob(Hispanic) = 1 – prob(non-Hispanic)
= the proportion of Hispanic /those of Spanish heritage in the sample,

once we know the actual ratio of those with surname A to those with surname B in our sample, under the above assumptions, by plugging in the other four known (subject only to sampling error) parameter values into Eq. (3), after straightforward simple algebra, we can directly calculate the Hispanic proportion in the sample which, of course, is what we want to find. *q.e.d.*