CSD Center for the Study of Democracy
An Organized Research Unit
University of California, Irvine
www.democ.uci.edu

Introduction

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.[1] If the (sur)names held by members of the group are relatively distinctive, and if we have the names of those in the population whose group composition we wish to estimate, by matching (sur)name with estimated ethnicity we may be able to derive estimates of group population shares, e.g., from lists of hospital patients, or of registered voters, or of purchasers of some particular commodity. In this essay, we only examine issues related to Spanish surname matching in the United States, [2]but essentially identical issues arise with respect to name matching for other ethnicities, e.g., 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 U.S.[3] 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).[4]

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.[5] 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 provides the proportion of self-identified Hispanics for each name for the country as a whole. This list is a public document, [6] and can be downloaded as a data file. Similar lists were generated by the Census Bureau for earlier periods,[7] and other surname matchup-lists have also been created by various entities,[8] 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 which are found to have high/highest proportions (or, in some applications, numbers) of Hispanics, treating any one with a name on the list as Hispanic and anyone whose name does not fall on the list as non-Hispanic.[9] 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 Census, 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.[10] 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 this data. Indeed, there is no developed theory to justify using any particular cutoff point as to how many names should be treated as Hispanic in some given application of surname matching technology.[11]

The focus of this essay is on using Bayes Theorem to model the relationship between surname and Hispanic ethnicity so as to generate an approach that improves substantially on present estimation practices. We consider how best to use surnames to identify the proportion of Hispanics in a given population or sample. We argue that virtually all previous efforts have been flawed because of a failure to recognize the importance of demographic context. To model the relationship between proportion Hispanic of the overall population and the level of Hispanicity of any given surname, we take the proportion of all Hispanics in the population who come from any given surname as given (from the U.S. Census), and the proportion of all non-Hispanics in the population who come from any given surname as given (from the U.S. Census), and then make use of Bayes Theorem. [12]

We show that there is no such thing as <u>the</u> proportion of bearers of a given name who are Hispanic. We show that the demographic context determines the size of the list (i.e., number of names classified as "Hispanic" as based on the chosen cutoff point for proportion Hispanic in the surname) which should be used. How Hispanic any given name turns out to be is a function of the overall Hispanicity (i.e., Hispanic proportion) of the population, which affects both the distribution of names and the conditional probability that the possessor of any given name will be Hispanic. Thus, the point at what we should draw the line between "Hispanic" surnames and "non-Hispanic surnames" turns out be a question that cannot be answered in general.

In particular, using Bayes Theorem, we show 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. Then, in the most original part of the paper, we offer a new pairwise ratio method to estimate the Hispanic proportion using the surname data. Using the phone directory, we test this method with data from four California cities which vary greatly in their proportion Hispanic. 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.

We state in the next section our main results in the form of propositions to which we provide informal proofs. Then, we illustrate these ideas with Census data using the matchup between surname and claimed Spanish heritage found in the sample name list (of names with more than 300 instances) generated by census staff in 2010. We believe that the methodology we offer has a wide variety of potential applications, such as the study of elections and of racial or ethnic differences in political participation, especially in situations where no survey data on some given set of elections is available (or where that data does not allow for reliable estimates of

racial/ethnic differences in voting). And we believe there are straightforward ways to adapt this methodology to contexts other than those involving "Spanish" surnames.

Modeling the Relationship Between Surname and Hispanicity in Different Demographic Contexts

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.

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

$non\bar{H}_n$ = 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.

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

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,

prob(Hispanic|name i) =

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

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

prob(Hispanic|name i) =

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

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

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

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.[13] *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.

> 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

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

Therefore,

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).

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

Which, in turn, after cancellation, simplifies to

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

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.*

Illustrating our main propositions with 2010 Census data

Illustrating Proposition 1

Using information compiled from the 2010 Census name and Spanish origin data, we make Proposition 1 concrete. 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.[14] 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.[15]

Table 1: Illustrative Surname List for Relationships Between Unconditional and Conditional Probabilities Linking Surname with Spanish Origin

surname	Count of pop (n= 222316554)	count of Hispanics	count of non-Hispanics	prop Hispanic	prop of all pop	rank on overall surname frequency (n= 53286)	prop of all Hispanics	rank on proportion Hispanic	prop of all non-Hispanics
ANDERSON	762394	12046	750348	0.0158	0.0034293	12	0.00040	31872	0.00389887
GARCIA	858289	779412	78877	0.9081	0.0038607	8	0.02610	1533	0.00040985
SAGRERO	433	430	3	0.9931	0.0000019	41730	0.00001	5	0.00000002
WIST	398	7	391	0.0176	0.0000018	43380	0.000000235	27422	0.00000203

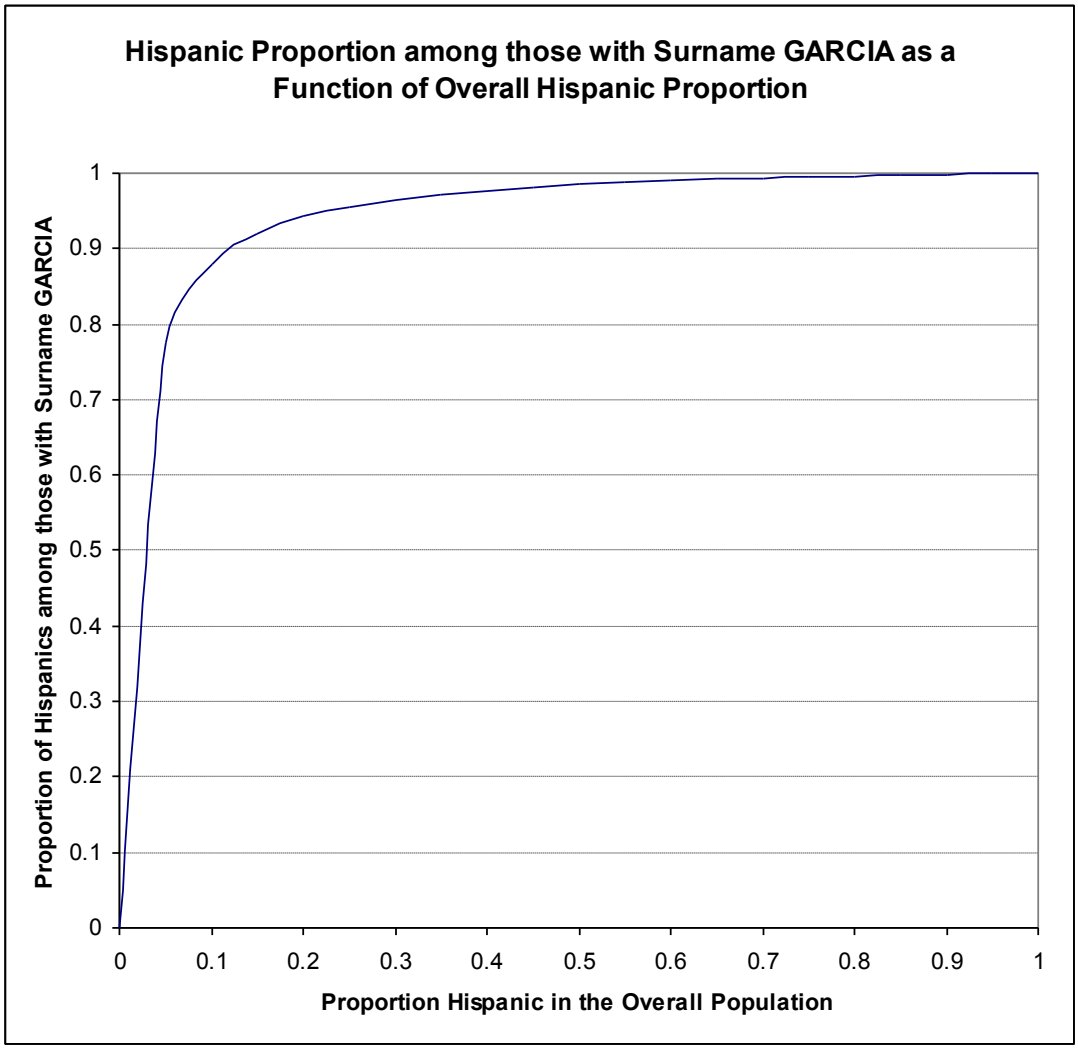
What we can immediately see from these illustrative examples is the need to distinguish proportion from raw count. For example, because ANDERSON is such a common surname, even though its percentage of Hispanics is low in the national sample, there are still far more Hispanic ANDERSONs than there are 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 Eq. (1), to obtain

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

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.[16]

Figure 1



Illustrating Proposition 2

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. The 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 error only occurs at some optimizing cutoff 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% 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, i.e., 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%. 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. 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 VERON, 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% of the people in the data set, i.e., the same fraction as the proportion of Hispanics in the data.

Figure 2

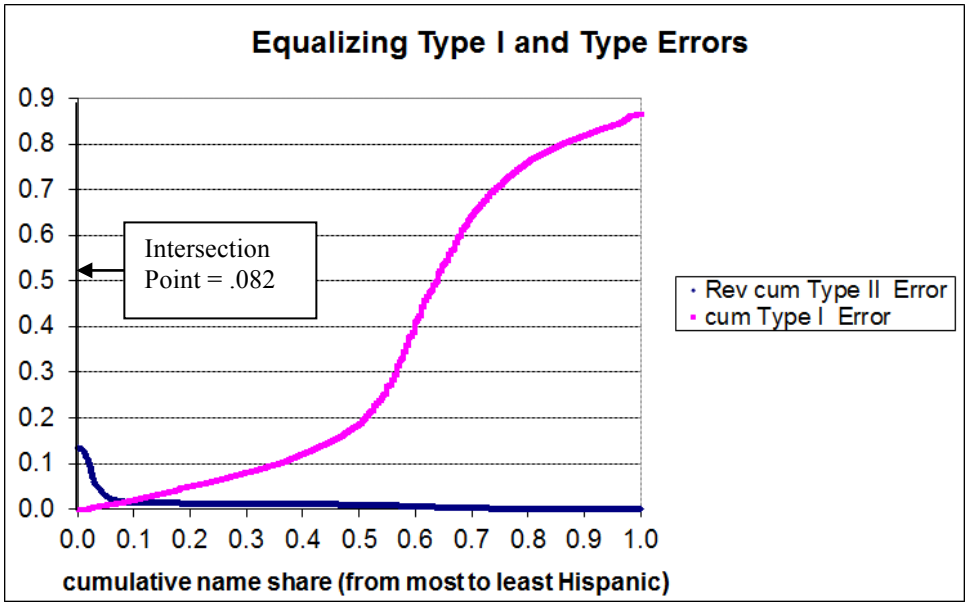
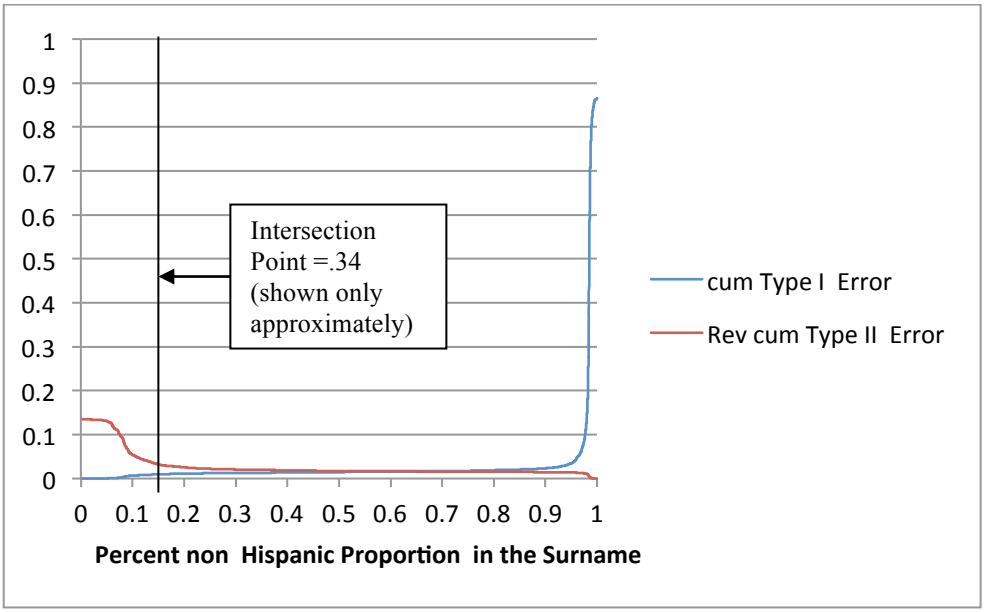


Figure 3: Equalizing Type I and Type II errors by Hispanic Proportion in the Surname (from most to least Hispanic)



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% Hispanic that are not 100% Hispanic), and we undercount Hispanics by 3,606,581 (in the 48,077 names that we count as 100% non-Hispanic that are not 100% non-Hispanic). So we are making many errors of both Type I and Type II; but these errors are cancelling 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.[17]

Illustrating Proposition 3

We show in Table 2 the optimal size of Spanish surname lists for various proportions of Hispanic in the overall population, ranging from 5% to 95% 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.

Table 2: Optimal Number of Most Hispanic Surnames to Treat as 100% Hispanic as a Function of Hispanic Population Proportion
(based on parameters in 2010 national census data for the subset with data on Hispanics)

Hisp fraction	optimal number of names
0.05	2620
0.1	3685
0.2	5724
0.3	8525
0.4	11530
0.5	15486
0.6	20011
0.7	24526
0.8	28198
0.9	31596
0.95	34440

We see from this table that, as shown in Proposition 3, 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.

We can also 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% Hispanic population share, the optimizing cutoff 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% Hispanic, those 4,310 surnames are held by a set of individuals who together constitute 13.43% of the national population, i.e., 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%, i.e., we would miss more than a third of all Hispanics. 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%, i.e., we would be estimating the Hispanic population to be about 115% 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%, about 135% of its actual size.

Illustrating Proposition 4

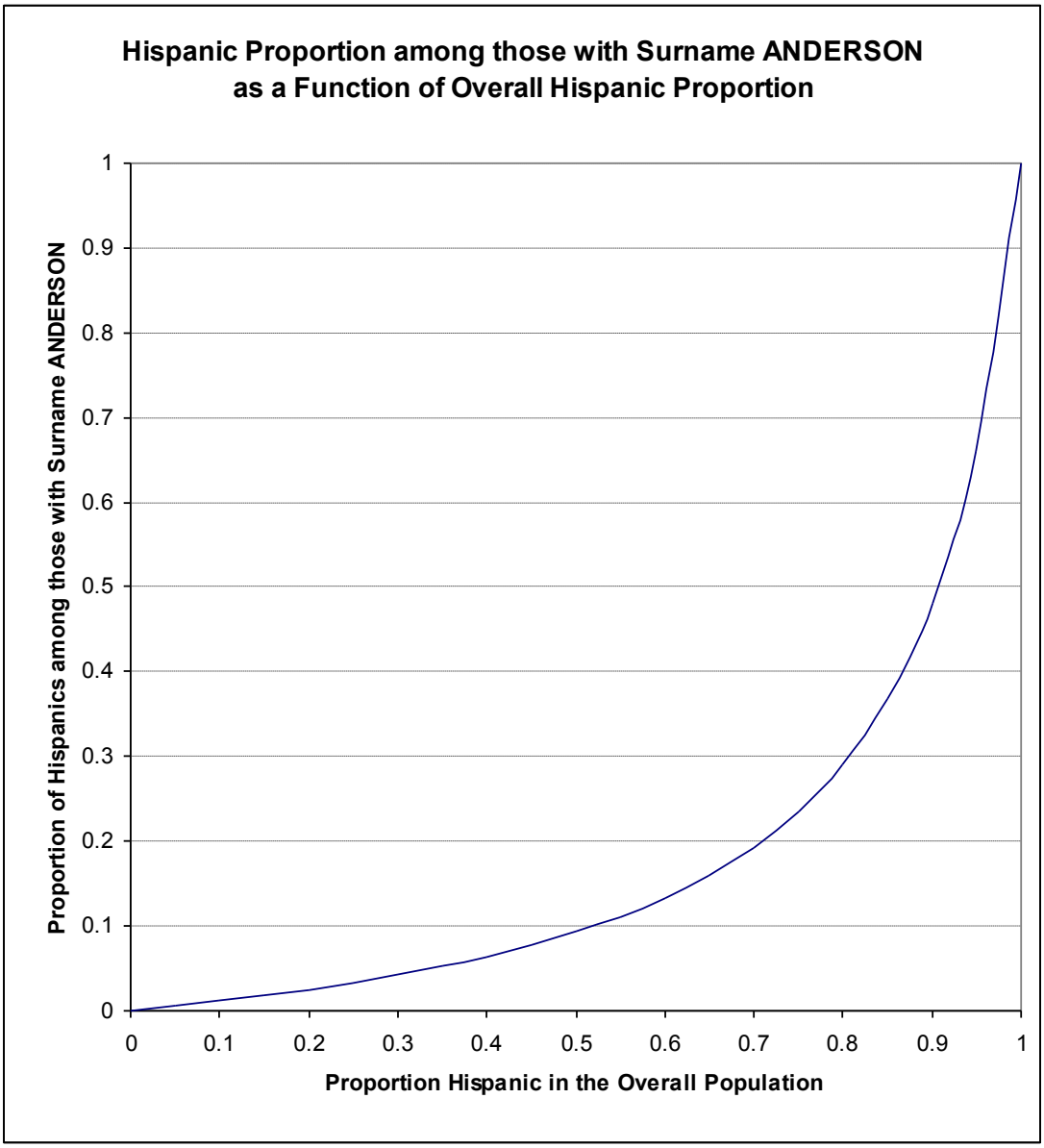
The most heavily Hispanic names in the U.S. contain a very high proportion of all Hispanics; indeed, 87.9% of all Hispanics have one of the 4,310 most Hispanic names. While we have already seen that the set of most Hispanic (in percentage terms) surnames ranges from 34% Hispanic to virtually 100% Hispanic, by Proposition 4, we see that it must also be true that Hispanics make up 87.9% of the set of people with one of these 4,310 names.[18] In fact, they do.

Illustrating Proposition 5

Let us, for illustrative purposes, compare ANDERSON and GARCIA. In the national population, which is 13.43% 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? The graph in

Figure 4 answers that question. What we see is that, in a population that is 0% Hispanic, there are 1/10th as many GARCIAs as there are ANDERSONs. In a population that is 100% Hispanic, there are estimated to be 65 times as many GARCIAs as ANDERSONs. The ratio hits 1 around 12% Hispanic in the population.

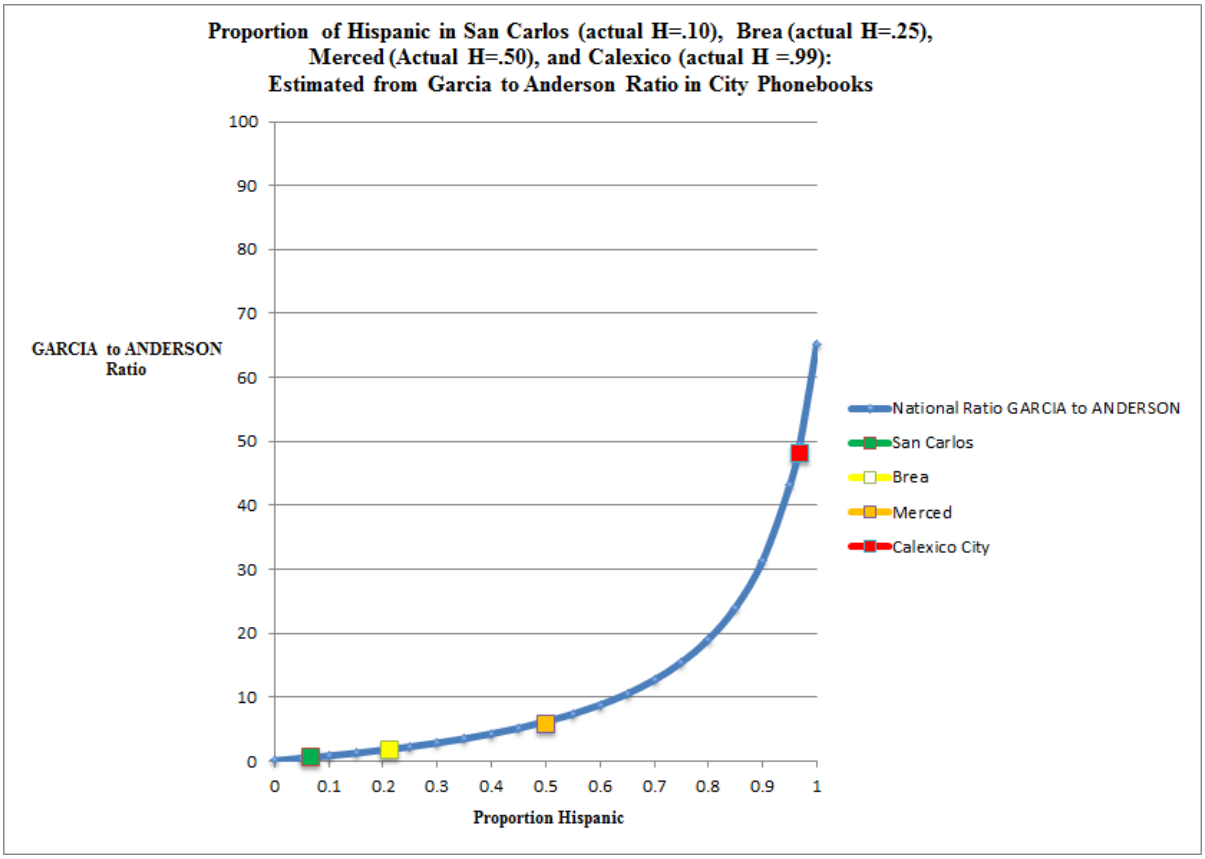
Figure 4



Above we use Figure 4 to show 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 4 what proportion Hispanic in the population would have been expected to give rise to that ratio.

By conducting this type of analysis for selected pairs of surnames[19] we can, we believe, set plausible bounds on the likely proportion Hispanic in the population whose ethnic characteristics we are seeking to estimate. By using this methodology in, say, a range of cities where we have reliable census estimates of their proportion Hispanic, we can test whether or not this approach is feasible. In Figure 5 we present results based on the GARCIA to ANDERSON ratio for four California cities – San Carlos, with a low Hispanic population of 10.1% (ca. 2010), Brea, with a moderately low Hispanic population of 25% (ca. 2010), Merced, with a moderately high Hispanic population with 49.8% (ca. 2010), and Calexico, with a very high Hispanic population of 98.6% (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.

Figure 5



By looking at the phonebook data we get a ratio of GARCIA to ANDERSONs for each of the four Californian cities**.** A ratio of .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 get to the estimated Hispanic population proportion that corresponds to that ratio in the national data. For San Carlos, the

estimated Hispanic proportion is around .067**;** for Brea, the estimated Hispanic proportion is .21; for Merced it is .48; and for Calexico it is .964. Considering (a) that we are projecting values derived from national data into particular cities in California, (b) that the phonebook data suffers from an unknown bias in terms of the relative proportions of Hispanics and non-Hispanics who are too poor to have land lines, and (c) that the phonebook data suffers 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 suffers 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 on-line phone book; it is truly remarkable to have the kind of fit shown here**:** .067 versus .10, .21 versus .25, .483 versus .498, and .963 versus .986.

Of course, we need to be careful about the realism of the assumption that prob(name i |Hispanic) and prob(name i |non-Hispanic) are the same in every population 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. For example, even though there are some non-Hispanic groups, e.g., Portuguese and Filipinos, who have a high incidence of "Hispanic" names, how far off we are going to be in particular applications 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 pairwise ratio method in particular, we have specifically 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

No Census publications about surname matching of which we are aware lays 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. As we have shown, a major problem with the standard use of Spanish surname matching to estimate Hispanic population is its failure to take into account the baseline demography. Since there is no such thing as the conversion rate of surname into Spanish origin self-identification, the likelihood is always context dependent. So, where we draw an optimal cutoff between Hispanic and non-Hispanic surnames will depend upon the nature of the demography in the area we are investigating. In particular, in seeking to balance Type I and Type II error, more names should be classified as Hispanic in areas of high Hispanic concentration.[20] We have also shown some properties of optimal thresholds, such as the relationship obtained under them between the proportion Hispanic among the set of names chosen and the proportion Hispanic in the total sample

Recognizing the potential problems with the usual approach to surname matching, we have also sought to offer an alternative approach to the search for a surname list that will best balance off Type I and Type II errors, using the idea of pairwise likelihood ratios of relative name occurrences**.** In empirical work investigating the applicability of this approach for four

cities in California that vary dramatically in their proportion Hispanic (from 10% to over 98%), we concluded that the very simple device of finding the ratio of GARCIAs to ANDERSONs, and projecting that ratio into the national census data to recover the corresponding Hispanic population, worked remarkably well. 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.

We believe this pairwise approach is one very much worthy of further investigation. In particular, if we need more precision, we can us this method as only a first approximation for the Hispanic proportion of the electorate. Then we would use the estimate derived 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 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: 383-398.

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.

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.

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, Colorado, April 1012.

Perkins, R. Colby, 1993 "Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results." U.S. Bureau of the Census, Population Division. Population Estimates and Projections Technical Working Paper Series, August.

Tversky, A. and D. Kahneman. 1982 "Causality and Attribution." In Kahneman, D., 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.

Endnotes

[1] 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 assitance 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 essay 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. Address feedback to bgtravel@uci.edu

[2] [2] 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 is 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; here, estimates of Hispanic share of the electorate based on surname matching techniques have been presented by demographers and other social scientists in a number of cases over the past several decades (see e.g., *Garza v. County of Los Angeles Board of Supervisors* 918 F. 2d 763 (9th cir. 1990); *Baldus et al. v. Government Accountability Board of Wisconsin*, Federal District Court, Case No. 11-CV-562 JPS-DPW-RMD, decided March 22, 2012.)

[3] For example. Bhavnani (2012) has used official records of election commissions in India to examine the effect of name and caste on voting behavior. To verify eligibility for seats reserved for Scheduled Castes, as well as for information gathering purposes, candidates in elections in India are required to report their status as a member or not a member of a Scheduled Caste (SC). Because Scheduled Caste names tend to be distinctive, Bhavnani has been able to generate an estimate of the likelihood that any given name (here the combination of first name and last name) will be that of someone from a Scheduled Caste. Similarly, Harris (2012) uses data on the surnames common in various ethnic groups to identify the changing ethnic distribution (Kalenjin, Kamba, Kikuyu, Kisii, Luhya, and Luo) 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.

[4] In fact, the most common error in using Spanish surname matching techniques, 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 Kahnemann (1982). We may characterize the Tversky and Kahnemann (1982) example as follows: A subject is told that in a given city 85% of the taxis are Green Cabs (painted green) and the remaining 15% 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 (non-color blind) 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% of the time and made an erroneous classification only 20% 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%. The correct answer, using Bayes Theorem, is that the probability equals .8*.15/ (.8*.15 + .2*.85) = .41. What subjects fail account for is the baseline proportions (.15 and .85) in doing their probability assessments. It is clear that (most) subjects do not really understand the concept of conditional probability.

[5] 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.

[6] As noted in the published research reports by Census Bureau personnel (e.g., Word and Perkins, 1996: i): "These data sets do not violate the confidentiality of Census responses. … The released files contain no subnational geographic data, nor is there any indication of first name or age of respondent."

[7] For example, "[I]n 1980, the Census Bureau published a list of 12,497 different 'Spanish' surnames. The central premise for including a surname on that list was the 'similarity' of that name's geographic distribution to the geographic distribution of the Spanish origin population within the U.S. The 12,497 surnames appearing on the 1980 Spanish surname list were culled from a data base of 85 million taxpayers filing individual federal tax returns for 1977" (Word and Perkins, 1996: 1). An updated list was prepared in 1996, using a different methodology, one that did not rely on geography and the complex methodology previously used for ascertaining which names belonged on the list, but simply linked ethnicity and name. While it was recognized that "the ideal data source for classifying surnames by proportion Hispanic would be the 1990 Census in its entirety," lack of a name code in the permanent Census record motivated the creators of the 1996 list to, instead, "use a very large sample data set that does link name (first and last) to individual 1990 Census records" (Word and Perkins, 1996: 1). This individual record file, which was originally created for the purpose of estimating undercount in the 1990 Census contained 7,154,390 person records, of which 5,609,592 records included both a valid surname and a response to the Hispanic origin question. However, Word and Perkins (1996: 2) opted to reduce the problem of "clustering" in terms of family names by limiting their universe to "the 1,868,781 Householder records that include valid responses to both surname and Hispanic origin. This "householder" data set contains 268,783 distinct surnames—167,765 occurring exactly one time." They refer to this list as a list of "householder" surnames.

[8] See e.g. http://www.family-crests.com/family-crest-coat-of-arms/surnames-7-7/common-spanish-surnames.html. This is a list of 660 names.

[9] 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 data base, 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, from the 1980s redistricting round to the 2010 redistricting round, every applications of Spanish surname matching of which the senior author of this paper is aware involved treating one set of names as if they were 100% Hispanic and all other names as if they were 0% Hispanic.

[10] 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

[11] 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.

[12] 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 the decision in *Garza* 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)

[13] 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.
[14] Recall, however, that all the names in the Census data set we use have at least 300 instances in the national population.

[15] 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 1500 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% 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. In the 2010 census data set, when we look at the correlation between surname count and surname proportion Hispanic, we find it to be -.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% of the Hispanic population in the U.S. as a whole, we would go down from 1500 names to just 113. These names would, on average, be 90.4% Hispanic in the national data set.

[16] Because GARCIA is a surname that has a higher proportion of all Hispanics (.0261) than it is has of non-Hispanics (.0004) among its members, the curve shown in Figure 1 is convex.

[17] 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.

[18] Of course, this equivalence of Hispanic proportion in the name set and proportion Hispanic in the data only holds for the name set which equates Type I and Type II error.

[19] We expect the pairwise ratios to be most predictive of the true Hispanic proportions where both numerator and denominator are common names, e.g., GARCIA and ANDERSON. Looking at ratios involving names that are highly uncommon may not be very helpful 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.

[20] 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 five percentage points, is tolerable. Word and Perkins (1996) simply caution those doing Spanish surname matching that the accuracy of any list varies with geography.