ELECTION LAW JOURNAL
Volume 13, Number 3, 2014

DOI: 10.1089/elj.2013.0190

Using Spanish Surname to Estimate Hispanic Voting Population in Voting Rights Litigation: A Model of Context Effects Using Bayes' Theorem

Bernard Grofman and Jennifer R. Garcia

ABSTRACT

We offer a simple use of Bayes' Theorem to model the relationship between surname and ethnicity in order to improve present expert witness practices in voting rights litigation. Our aim is to show how to better estimate the overall Hispanic share of the electorate to determine realistic opportunity to elect candidates of choice. We show that there is no such thing as *the* proportion of bearers of a given name who are Hispanic. How "Hispanic" any given name turns out to be is a function of the overall Hispanicity of the population, which affects both the distribution of names and the conditional probability that the possessor of any given name will be Hispanic. Because of this, the number of names on a surname list (say that of registered voters) that should be counted as Hispanic is not fixed, but rather varies by demographic context. We show how to identify the optimal size of a surname list by balancing false positives and false negatives. We also provide some "quick and dirty" approximation methods for estimating the size of an optimal surname list. For example, the optimal number of names needed for a national sample, which is 13.4 percent Hispanic, is roughly 4,300 names. Too many names and you overstate Hispanic population; too few and you understate it. This list of 4,310 surnames, rather counterintuitively, includes all surnames whose holders have more than 34 percent probability of self-identifying as of Spanish heritage on the Census. However, we also show that, despite the existence of both false positives and false negatives, ecological inference of racial bloc voting (RBV) patterns using surname-based estimates of the Hispanic share of the electorate at the voting tabulation unit level as the independent variable will usually give us results that are more robust to error in list size than calculations of overall Hispanic levels. In the former case, the two types of error will tend to occur in geographic locations in ways that limit their consequences for the accuracy of RBV estimates.

I. INTRODUCTION

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 of Spanish origin on the Census. Based on the 2010 Census, the Census Bureau has created a publicly available list of common U.S. surnames (names with greater than 300 instances), which also provides the proportion of self-identified Hispanics for each name.[1] 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.

Bernard Grofman is the immediate past director of the Center for the Study of Democracy and the Jack W. Peltason Endowed Chair Professor of Political Science and adjunct professor of economics in the School of Social Sciences at the University of California, Irvine in Irvine, CA. Jennifer R. Garcia is associate staff at the Center for the Study of Democracy at the University of California, Irvine in Irvine, CA.

[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 in Appendix A, Figure A1.

In voting rights litigation, involving issues of Hispanic vote dilution,[2] expert witness testimony often involves surname matching. A list of names, most commonly registered voters, is matched against a list of the most common Spanish surnames derived from Census data,[3] providing an estimate of the Hispanic proportion of the list. When done for a district as a whole, the surname-based estimate is most often used to examine whether or not Hispanics constitute a majority of the (potential) electorate,[4] or have a realistic opportunity to elect candidates of choice in the district (e.g., as a majority of the voters in a major party primary).[5] The other use of surnames among those registered or

[2]There are numerous other applications of name matching, but we will limit ourselves to voting rights applications, and we limit ourselves to the use of name matching to identify Hispanics. For example, Bhavnani (2012) has used official records of election commissions in India to examine the effect of name and caste on voting behavior in India. 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. 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 group 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.

[3]Lists similar to the 2010 surname matching list were generated by the Census Bureau for earlier periods, at least as far back as 1980, but the methodologies to generate these lists have varied 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. Although other surname matchup-lists have been created by various entities (see e.g., a list of 660 names found on <http://www.family-crests.com/family-crest-coat-of-arms/surnames-7-7/common-spanish-surnames.html>), virtually all surname matching done in situations where legal issues are involved has drawn in some fashion on a list prepared by the Census.

[4]The ability to demonstrate the potential for creating an additional district with an Hispanic citizen voting age (CVAP) majority is, under current case law, a necessary condition for having a viable Section 2 voting rights claim. In *Voinovich v. Quilter*, 507 U. S. 146, 154–155 (1993) the Supreme Court held that a "numerical, working majority of the voting age population" was a prerequisite for a Section 2 challenge, a view reaffirmed in *Bartlett v. Strickland*, 556 U.S. 1 (2009). However, until the 2010 redistricting round, CVAP data from the Census was not available until after the time when states needed to complete their redistricting plans, so it thus was very difficult (if not impossible) to use a fifty percent CVAP cutoff to determine the viability of a Section 2 challenge. But, if a majority of the registrants are Hispanic, this provides *prima facie* evidence that Hispanics have a realistic opportunity to elect candidates of choice, since normally the Hispanic proportion of registrants is lower than the Hispanic proportion of CVAP. Currently, after the abolition of the long form of the census which was administered to one household in six, the only direct estimates of CVAP come from the American Community Survey (ACS), which is administered every two years, but to a much smaller sample than the old long form, and estimates of CVAP based on these small samples have some severe problems (Persily, 2011). Using pooled ACS data to determine if there is an Hispanic CVAP majority district leads to complex estimation issues because (a) data is collected at different points in time, with corrections needed for aging of the population and potentially also for population movement, (b) smaller sample size leads to larger sample error and, perhaps most importantly for redistricting purposes, (c) data is not collected at the level of the small units of census geography which must be aggregated to form political districts, thus requiring projections of statewide or countywide or citywide CVAP data into local areas to provide estimates of the conversion rates of Hispanic and non-Hispanic voting age populations into citizen voting age populations (Persily, 2011). These complications introduce estimation error and their accuracy become the basis of considerable (and highly technical) dispute among expert witness, as happened, for example, 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. Thus making use of Spanish surname-based estimates of voter registration (or turnout) to measure the Hispanic share of the electorate, rather than relying solely on CVAP estimates may be desirable.

[5]In *Page v. Bartels*, 248 F.3d 175 (3d Cir. 2001) potential to control the Democratic primary was used as the litmus test for a "realistic opportunity to elect candidates of choice" (see discussion of this concept in Grofman, Handley and Lublin, 2001; see also Grofman 2006).

voting is to estimate the Hispanic share of the electorate at the voting tabulation (polling district) level. These estimates are then used in ecological inference or ecological regression methods as the independent variable to calculate levels of racial bloc voting.

Estimates based on Spanish surname matching of registration or turnout data, if available and used correctly, can provide a better estimate of the Hispanic proportion of the actual *electorate* than the Hispanic population share of the voting age population or even the Hispanic citizen voting age population share (CVAP).[6] Similarly, for ecological inference that requires comparing Hispanic support levels for candidates against the Hispanic proportion across real units such as voting tabulation districts (polling stations),[7] using surname data on registered or actual voters generates estimates of the Hispanic share of the electorate at the precinct level that has been shown, *ceteris paribus*, to be more accurate than the same type of ecological inference based instead on the Hispanic population share (Grofman and Barreto, 2009).[8]

In the redistricting arena, from the 1980s redistricting to the 2010 redistricting, every application of Spanish surname matching, of which the senior 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. The obvious issue with this form of surname matching estimate is that for any list of surnames we pick to identify as "Hispanic" (with all other surnames being treated as "non-Hispanic"), there are both Type I and Type II errors.[9] Type I errors occur when we classify someone with a "Hispanic" surname as Hispanic who is not actually Hispanic, a false positive.[10] Type II errors occur when we classify someone with a "non-Hispanic" surname as non-Hispanic even though that person is actually Hispanic, a false negative.[11]

We take the Census list of names, and the accuracy of their *national-level* data on the proportion of those with a given surname who self-identify as Spanish heritage, as given. The national sample used for the Census' most recent surname matching list is 13.4 percent Hispanic. However, because the populations that are examined in particular voting rights cases almost certainly have a different Hispanic proportion than 13.4 percent, we need to be sensitive to how this difference can affect the estimates of the Hispanic population we derive from surname analysis. A key question, and one that has not adequately been dealt with in the previous literature on surname matching, is how to optimally determine the size of the list of names we use for surname matching purposes.

The size of the surname list we choose matters. The more names on the list, the higher our estimate of the underlying Hispanic proportion; the fewer names on the list, the lower our estimate of the underlying Hispanic proportion. Moreover, the size of the list affects not just how many Hispanics we estimate there to be, but also the magnitude of both Type I and Type II errors. The surname lists that have been used for academic as well as legal purposes have varied greatly in size. 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

[6]If we are trying to estimate the voting behavior of the actual electorate, the closer our estimates of the proportion of a given population are to their actual proportions in the electorate, *ceteris paribus*, the more accurate we will be. For Hispanics, there is a substantial dropoff from overall population to proportion of turnout that comes from the dropoffs from population to voting age population, and then from voting age population to CVAP, and then from CVAP to registration, and then from registration to turnout, and then, quite possibly, from turnout to *roll-on*, i.e., the proportion of those at the polls who cast a valid ballot for a given office. Also see earlier footnote discussing the difficulties of generating reliable estimates of CVAP from the American Community Survey (ACS) data.

[7]See e.g., Loewen (1982) for a discussion of ecological regression, extreme case analysis, and the Duncan-Davis *method of bounds*, and King (1997) for a discussion of ecological inference. The latter technique (or extensions of it) has become the preferred technique of most voting rights experts, but it is often complemented by use of ecological regression, and/or extreme case analysis, and/or the Duncan-Davis *method of bounds*.

[8]Our concern in this article is not to debate the issue of the accuracy of ecological inference techniques (see e.g., Grofman, 1993; Owen and Grofman, 1997; Grofman and Barreto, 2009), but only with how to improve their reliability by basing them on the most appropriate estimates of the independent variable, the relevant Hispanic population.

[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 (more than 50,000 relatively common surnames in the 2010 Census list), to be compared against the names in the data set. Invariably, those doing the matching limit themselves to a smaller and more manageable set of names.

[10]Here, by "actually Hispanic," we simply someone who self-identifies as of Spanish heritage on the Census.

[11]See previous footnote.

a 2012 federal court case, *Baldus v. Wisconsin Government Accountability Board*, also used Census data, but made use of only 639 names.

While surname matching has been used in some academic work on generating ethnic composition estimates for particular units of geography in the U.S. (see e.g., Abrahamse, Morrison, and Bolton, 1994; Barreto, Segura, and Woods, 2004), the vast bulk of the work using this methodology has been for litigation, or done under contract with redistricting authorities in advance of litigation. Because of this, much of the data analysis conducted is treated as proprietary, and only limitedly available in the summary form of expert witness reports that are difficult to find. Consequently, there has been little discussion of statistical issues involving the application of surname matching in the voting rights arena in the social science literature. Moreover, while the Census provides information on surname ethnicity characteristics, and although Census staff have identified smaller subsets of "heavily Hispanic" names, the Census does not offer "best practices" advice on how to make use of this data.[12] Thus, there is no developed theory that indicates how many names should be treated as Hispanic in any given situation using surname matching. It is this gap that we seek to fill here. In this article, we show how to conduct more reliable surname matching. Additionally, we demonstrate the implications of our findings for how the Census surname list should be used by expert witnesses in voting rights litigation.

By using elementary mathematics, based on Bayes' Theorem, we examine the nature of Type I and Type II errors in surname matching. We also provide a straightforward proposition about how to find the optimal cutoff for deciding how many "Hispanic" names to include on a list (with names not on this list being classified as "non-Hispanic"). Key to that calculation is a simple, but not at all well understood point; namely, that there is *no* unique conditional probability that someone with a given name will be Hispanic. Rather, this probability differs with demographic context.

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 which to draw the line between "Hispanic" surnames and "non-Hispanic" surnames is a question that cannot be answered in general. More specifically, we show that the optimal number of (most heavily Hispanic) names to count as "Hispanic" vary with the demographic context in a way that can be specified precisely in terms of exactly balancing off Type I and Type II errors, and we show how/why such context effects are inescapable. We also show that this optimization process requires us to take into account both information about how Hispanic a name is (i.e., the proportion of those with that surname who are Hispanic), and the frequency of this name within the Hispanic population.

We would note that, while our basic results are not mathematically deep, some of their implications are quite counterintuitive. One point that must be understood is that correctly classifying *individuals* as Hispanic or non-Hispanic by looking at their surname, on the one hand, and accurately estimating the *overall proportion* Hispanic in some population (such as registered voters) using surnames, on the other, are different problems and have different solutions. If we wish to maximize predictive accuracy (i.e., minimize the sum of errors made), with respect to the question of which *individuals* are Hispanic, and we wish to treat Type I and Type II errors as equally pernicious, we simply posit that any surname whose holders we believe to be at or above 50 percent Hispanic be classified as "Hispanic" and any surname whose holders we believe to be less than 50 percent Hispanic[13] be classified as "non-Hispanic."[14]

[12]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." Unfortunately this very important 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.

[13]We intentionally use the term "believed to be" because, as we show below the proportion of those with a given surname who are Hispanic is not a fixed parameter but rather varies with demographic context. If we are uncertain as to that context it might be possible to think of this conditional probability as an estimate within certain confidence bounds. That is an issue we hope to explore in future work.

[14]If we do not wish to treat Type I and Type II errors as equally pernicious we can opt to try to minimize a weighted sum of the occurrences of each of these two types of error. Here the weights reflect the relative importance we attach to the two forms of error.

However, we show that *aggregate level* predictive accuracy of the proportion Hispanic in a population requires us to set Type I error equal to Type II error.[15] This optimization strategy may well lead us to treat as 100 percent "Hispanic" some surnames that have less (perhaps even much less) than a 50 percent Hispanic share. Or, it may lead us to treat as "non-Hispanic" some surnames whose holders are more (perhaps even much more) than 50 percent Hispanic as non-Hispanic. In other words, optimizing overall accuracy may involve (deliberately) getting a lot of the individual level predictions wrong—in both directions. For example, we show that if you apply the usual name-matching tool to the national data to get the Hispanic proportion right, you should classify all names that are least 34 percent Hispanic as if they were 100 percent Hispanic and count all names that are less than 34 percent Hispanic as if they were zero percent Hispanic.[16] Moreover, when you do so, you make seven million errors in individual classification, and wrongly estimate the Hispanicity of almost one in eight Hispanics. But, that is still the best thing to do, since the estimate you derive of proportion Hispanic in the sample is accurate to four significant figures.

To better understand the difference between optimizing predictive accuracy at the individual level and predictive accuracy at the aggregate level, a simple example is useful. Consider the data shown in Appendix Table A1 from a hypothetical universe where there are only five different surnames. These five names vary both in the proportion of those who have the name and the proportion of holders of those names who are Hispanic.

In this example, the population is 70 percent Hispanic. If we classify all surnames whose members are more than 50 percent Hispanic as "Hispanic," and all surnames whose members are less than 50 percent Hispanic as "non-Hispanic," we will erroneously identify 13 percent of the sample as Hispanic who are not Hispanic (Type I error), and three percent of the sample as non-Hispanic who are in fact Hispanic (Type II error). Given the limited information at our disposal, this is the best we can do. In *toto*, we erroneously classify 16 percent of the sample, judged in terms of *individual level classifications*. Note, however, that the number of Type I errors and the number of Type II errors are far from identical.

Now imagine that we classify names one and two as "Hispanic" and the remaining three names as "non-Hispanic," despite the fact that 70 percent of those who hold name three identify as Hispanic. In this case, we estimate the Hispanic population in the sample as 70 percent, which is exactly right. Of course, by classifying name three as "non-Hispanic" we make seven errors per 100 names. So, our errors in individual classification rise from three percent to seven percent for this name, thus increasing our total misclassifications at the individual level from 16 percent of the sample to 20 percent of the sample. But now half of that 20 percent is in the form of Type I errors and half in terms of Type II errors. Therefore, the number of Type I errors exactly equals the number of Type II errors.

Next we turn to an explication of how Bayes' Theorem allows us to derive such a rather surprising result about optimal surname matching.[17] We then consider several different ways in which pooling information about more than one surname may allow us to develop approximations or bounds on the Hispanic proportion in the list we are evaluating.

II. MODELING THE RELATIONSHIP BETWEEN SURNAME AND "HISPANICITY" IN DIFFERENT DEMOGRAPHIC CONTEXTS

Before we can say anything about how large of a list of surnames to use in the absence of reliable information about the proportion Hispanic in the population of interest, we need to explain why equalizing Type I and Type II error rates is desirable. This is the goal of the first subsection below. After we offer that intuition, we then turn in the

[15]Note that equalizing Type I and Type II error as the optimizing strategy is very different from the usual situations involving Type I and Type II errors, e.g., medical diagnostics, or jury decision making, or decisions to "stop and frisk," where optimization requires minimizing one or the other type of error, or some weighted function of the two.

[16]Yes, that really is 34.

[17]Bayesian ideas are briefly mentioned in some Census publications without a specified model or empirical analyses, and some expert witnesses in the 1990s considered a Bayesian approach to Spanish surname analysis but dropped it after the 1990 Circuit Court decision in *Garza v. Board of Supervisor of Los Angeles County*, because it appeared that, in that litigation, federal courts had accepted the validity of simply using the 12,000+ census Spanish surname list (personal communication, Kenneth McCue, October 2012).

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*	*Prop Hispanic*	*Prop of all population*	*Rank on overall surname frequency (n = 53,286)*	*Prop of all Hispanics*	*Rank on prop Hispanic*	*Prop 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

Prop, proportion.

succeeding subsection to explain Bayes' Theorem, and use it to show how we can model the changes in the conditional probability that the holder of a given surname is "Hispanic" as a function of the overall Hispanicity of the sample.

Calculating an optimal size of the Spanish surname list of names: Balancing Type I and Type II errors

From information compiled from the 2010 Census name and Spanish origin data set, 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.[18] For each surname, we provide both raw counts and percentage data. We also provide conditional probabilities both of the likelihood that, in this data set, a given name is "Hispanic" ("non-Hispanic") and of the likelihood that a Hispanic (non-Hispanic) has a given name. That is to say, 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 data set, its proportion of all the Hispanics in the data set, and its proportion of all the non-Hispanics in the data set. And we also provide some data on where a given surname ranks with respect to various characteristics of the national data.

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 SAGERO are about 60 times more likely to be Hispanic than are those named ANDERSON.[19]

As noted earlier, for practical reasons of manageability, 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 in the larger data set (or sample) that is being used. 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.

When we make Type I errors we overestimate the Hispanic population; when we make Type II errors we overestimate the non-Hispanic population. For example, if we were to classify those with last

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

[19]In the national data set, the relationship between how numerous is a surname and how likely it is for individuals with that surname 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 first effect would lead us to expect a positive correlation between overall surname count and surname proportion Hispanic; the second would lead us to expect a negative correlation. 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 to capture 50% of the Hispanic population, from over 1500 to just 113. These names would, on average, be 90.4% Hispanic. We consider the option of looking at names that are both common and highly Hispanic in a subsequent section.

name YEPIZ as "non-Hispanic," which is the 640th most-Hispanic name of the list compiled from the Census sample data, we might expect to commit a Type II error over 94 percent of the time. On the other hand, such errors would be rare because YEPIZ, though a heavily Hispanic surname, is a rare surname even among Hispanics. Thus, it would seem that any analytic method to find an optimal cutoff must combine information both about how Hispanic a name is (i.e., the proportion of those with that surname who are Hispanic), and how likely such a name is to be found in the Hispanic population and the population as a whole.

However, as we will see, if we set the threshold appropriately as to what names to include, then the mistakes (Type I errors, false positives) we make by including non-Hispanics in the set of names we assign to the category "Hispanic" will (precisely or almost precisely) equal the mistakes (Type II errors, false negatives) we make by including too many Hispanics in the set of names we assign to the category "non-Hispanic." But that is true only for the optimal threshold, and our intuitions about where that threshold should be set can be quite misleading, as shown in the examples below.

For any real world data set, not all surnames are equally likely and the Hispanic and non-Hispanic populations are not identical in size. So, how do we find the optimizing threshold for any particular distribution? We first illustrate several methods using our national data set, and then consider how to generalize from these approaches to a distribution whose Hispanic population is not known.

To be able to take into account the different population shares and the different name distributions of Hispanics and non-Hispanics in our data set, after sorting the data from most Hispanic to least Hispanic, we create a graph with two lines. One is a line showing the percent of total population (or the raw number of people) whom we wrongly classify as Hispanic as we lower the threshold for the degree of Hispanicity needed to classify a surname as "Hispanic"—which occurs as we move the threshold to the right. The second is a line showing the percent of total population (or the raw number of people) whom we wrongly classify as non-Hispanic as we raise the threshold for the degree of non-Hispanicity needed to classify a surname as "non-Hispanic"—which also occurs as we move the threshold to the right. For any x value (i.e., any surname), the y value on the first of these lines gives us the cumulative amount of Type I error (non-Hispanics wrongly classified as Hispanic) if we classify all those with surnames to the left as "Hispanic" and all those with surnames to the right as "non-Hispanic." For any x value (i.e., any surname), the y value on the second of these lines gives us the cumulative amount of Type II error (Hispanics wrongly classified as non-Hispanic) if we again classify all those with surnames to the left as "Hispanic" and all those with surnames to the right as "non-Hispanic."

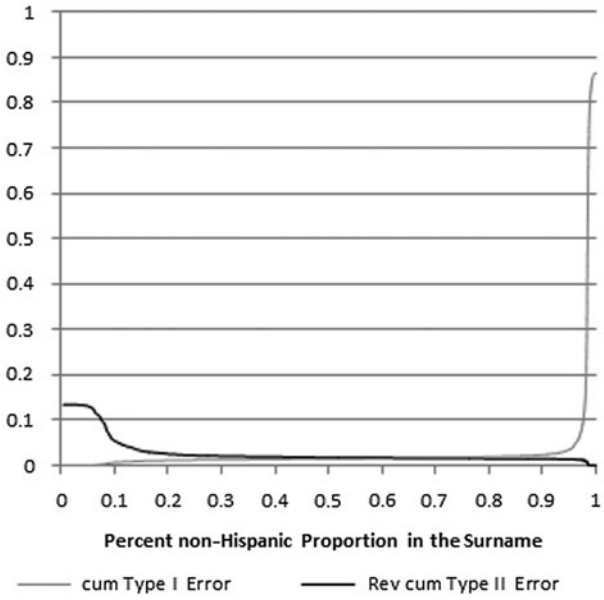

FIG. 1. Equalizing Type I and Type II errors by Hispanic proportion in the surname.

The point where the two lines intersect in Figure 1 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, i.e., classify all names above this threshold as Hispanic and all names below this threshold as non-Hispanic, then we will be correctly identifying the "true" Hispanic population proportion. For instance, 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, if we succeed in picking the optimal threshold, we will obtain a value of roughly 13.43 percent.[20]

[20]We may not get the exact answer to more than several significant figures because of what we might call "lumpiness" effects, i.e., for any given name, with a certain number of instances of that name observed in the data, we must classify the name (and all its instances) as either Hispanic or non-Hispanic.

Figure 1 shows the intersection point where Type I and Type II errors are equalized in terms of the proportion non-Hispanics (Hispanics) among holders of the surname. Here we find that the intersection point of the two lines is at roughly 66 percent non-Hispanic (i.e., at 34 percent Hispanic). What this means—and we think the finding more than a little counterintuitive—is that the optimizing strategy for Hispanic surname matching with a national sample is to count all names that are at least 34 percent Hispanic as if they were 100 percent Hispanic and to count all names that are less than 34 percent Hispanic as if they were 0 percent Hispanic. In other words, we are including in the "Hispanic" category some names that are much less than 50 percent Hispanic.

Alternatively, we could show the intersection in terms of the proportion of surnames (figure omitted for space reasons). Here we find that the intersection point of the two lines is at about the eighth percentile of all surnames. Since there are 53,286 surnames in the database, this means that we should treat somewhat over 4,000 names (4,310 to be precise), the ones that are most Hispanic in their percentages, as "Hispanic," and treat the remaining surnames as 100 percent "non-Hispanic." Thus, we need only check the data set for the frequency of occurrence of these 4,310 names to accurately estimate the Hispanic proportion in the data. Of course, both of these methods must give us the same answer. That is, when we use the 4,310 names which are the most heavily Hispanic, this set consists of those names that are 34 percent or more Hispanic in the national data set.

When we put the cutoff at the 4,310th name (VARON), we over-count 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, both Type I and Type II, but these errors are cancelling out. Remarkably, but not surprisingly, 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 (i.e., the same fraction as the proportion of Hispanics in the data).

We note that these individual level mistakes, more than seven million of them, equal 24.2 percent of the total Hispanic population in our sample, though the Type I error rate is only 3,606,488/29,863,836 (= 12.1 percent). In other words, to get the correct aggregate proportion of Hispanics nationally using the standard surname matching method requires us to *mis*classify almost one in eight Hispanics.

In order to make sense of this result, we remind the reader that 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 weighted average) of Type I and Type II errors. For aggregate optimization purposes, how many (what proportion of) individuals we wrongly classify is essentially irrelevant. It is acceptable, 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.

This leads us to our first major analytic finding. We state this result informally. Formal notation and proofs of this proposition, and the two others stated in the article, are given in the Mathematical Appendix, which can be found on the senior author's website.[21]

> *Proposition 1:* If we array names from most "Hispanic" to least "Hispanic," and we treat the first s names as 100 percent "Hispanic" and the remaining names (from the (s + 1)th to the Nth) as "non-Hispanic," to optimize the predictive accuracy of the cutoff (size of the surname list, s, classified as Hispanic) we find s so that the number of false positives equals the number of false negatives.

There is a further analytic proposition that we can derive, which is a bit less intuitive.

> *Proposition 2:* If we array names from most "Hispanic" to least "Hispanic" and we treat the first s names as 100 percent "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 average Hispanic share of the population among the names from the first to

[21]The Mathematical Appendix can be accessed at <http://www.socsci.uci.edu/~bgrofman/>.

the sth name equals the proportion of the total Hispanic population constituted by the people with those surnames.

The meaning of Proposition 2 may not be immediately apparent, so an example should help. The most heavily Hispanic names in the U.S. contain a very high proportion of all Hispanics. Indeed, 87.9 percent of all Hispanics have one of the 4,310 most "Hispanic" names. But the "optimal" set of most "Hispanic" (in percentage terms) surnames ranges from 34 percent Hispanic to virtually 100 percent Hispanic. By Proposition 2, we see that it must also be true that Hispanics make up 87.9 percent of the set of people with one of these 4,310 names.[22] In fact, they do. This equivalence only holds for the optimal value of the threshold, s.

In principle, the analytic results above show us how to calculate the optimal size of a Spanish surname list. However, there is a critical problem with making use of the idea of equating Type I and Type II error. As we show below, using Bayes' Theorem, the probability that someone with a given surname is actually Hispanic varies with the demographic context. If we knew the demographic context, then we could specify that probability. We could then use that information to decide on the optimal size of a Spanish surname matching list. But, if we truly knew the demographic information, we would not need to be using surname matching to estimate the Hispanic proportion in the first place. Nonetheless, before we seek to address this "chicken and egg" problem, we need to spell out *exactly how* the probability that someone with a given surname is actually Hispanic varies with demographic context. This we do in the next two subsections. Then, in the final subsection, we consider ways to plausibly approximate the right answer to the Hispanic demographic context.

Using Bayes' Theorem to show how demographic context affects the Hispanicity of any surname

To reconstruct, via Bayes' Theorem, the proportion Hispanic of .0176 shown for the surname WIST, we need to know the proportion of all Hispanics in the sample that come from that name, the proportion of all non-Hispanics in the sample that come from that name, and the proportion of Hispanics in the population. The first two of these numbers are given in Table 1. The third we know to be .1343 for our Census data set.

The basis of Bayes' Theorem is the identity

$$\text{prob(Hispanic}|\text{name i)}^{*}\text{prob(name i)} = \text{prob(name i}|\text{Hispanic)}^{*}\text{prob(Hispanic)}$$

From this identity we derive the equation

$$\text{prob(Hispanic}|\text{name i)} = (\text{prob(name i}|\text{Hispanic)}^{*}\text{prob(Hispanic))}/\text{prob(name i)}$$

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 that

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

This last equation, which we will label Equation 1, is the most familiar form of Bayes' Theorem for a dichotomous variable. Substituting in the values from Table 2, along with the given overall Hispanic population proportion of .1343, we get

$$\text{prob(Hispanic}|\text{WIST)} = .0176 = \frac{0.000000235 * 0.1343}{0.000000235 * 0.1343 + 0.00000203 * .8657}$$

But, 1.76 percent is the correct answer only when we have a 13.43 percent Hispanic base population.

What would happen, if the Hispanic base population were, instead 50 percent? If the share of overall Hispanic population and the share of overall non-Hispanic population coming from the name WIST does not change, all we need do is substitute .50 for .1343 in Equation 1. Doing so, we get prob (Hispanic|WIST) = .1038. When we increase the population proportion of Hispanics, both the numerator

[22]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.

and the denominator of Equation 1 change, thus changing the ratio. Indeed, we can readily calculate that when the population is 50 percent Hispanic, the likelihood of WIST being a Hispanic name is estimated to be roughly six times higher than what we find for the national data, where under two percent with the surname WIST are Hispanic. When the Hispanic population rises, even names with only few Hispanics now have a higher Hispanic proportion, since there are more Hispanics and fewer non-Hispanics with that surname. Still, not until the Hispanic population is around 85 percent will the proportion Hispanic among those with the surname WIST be expected to be as high as half. Of course, when a population is 100 percent Hispanic, all the WISTs (if any are present) are also Hispanic, just as when a population is zero percent Hispanic, all the WISTs in that population are non-Hispanic.[23]

We show in Figure 2 how, under the given assumptions, the Hispanic proportion of those with the surname WIST changes with the overall proportion of the population that is Hispanic. The calculations that generate this figure come from Equation 1 above and the data for WIST in Table 1. This figure clearly shows that there is not a single parameter value for the proportion of those with the surname WIST who are Hispanic.

What is true for WIST is true for all surnames. For example, because GARCIA has such a high ratio of the first of these two conditional probabilities to the second, even when there are relatively few Hispanics in the overall population, GARCIA is a surname with a substantial proportion of bearers who are Hispanic. 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 (figure omitted for space reasons).[24]

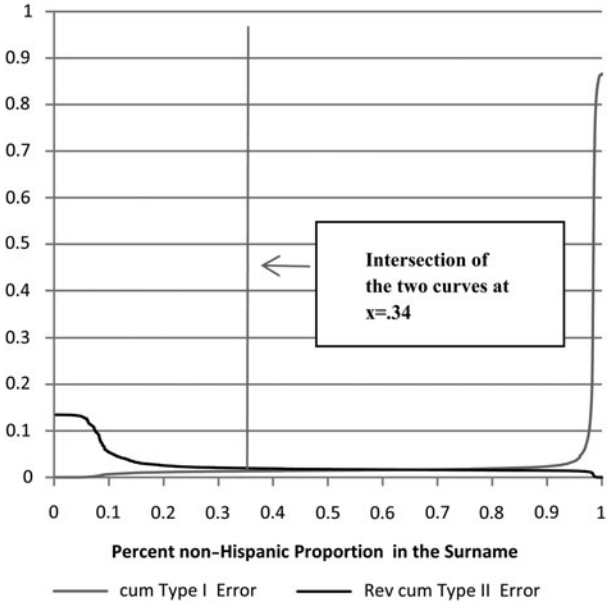

FIG. 2. Hispanic proportion by surname.

Showing how the accuracy of estimate varies with the number of surnames we use for the matching

We have shown that, for our national data subset, with a 13.43 percent 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 percent "Hispanic," those 4,310 surnames are held by a set of individuals who together constitute 13.43 percent of the national population (i.e., the actual proportion). But, what happens if we use a smaller (or a larger) number of surnames to estimate the national Hispanic population via surname matching? If we were to use only the top 639 Hispanic percentage surnames in our data set, we would estimate the national Hispanic population to be only 8.39 percent, missing more than a third of all Hispanics.[25] If we were to use the top 8,000 names in Hispanic percentage we would estimate the national Hispanic population to be 15.56 percent, enhancing the estimated Hispanic population to be about 115 percent of what it is in actuality. If we were to use 12,497 names for 2010 data, which is the number most often used in studies done in the 1980s, we would estimate the national Hispanic population to be 18.19 percent, about 135 percent of what it actually is. As we increase

[23]But, as the population grows more heavily Hispanic, the proportion of those with the surname WIST will decline. See discussion below.

[24]The omitted figure generates a curve that is convex, rather than the concave curve shown in Figure 3. The curve in Figure 3 is concave because WIST is a surname that has a lower proportion of all Hispanics than it is has of all non-Hispanics among its members.

[25]The name list of 639 names used in expert witness testimony in *Baldus* is different from the top 639 names in the 2010 list we have sorted according to percentage Hispanic among those with the given name, since it largely excludes rare but overwhelmingly Hispanic names.

the size of the name list our inaccuracy grows. For example, for 20,000 names, we would estimate the national Hispanic population at around 23 percent, coming close to double the true proportion (figure omitted for space reasons). Thus, at least when Hispanic population is low, using more names is not necessarily better.

As noted above, the 4,310 result shown in Figure 2 applies only to the entire national data set, where we have a 13.43 percent Hispanic population overall. The question still remains as to how the number of (most Hispanic) needed to equalize Type I and Type II errors changes with the proportion Hispanic in the population. In Table 2, we show 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 non-Hispanic populations as is true in the 2010 national data set.

We see from Table 2 that 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 relatively small set of surnames. In contrast, we need many names to accurately assess the Hispanic population proportion when that proportion is high. We can formalize the intuition that leaps out at us from Table 2 in the form of a very general proposition.

> *Proposition 3:* As the size of the Hispanic population increases, the optimal name list must also grow in size.[26]

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)

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

What this means, in practice, is that it is a bad idea to use small surname list in areas where there are many Hispanics, since that list will certainly underestimate the Hispanic population. On the other hand, as we have seen, using a large surname list when there are few Hispanics will overestimate the Hispanic population share.

Approximating the optimal size of the Spanish surname list

We have previously called attention to the "chicken and egg" problem associated with the need of taking into account the demographic context before we can specify the size of the surname list needed to exactly equalize Type I and Type II errors in the population of interest. Here we suggest three different ways to provide an approximate answer to this question, and thus be able to "triangulate" by picking a surname list size that is reasonably appropriate for the initially estimated Hispanic proportion in any given list.

The first and most obvious of these is to find an upper bound on the Hispanic population proportion in the unit or units (whole political jurisdictions, particular districts, the entire set of voting tabulation units) for which we need to generate estimates of the Hispanic proportion of the electorate, simply by looking at data on Hispanic population or Hispanic voting age population. In voting rights litigation, such data is either directly available, or can be constructed by matching census geography with the political geography whose behavior we wish to examine.

The second approach involves using the identity shown in Proposition 2 to give us a "quick and dirty" way to put lower bounds on the Hispanic population in a given data set by looking at a small set of overwhelmingly "Hispanic" surnames and the proportion of the total Hispanic population that those names can be expected to constitute.

We show in Table 3 the ten most common "Hispanic" surnames nationally circa 2010, and the proportion of the total Hispanic population for each.

[26]Because the exact numerical results depend upon both the distribution of names in the population and the conditional probability of the holder of any given name being Hispanic, the proposition is stated in very general terms since we do not have a way to state it in the form of "an increase in the size of the list of k% yields an increase in aggregate accuracy of classification of z%."

Table 3. Ten Most Common Highly Hispanic Surnames

Surname	*Count of population (n = 222, 316,554)*	*Count of Hispanics*	*Count of non-Hispanics*	*Prop Hispanic*	*Prop of all population*	*Rank on overall surname frequency (n = 53,286)*	*Prop of all Hispanics*	*Rank on prop Hispanic (n = 53,286)*	*Prop of all non-Hispanics*
GARCIA	858,289	779,412	78,877	0.9081	0.0039	8	0.0261	1,533	0.0040
RODRIGUEZ	804,240	745,530	58,710	0.927	0.0036	9	0.0250	1,031	0.0039
MARTINEZ	775,072	710,896	64,176	0.9172	0.0035	11	0.0238	1,302	0.0037
HERNANDEZ	706,372	662,648	43,724	0.9381	0.0032	15	0.0222	741	0.0034
LOPEZ	621,536	568,768	52,768	0.9151	0.0028	21	0.0190	1,357	0.0030
GONZALEZ	597,718	561,795	35,923	0.9399	0.0027	23	0.0188	689	0.0029
PEREZ	488,521	447,729	40,792	0.9165	0.0022	29	0.0150	1,322	0.0023
SANCHEZ	441,242	404,972	36,270	0.9178	0.0020	33	0.0136	1,285	0.0021
RAMIREZ	388,987	364,364	24,623	0.9367	0.0017	42	0.0122	773	0.0019
TORRES	325,169	296,424	28,745	0.9116	0.0015	50	0.0099	1,447	0.0015
TOTAL	6,007,146	5,542,538	464,608	0.9227	0.0249		0.18564		0.0288

Prop, proportion.

In this very small subset of only ten names that are both heavily Hispanic, with a mean Hispanic percentage of 92.3 percent, and relatively common, with all of the ten names ranking from eighth place to fiftieth place among the 52,386 names on our list, we find nearly 18.6 percent of all Hispanics have one of these ten names. If we count the number of people in a data set who have these names, we can first compensate for the limited number of names in our subset by taking into account (dividing by) the proportion of all Hispanic names that those names comprised. For the national sample, we would take 6,007,146, and divide by .18564, to get 32,359,114. But, we also need to correct for the fact that not all of those with these ten names are Hispanic. So, our next step is to multiply 32,359,114 by the mean Hispanic proportion in the set of ten surnames. For the national sample, we know that this is .9227. So, we would reproduce the national Hispanic number as 32,359,114 * .9227 = 29,857,754, which is identical to the total Hispanic population in our sample (29,863,836) to within four significant digits. [27] Now, to estimate the Hispanic percentage, we simply divide 32,359,114 * .9227 by the total population. In the national sample, this number is 222,316,554, and so we recreate the true Hispanic percentage of 13.43 percent.

For a sample of unknown Hispanicity, as long as the Hispanic proportion is non-trivial (say, above 10 percent), we can expect that the Hispanic share of these 10 surnames will be at least 90 percent, since these are very heavily Hispanic surnames.[28] If we do not know the true proportion of Hispanics in our surname list, we can use .90 as a lower bound on the proportion Hispanic among these 10 surnames, and then look at how common these surnames are in the list of names whose Hispanicity we are examining. If we did this for our national data, using .90 rather than .9227, we would be estimating the population as .90/.9227 its actual Hispanic share. This translates into a lower bound for the percentage Hispanic in the sample of 13.1 percent, which may be compared to the real answer of 13.4 percent. But, we can also use these same 10 names to give us an upper bound by assuming that, in very highly Hispanic areas, these names are 100 percent "Hispanic." Now we would multiply 13.4 by 1.00/.9227, to get a value of 14.5 percent Hispanic among the names on our list. The true value lies in between. The upper bound we get using the first of these two approaches can be compared with the upper and lower bounds we get from the second approach.

However, we do not have to stop there. Once we have these bounds on our estimate of the Hispanic population in the list, we can derive upper and lower bounds on the size of the optimum list of surnames to classify as "Hispanic" by inspection of Table 2 (or doing calculations from the formula used to construct that table). By then applying surname lists much longer than only ten names, where the number of names chosen reflects the upper and lower bounds we derived from Method 2 and/or Method 1, we can expect to improve the accuracy of our (bounds on

[27] Any imprecision occurs because we did not carry out some of our arithmetic to enough significant figures.

[28] See our earlier discussion of Bayes' Theorem. Areas that are less than 10 percent Hispanic are unlikely to be involved in voting rights litigation affecting the Hispanic community.

the) estimate of proportion Hispanic in the list because we are reducing sampling error by using a larger set, and plausibly chosen, of names.

There is, however, a third way to generate an approximation to the demographic context that we can use to specify a plausible surname list size. This method, original to the present authors (described in more detail in Grofman and Garcia, 2014), like the second method, involves combining data on more than one surname, but in this method we look at ratios of pairs of names rather than averages of several.

For any given proportion Hispanic in a list, using Bayes' Theorem, and taking the national proportions, as shown in Equation 1, we can find the proportion of Hispanics among, say, all ANDERSONs, a common, but primarily "non-Hispanic" surname, and we can find the proportion of Hispanics among, say, all GARCIAs, a common, but primarily "Hispanic" surname. If we then observe a given ratio of GARCIAs to ANDERSONs in a population, the proportion Hispanic in the population that would have been expected to give rise to that ratio can be calculated from repeated applications of Equation 1, and some algebra. We show in Figure 3 the result of calculations done by combining the information from the calculations using Bayes' Theorem for each of the two surnames.

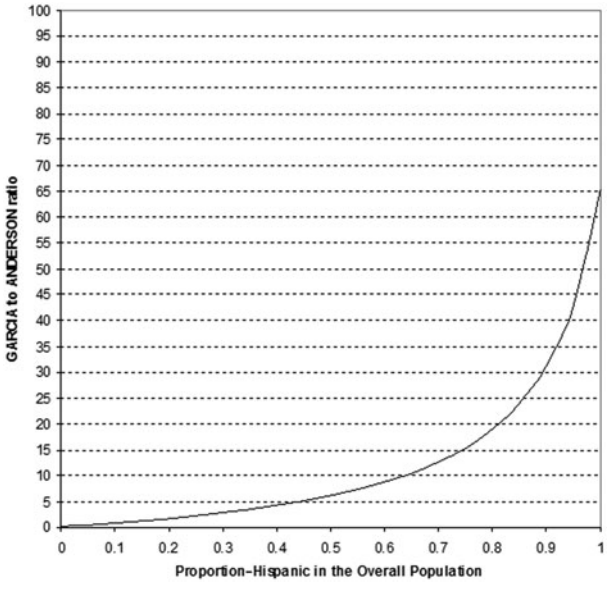

FIG. 3. Surname ratios. Ratio of those in the population with surname GARCIA to those with surname ANDERSON as a function of overall Hispanic proportion.

What we see from Figure 3 is that, in a population that is zero percent Hispanic, there are 1/10th as many GARCIAs as there are ANDERSONs. While in a population that is 100 percent Hispanic, there are estimated to be 65 times as many GARCIAs as ANDERSONs. The ratio hits one around when the Hispanic population reaches about 12 percent of the population. In a like manner, we could calculate the relative sizes of any two surname populations. By conducting this type of analysis for appropriately selected pairs of surnames, and plugging the ratio we obtain into Figure 3 (or its equivalent for the expected ratio values for other surname pairings) to find the Hispanic population proportion that would be expected to give rise to that ratio, we can, we believe, set plausible bounds on the likely proportion Hispanic in the population whose ethnic characteristics we are seeking to estimate.[29]

III. DISCUSSION

We began our article with the claim that the proportion Hispanic for a given surname was not a single quantity, but changes with context. Even for those who understand basic probability theory, it is very easy in a real world situation to fail to process the relevant information in the appropriate way. Unfortunately, as we noted previously, no Census publications about surname matching of which we are aware clearly lay out exactly how prob(Hispanic| name i) can be expected to vary with prob(name i|Hispanic) and prob(Hispanic) in the data set.[30] Furthermore, there does not appear to be an academic article that does so clearly either. As we have

[29]In follow-up work to this article we are empirically testing the ratio approach, and comparing it to the results we get from the two other approaches identified above, using surname data from California cities with a wide range of Hispanic populations (Grofman and Garcia, 2014).

[30]The closest that we have found are Passel and Word (1980) and Word and Perkins (1996). The former suggests 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.

shown, a major problem with the standard use of Spanish surname matching to estimate the Hispanic population is its failure to take into account the baseline demography.[31] 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 when dichotomizing for purposes of surname matching will also depend upon the nature of the demography in the area we are investigating.

Simply understanding that the optimal cutoff (size of the name list to be used) is based on setting Type I errors equal to Type II errors (Proposition 1) is a useful contribution to the literature on surname matching as it has been applied in the voting rights arena. This is especially true since it is so basic to understanding the logic of surname matching, and appears to be not really understood. Furthermore, Proposition 2, which is much less intuitive, shows that in order to create the most accurate surname list for any given population, the average Hispanicity of the names we choose to label as "Hispanic" must be the same as the proportion of all Hispanics who have that surname. Proposition 3 shows that, to maximize predictive accuracy, as the size of the Hispanic population increases, the optimal name list we choose must also grow in size.

Our results have direct implications in two different domains of voting rights litigation: demonstrating whether or not minorities in a given district are of a sufficient size to have a realistic opportunity to elect candidates of choice and using ecological inference methods to estimate racial bloc voting.

First, our results demonstrate that if we fail to take into account baseline proportions of Hispanics and non-Hispanics, we can misestimate the overall Hispanic share by using an inappropriately sized list of surnames. In his testimony as an expert witness for the Hispanic plaintiffs in *Baldus*, Professor Kenneth Mayer (2011) made use of a list with only 639 names when estimating the proportion of registrants who were Hispanic.[32] As Professor Mayer acknowledged in his trial testimony, his classification method would have classified many of the Hispanic candidates and office-holders in Milwaukee, the jurisdiction in question, as non-Hispanic because their names did not appear on the relatively short list of surnames he used. Thus, his classification scheme would appear to lack face validity.

However, our concern with this testimony in *Baldus* does not lie with a failure to correctly predict whether or not particular individuals are or are not Hispanic. As we have shown, aggregate level accuracy may be maximized even though we wrongly predict a substantial proportion of all individual cases (see previous discussion). Rather, our concern stems from the inadequate attention given to what proportion of the surnames on the Census list should be treated as "Hispanic." Had Professor Mayer's testimony about the proportion Hispanic among registrants been a key factor in the outcome of this litigation, the trial court would have based its conclusions on erroneous findings, since the actual Hispanic proportion among registrants was much higher than he estimated.

But Professor Mayer also indicated in his trial testimony that he had not relied on this surname analysis of registration lists in reaching his conclusion that the district created by the State of Wisconsin under challenge from plaintiffs would not

[31]The intuition as to why this must be so is connected to the "blue cab, green cab" probability misassessments called attention to by Tversky and Kahnemann. We may characterize the Kahnemann and Tversky (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 that did it 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 to do is to take into account 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.

[32]While the name list of 639 names used by Professor Mayer is different from the top 639 names in the 2010 list we have sorted according to percentage Hispanic among those with the given name, any set of only 639 names will still underestimate the Hispanic proportion *in heavily Hispanic areas* if we take these 639 names as 100 percent Hispanic and all other names as zero percent Hispanic. On the other hand, since the Hispanic proportion in our national sample is only slightly above 13 percent, 639 names can actually overstate the proportion Hispanic at the national level, if we omit low frequency names.

actually offer Hispanics an equal opportunity to elect candidates of choice. Instead he relied on his analysis of estimated citizen voting age population in the proposed new district, and on the outcomes of hypothetical elections using real election data "projected" into the proposed new district. Experts for both sides in the case agreed that the new district would not contain an Hispanic citizen voting age majority, and the trial court held this fact dispositive.[33] Thus, the erroneous registration calculations based on too small a list of surnames did not affect the court's decision in the case. [34] But it might well in others.

The second area of voting rights litigation to which our results are relevant is evaluating whether or not errors that are generated by the presence of both Type I and Type II errors cause substantial problems for ecological inference analyses used for racial bloc voting analyses based on surname analysis (i.e., ones using surname estimates as their independent variable to reflect Hispanic share of the electorate). Here our work has somewhat surprising implications.

Consider first ecological inference or ecological regression methods. If we use a single list of names to infer Hispanicity in every ecological unit, and as before, treat names as either zero percent or 100 percent "Hispanic," it is clear that we will be making mistakes everywhere, except in the ecological units that have a 13.4 percent Hispanic population. Regardless of where we put the cut off for what counts as a "Hispanic" surname, there will be error in the estimates of the independent variable, the Hispanic proportion in the precinct. In areas that are highly non-Hispanic, both the "Hispanic" and the "non-Hispanic" surnames will be less "Hispanic" than they are on average. On the other hand, in areas that are highly Hispanic, both the "Hispanic" and "non-Hispanic" surnames will be more "Hispanic" than they are on average.[35] Thus, errors will be most common where they are least important. Moreover, the errors that we get will tend to be counterbalancing in terms of their effects on our ecological estimates of differences between the groups in their voting patterns regardless of whether we use Goodman's ecological regression approach or methods of ecological inference derived from the work of Gary King (1997).

Now let us consider the implications of miscalculating the optimal cutoff for homogeneous case analysis of racial bloc voting patterns and for understanding turnout differences between Hispanic and non-Hispanic voters, where we infer the behavior of Hispanic voters from the most heavily Hispanic precincts and the behavior of non-Hispanics from the most heavily non-Hispanic precincts. Here, if we use a single list of names to infer Hispanicity in every ecological unit, then almost certainly we will again be understating the Hispanic proportion in the most heavily Hispanic units and overstating the Hispanic proportion in the most heavily non-Hispanic units. But, if we treat the most homogenous units as accurately reflecting the true means for the two groupings, the fact that there are both Type I and Type II errors means that the units are somewhat more homogenous than they might first appear, and thus homogenous case analysis is more accurate than we might have thought it to be.

---

[33]While experts for both sides agreed that the proposed district did not as of the Census data collection have a majority Hispanic citizen voting age population, there was dispute about how rapidly the district might become majority Hispanic in CVAP over the course of this decade.

[34]In the view of one of the present authors, who testified for the state in *Baldus*, had the registration data, especially that for those who registered as Democrats, been correctly analyzed, and had the legal standard applied been that in use in *Page v. Bartels*, 248 F.3d 175 (3d Cir. 2001), where potential to control the Democratic primary was used as the litmus test for a realistic opportunity to elect candidates of choice, the outcome in *Baldus* might have been different. Under the time pressures of litigation, however, no actual analysis of Hispanic share of Democratic registrants was presented by experts for the state, and assertions to the effect that Hispanics would be a clear majority of the Democrats in the new district and could confidently be expect to select the Hispanic incumbent who would be running in the reconfigured district as the Democratic nominee, and that this incumbent would still easily carry the new district because it remained heavily Democratic in registration were, in the absence of any hard data, (quite reasonably) dismissed by the trial court as merely speculative.

[35]The nature of the context specific errors using surnames has resemblance to errors generated by not correcting Hispanic CVAP, VAP, or registration share for differential turnout rates among Hispanics and non-Hispanics (Owen and Grofman, 1997). These errors, too, will vary with demographic context. Also, we might expect that instances of intermarriage among Hispanics and non-Hispanics was more likely to be found in areas of either low Hispanic population proportion or areas that were more ethnically/racially mixed, which would reduce the importance of the contextual effect when we did ecological inference, since the most heavily Hispanic areas would tend to have a similar ethnicity among marriage partners. And, of course, increasingly, professional women retain their maiden name. A full exploration of the link between errors in estimating the underlying Hispanic population and errors in ecological inferences based on those estimates is, however, well beyond the scope of the present essay.

Thus, despite the existence of both false positives and false negatives, we suggest that ecological inference of racial bloc voting (RBV) patterns, using surname-based estimates of Hispanic share of the electorate at the voting tabulation unit level as the independent variable, will usually give us results that are more robust to error in list size than surname-based calculations of overall Hispanic proportions. In the former case, the two types of error will tend to occur in geographic locations in ways that limit their consequences for the accuracy of RBV estimates using ecological regression, ecological inference, or of estimates of RBV derived from homogeneous case analysis.[36] This leads us to assert that understanding the implications of Bayes' Theorem as applied to surname matching actually increases our confidence in RBV results reported by competent expert witnesses in voting rights litigation of the past 50 or so years.[37]

To be clear, we wish to emphasize that, in the paragraphs above, we are not making claims about the overall accuracy of ecological inference, but only about the relative reliability of surname-based analyses as opposed to population based ecological analyses. It is not a goal of this article to propose something *superior to* ecological inference, or to critique ecological methods, or evaluate their accuracy (see, however, Grofman, 2000; Grofman and Barreto, 2009; Owen and Grofman, 1997). Rather we seek to improve the accuracy of existing tools for ecological inference by improving the estimate of the independent variable, proportion Hispanic, used in the analyses.

Of course, to apply any of the surname matching techniques we have described above to a population of unknown Hispanicity requires us to act as if the conditional probabilities found in the national sample that give us the proportion of Hispanics who share any given surname, $p(name_i|H)$, would also hold, at least approximately, in the list under scrutiny. However, this approximate constancy assumption is not nearly as strong as the assumption standard in the existing surname matching literature that the conditional probability of someone with a given surname being Hispanic, $p(H|\ name_i)$, is a constant across all subpopulations, regardless of that subpopulation's demographic characteristics. We know this assumption to be false. In contrast, assuming that the name distribution among Hispanics is the same in all Hispanic populations (subject only to sampling error), while a strong assumption, is one that can be checked empirically. And in those instances where we have good reason to believe it false (e.g., among Cuban Americans), we can create surname lists that are appropriately tailored.

We have begun work to assess the plausibility of this assumption for a set of California cities varying considerably in their Hispanic population. Despite all the possible complicating factors,[38] for the several cities we have studied to date, taking $p(name_i|H)$ from nationwide Census data and then looking at surname lists from those cities appears to work well in allowing us to very accurately

[36]The senior author of this article derived this insight over two decades ago, when working with the demographer William O'Hare in the case of *Garza v. County of Los Angeles Board of Supervisors*, 918 F. 2d 763 (9th Cir. 1990). In that case, with the assistance of Robert Kengle of the Civil Rights Division of the U.S. Department of Justice, we obtained a special run of data from the Census for portions of Los Angeles County, coded at relatively small units of tabulation, that had counts of those who described themselves in each unit as of Spanish heritage on the Census. By running surname estimates for those same units based on lists of registered voters we could see how errors in surname mathching varied across units of differing ethnic demography.

[37]As we have seen, to get the surname-based estimate right, we need fewer names in the less Hispanic areas and more names in the more Hispanic areas. Thus, despite the implications of our remarks above, that using the "wrong" cutoff matters less for purposes of racial bloc voting analysis than it does for estimating the Hispanic proportion of some larger unit, such as a district, in dealing with ecological units of varying levels of likely Hispanicity, it would, in principle, be better to provide surname estimates of the independent variable (e.g., Hispanic registration) by using a list of names of variable length whose cutoff value (list size) is sensitive to the demographic context. But, in practice, the marginal gains in accuracy might be more than outweighed by the costs and complexity of these more refined analyses.

[38]First, 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. Second there can be intermarriage across ethnic lines. What intermarriage rates do is cause the potential for both Type I and Type II errors to increase, the latter error occurs where an Hispanic women marries a non-Hispanic man and takes her husband's name; the latter where an Hispanic man marries a non-Hispanic woman who takes his name. Third, the distribution of non-Hispanic names may also vary with the proportion Hispanic, e.g., black or Asian American populations may be more likely to be proximate to populations with high concentrations of Hispanics and these groups have a surname distribution that is different from the general non-Hispanic population. Finally, there are some non-Hispanic groups, e.g., Portuguese and Filipinos, who have a high incidence of "Hispanic" names, making it difficult to apply standard surname lists in areas where there a large number of members of these groups (Passel and Word, 1980; 10–11). But how far off we are going to be in particular applications is a matter that can only be studied empirically.

derive estimates of the Hispanic proportion. Thus, for California, where we have a predominantly Mexican American Hispanic population, as is true for the country as a whole, many of these potential confounds simply prove not that important. For example, groups like Filipinos have only minuscule ($<2\%$) proportions in the California cities we have looked at, too small to affect our estimates.[39] While it is still quite premature to evaluate the usefulness of the three approaches discussed in the article in terms of their being able to provide quick but surprisingly accurate estimates of Hispanic populations, our preliminary empirical work using the ratio method (Grofman and Garcia, 2014) is very encouraging.[40]

One last point: While this essay deals directly only with Spanish surname matching, identical issues arise with respect to name matching for other U.S. ethnicities (e.g., Asian Americans: Abrahamse, Morrison, and Bolton, 1994). The methods (and cautionary notes) we provide are general ones that have applicability to any type of name matching, and are not restricted to the U.S.[41]

ACKNOWLEDGMENTS

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 (UCI) 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 Department of Justice; other related work was done 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.

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.

Duncan, Dudley and Beverley Davis. 1953. An Alternative to Ecological Correlation. *American Sociological Review* 18: 665–666.

Goodman, Leo. 1953. Ecological Regression and the Behavior of Individuals. *American Sociological Review* 18(6): 663–664.

Goodman, Leo. 1959. Some Alternatives to Ecological Correlation. *American Journal of Sociology* 64: 610–625.

Grofman, Bernard. 2000. A Primer on Racial Bloc Voting Analysis. In Nathaniel Persily (ed.) *The Real Y2K Problem: Census 2000 Data and Redistricting Technology*. New York: The Brennan Center for Justice, New York University School of Law.

Grofman, Bernard. 2006. Operationalizing the Section 5 Retrogression Standard of the Voting Rights Act in the Light of *Georgia v. Ashcroft*: Social Science Perspectives on Minority Influence, Opportunity and Control. *Election Law Journal* 5(3): 250–282.

Grofman, Bernard and Matt A. Barreto. 2009. A Reply to Zax's (2002) Critique of Grofman and Migalski (1988): 'Double Equation Approaches to Ecological Inference When the Independent Variable is Misspecified.' *Sociological Methods & Research* 37(4): 599–617.

Grofman, Bernard and Jennifer Garcia. 2014. Using Spanish Surname Ratios to Estimate Proportion Hispanic via Bayes' Theorem. Working paper, University of California, Irvine Center for the Study of Democracy.

Grofman, Bernard, Lisa Handley and David Lublin. 2001. Drawing Effective Minority Districts: A Conceptual Framework and Some Empirical Evidence. *North Carolina Law Review* 79:1383–1430.

Grofman, Bernard N., Michael Migalski, and Nicholas Noviello. 1985. The 'Totality of Circumstances' Test in Section 2 of the 1982 Extension of the Voting Rights Act: A Social Science Perspective. *Law and Policy* 7(2): 209–223.

[39]And, as noted above, if there were important subpopulation distributions in the commonness of surnames we could almost certainly compensate for them, e.g., by developing a more appropriate surname list.

[40]Further empirical investigations of the accuracy of these approximation techniques, take us well beyond the primarily theoretical scope of the present article, and must be left to future research.

[41]See footnote 2.

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

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

Loewen, James. 1982. *Social Science in the Courtroom*. Lexington MA: Lexington Books.

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.

Owen, Guillermo and Bernard Grofman. 1997. Estimating the Likelihood of Fallacious Ecological Inference: Linear Ecological Regression in the Presence of Context effects. *Political Geography* 16(8): 657–690.

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.

Persily, Nathaniel. 2011. The Law of the Census: How to Count, What to Count, Whom to Count, and Where to Count Them. *Cardozo Law Review* 32(3): 755–789.

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.

Address correspondence to:
Bernard Grofman
Center for the Study of Democracy
School of Social Sciences
University of California, Irvine
Irvine, CA 92697

E-mail: bgtravel@uci.edu

APPENDIX A

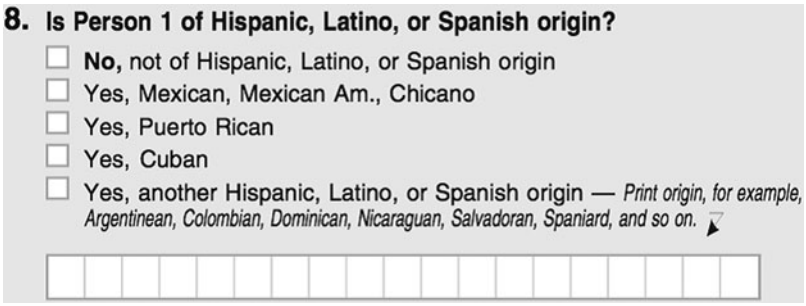

FIG. A1. Spanish Origin Census Question.

TABLE A1. HYPOTHETICAL UNIVERSE WITH ONLY FIVE SURNAMES

	Surname i percentage of all voters	*Percentage Hispanic of those with surname i*	*Type I errors if use >50% rule*	*Type II errors if use >50% rule*	Combined Errors
Name 1	40	90	4		0.36
Name 2	30	80	6		0.24
Name 3	10	70	3		0.07
Name 4	10	20		2	0.02
Name 5	10	10		1	0.01
TOTALS	100		*13*	*3*	*0.7*