SURNAME ANALYSIS FOR ESTIMATING LOCAL CONCENTRATION OF HISPANICS AND ASIANS

Allan F. Abrahamse
Peter A. Morrison
Nancy Minter Bolton

P-7844

Published 1993 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

SURNAME ANALYSIS FOR ESTIMATING LOCAL CONCENTRATION OF HISPANICS AND ASIANS*

by

Allan F. Abrahamse
Peter A. Morrison
The RAND Corporation

and

Nancy Minter Bolton
Bolton Associates

ABSTRACT

Surname analysis is a potentially useful technique for identifying members of particular racial, ethnic, or language communities within a population. This paper reviews the existing state of the art for identifying persons of Hispanic or Asian origin, based on surnames distinctive of each group. We discuss the logic of surname analysis, describe several available surname dictionaries, and illustrate their applications in local redistricting. Results of our ongoing validation studies suggest promising future directions for improving accuracy and broadening applications.

---

*Revised version of paper presented at the annual Population Association of America Meetings, Cincinnati, April 1, 1993. Views expressed here are the authors' own, not necessarily those of RAND.

I. INTRODUCTION

Surname analysis is a technique demographers can use to identify the members of particular racial, ethnic, or language communities within a population. Insofar as a particular surname belongs uniquely to a particular racial or ethnic group, it is possible to identify its holder's probable membership in the group by using well-formulated surname dictionaries. Several such dictionaries, in varying stages of refinement, now exist for such applications.

Surnames appear in many computerized administrative files of licensed drivers, registered voters, business owners, vital records, airline passengers, paramedic emergency responses, and so forth. Using surnames to identify groups has a variety of potential applications:

- pre-identifying potential members of certain communities--for example, every business owner who *might* be Korean, for mailing out a screening survey, or every licensed driver likely to be Vietnamese, for evaluating the selection of jurors;
- estimating the absolute size of a community--for example, the number of Spanish-surnamed registered voters, for deciding how many political flyers to print;
- comparing average emergency response times--for example, how quickly paramedics respond to "911" emergency calls for Hispanic and non-Hispanic persons.
- gauging relative yield--for example, the share of all flyers vs. first-class ticket buyers with Japanese surnames;
- estimating the comparative magnitude of potential markets--for example, for new life insurance policies among Hispanic and non-Hispanic newlyweds, identified by surnames on marriage certificates.

Applications turn up occasionally in the media, as when *Time* magazine reported the most common surname among new homebuyers in Orange County, California, to be the Vietnamese surname "Nguyen".

Our own interest in this admittedly esoteric subject originated in a pragmatic need: to estimate how many of the registered voters in a jurisdiction are Hispanic. As our interests broadened to Asian nationalities, we discovered new possibilities but also inherent limitations. For example, "Kim" is a uniquely Korean surname; about 22 percent of all Korean-Americans are named Kim (Shin and Yu, 1984). Thus, the "Kims" of the world offer an encouraging headstart in spotting Korean registered voters. "Lee", by contrast, is the surname of many Chinese persons; yet many persons named Lee are not Chinese.

This paper reviews the existing state of the art for identifying persons of Hispanic or Asian origin within a population based on surnames distinctive of each group. We discuss the logic of surname analysis, describe several available surname dictionaries, and illustrate their applications based on our experience with local redistricting. Drawing on the results of our ongoing validation studies, we suggest further avenues for improving accuracy and broadening applications.

BACKGROUND

The law designates protected minorities by such labels as "Hispanic" or "Asian and Pacific Islander." The populations themselves, of course, are far from monolithic: they consist of various ethnic and national subgroups that may differ politically, socially, and culturally. These populations generally are identified using census data, which precisely measure their local size and composition decennially.

In the years after a census, it is often necessary to gauge the presence of particular racial, ethnic, and language communities within a locality. Jury commissions concerned with minority representation inquire about the relative concentration of, say, Hispanics among the pools of registered voters and licensed drivers from which jurors are drawn. The potential electoral strength of, say, Asians can be gauged by observing how their demographic presence translates into a presence among the voters who turn out in an election. Making that observation

entails comparing the percentage of actual voters with Asian surnames to the corresponding percentage of all eligible registered voters.

USEFUL SURNAME DICTIONARIES

The Census Bureau has developed and tested several dictionaries, the most familiar of which is its list of 12,497 Spanish surnames for identifying persons of probable Hispanic origin (Passel and Word, 1980). That list has well-documented detection characteristics (Perkins, 1993), and we have used it in numerous applications to gauge the makeup of local California electorates. In Section II ahead, we present two representative illustrations.

No comparably refined surname list exists for the Asian population, although the Census Bureau conducted exploratory work to develop one (Passel et al., 1982) and later extended that effort in applications designed to pre-identify minority business owners (U.S. Bureau of the Census, 1988).

A related effort by Donnelley Marketing Information Systems (DMIS) also extends the Bureau's earlier research. The genesis of the DMIS Asian Surname File was an early Census Bureau list of 1,010 Asian surnames. DMIS refined and expanded that list using complementary information about (1) Asian-like surnames of the residents of census block groups known to be inhabited predominantly by Asian households, and (2) the association between surnames and people known to be born in Asia. This combination of "Asian-neighbored" and "Asian-born" information provided the basis for enlarging the original list of 1,010 surnames to the current DMIS list of 9,814 surnames, many of which are further associated with one or more specific Asian nationalities.

The Appendix furnishes additional technical detail on these lists.

LIMITATIONS OF SURNAME ANALYSIS

The use of surnames to identify people's origins has inherent conceptual and technical limitations. First, not all Spanish-surnamed (or Asian-surnamed) persons are of Spanish (or Asian) origin; conversely, not all persons of Spanish (or Asian) origin necessarily bear Spanish (or Asian) surnames. Relying on a list of names to infer

nationality, then, exposes one to two types of error: (1) "false positives," e.g., classifying a non-Asian person as "Asian" because his or her surname happens to be on the list; and (2) "false negatives," e.g., classifying an Asian person as "non-Asian" because his or her surname is not listed.

Second, members of one Asian nationality--Filipinos--often have Spanish surnames. Filipinos with such names as Bacerra, Delmundo, Escober, Hermoso, or Quines would be misidentified as Hispanic using the Census Bureau's Spanish surname list.

Third, an Asian (or Spanish) woman may relinquish her maiden Asian (or Spanish) surname or acquire a distinctive surname from her husband. Alternatively, either or both spouses may elect to hyphenate their surname (e.g., "Fong-Torres"), thereby confounding precise classification.

Fourth, particular surnames can be highly misleading in some contexts. Persons with the common surname "Lee," for example, are likely to be Korean or Chinese if they reside in a predominantly Asian neighborhood but not if they live in, say, Williamsburg, Virginia. Likewise, the Asian surname "Ohara" could easily misidentify persons living in predominantly Irish neighborhoods.

A surname, then, is an inexact and sometimes misleading identifier of its bearer's race or nationality. In statistical terms, though, surnames can potentially identify a high proportion of the members of a certain group within the general population.

II. APPLICATIONS TO LOCAL REDISTRICTING

Electoral redistricting undergoes stringent legal scrutiny with respect to minority voting rights. Legislative bodies elected in single-member districts must be closely attentive to how electoral opportunities are created or obstructed by the election district boundaries they authorize. Surname analysis is directly applicable to these concerns: It affords an objective way to gauge certain facets of electoral makeup from computerized records of registrants and voters. In this section, we illustrate two such applications.

1. GAUGING A GROUP'S POTENTIAL VOTING STRENGTH

On occasion, it is necessary to confirm that the actual voting strength of a "protected group" in a proposed election district exceeds some level. Although census data precisely measure the size and composition of a district's voting-age citizen population, the electorate consists of the registered voters (or, more specifically, those who actually turn out) within that voter-eligible population. A particular group's electoral strength, then, is partly a function of how fully its demographic presence translates into a presence among the district's voters.

We use surname analysis to gauge the potential voting strength of Hispanics in proposed election districts. Classifying those surnames can indicate the approximate makeup of local electorates, revealing the extent to which Hispanics' demographic presence is an electoral presence as well. Gauging that electoral presence can confirm the attainment of minority electoral opportunities or indicate how large a gap remains.

For example, it was necessary to gauge the potential voting strength of Hispanics in two newly-rebalanced Salinas City council districts proposed for adoption (Table 1). Districts 1 and 2 were intentionally designed to concentrate Hispanic voting strength, and each had a predominantly Hispanic voting-age population (81 percent and 73 percent, respectively). The unanswered question was whether a council district's majority-Hispanic appearance translated into an

actual majority of present-day electors, given that many voting-age Hispanics in Salinas are non-citizens ineligible to vote. The County Registrar of Voters furnished a publicly-available computerized file of the names of all registered voters, by voting precinct. Surname analysis of these voter names (aggregated up to council districts) demonstrated that Hispanics were indeed a majority of present-day electors in Districts 1 and 2: Spanish-surnamed persons constituted 60 percent and 54 percent, respectively, of all registrants as of 1991. Clearly, Hispanics constituted a majority of the voting-age *citizens* in each district.

Table 1

Indices of Hispanic Voting Strength in Proposed New City Council Districts: Salinas, 1990

COUNCIL DISTRICT	% of Population Hispanic		% of Registrants Spanish Surnamed
	All Ages	18+	
1	84.9%	81.2%	60.4%
2	77.9%	73.0%	53.7%
3	36.1%	31.0%	15.1%
4	33.1%	29.1%	17.7%
5	28.6%	25.1%	17.1%
6	42.8%	37.8%	27.0%
TOTALS	50.6%	44.4%	26.0%

SOURCES: Census of Population and Housing, 1990: PL 94-171 data; Official June 1991 voter registration tapes.

2. CLARIFYING CANDIDATE SUPPORT

Surname analysis also can yield important insights into the ethnic composition of the voters in particular election precincts. The City of Oxnard, CA, affords an illustration.

During the 1980s, an Hispanic candidate ran for city office but lost repeatedly. He challenged Oxnard's at-large system of electing city councilmembers, claiming that the system itself blocked Hispanics in general from getting elected. Surname analysis enables us to evaluate this assertion empirically.

The surnames of all persons who had voted in one particular election were coded and tabulated to gauge the composition of the electors in each of Oxnard's 66 voting precincts. In that election, the Hispanic candidate had lost in every single precinct. Surname analysis of the actual voters revealed that in several precincts, the vast majority of persons voting had Spanish surnames. Clearly, then, this candidate had failed to attract the votes of electors who were predominantly Hispanic. In the very same election, though, another Hispanic candidate (running for a different office) won in 65 of the 66 precincts. Among the 65 were several precincts where the vast majority of voters were non-Spanish surnamed. This other Hispanic candidate, in contrast, clearly had attracted the votes of many non-Hispanics (as well as Hispanics).

Here, surname analysis helps clarify the underlying realities about a city's at-large election system and its capacity to elect Hispanics. It documents that one Hispanic candidate could in fact attract voter support in mostly non-Hispanic precincts, thereby winning, whereas a different Hispanic candidate could fail to attract voter support in mostly Hispanic precincts, thereby losing.

III. VALIDATION STUDIES OF ASIAN SURNAME DICTIONARIES

Research applications of surname analysis like those just illustrated depend on validation studies to establish how the dictionaries perform in different local contexts.[1] In our experience, detection rates may vary widely and systematically at the census-tract level, depending on the target group's relative size and the fraction who are foreign-born. For Asian nationalities especially, detection rates are heavily dependent on neighborhood context. Common Asian surnames (e.g., "Lee" and "Park") can generate excessive numbers of false positives in non-Asian neighborhoods where most residents with such surnames are not Asian.

This section reports selected findings from our own exploratory studies of two Asian surname dictionaries: the Census Bureau's dictionary, developed for the purpose of pre-identifying small business owners who *might* be Asian; and the Donnelley Marketing Information Systems applications-oriented dictionary, aimed at identifying Asians without incurring excessive false positives or false negatives. These two dictionaries exemplify not only distinctive purposes but distinctive approaches to constructing surname dictionaries--notably the DMIS approach, which exploits neighborhood context information (see Appendix for details).

Our evaluation studies compare both dictionaries, using California death certificates for 1989. The focus of our inquiry is on the detection of Asian decedents (1) to gauge how accurately these surname dictionaries can detect the racial/ethnic identity of decedents officially identified as Asian on the certificate itself; and (2) to elucidate the detection characteristics of each dictionary and its attendant limitations. Our results to date are revealing, although hardly generalizable beyond California.

---

[1] The most recent national validation study of the Spanish surname list is reported in Perkins (1993). We know of no prior validation studies of the DMIS Asian surname list.

Our analysis is based on 219,182 death certificates, shown in Tables 2 and 3. Both tables compare the Census Bureau's pre-identification-oriented Asian surname dictionary and the DMIS estimation-oriented Asian surname dictionary. We have tested two variants of each dictionary, defined as "broad" and "narrow." The broad version casts a wide net (e.g., by including "Lee" despite the many false positives that result). The narrow version is smaller and more focused on those fewer names with the highest probability of being Asian. For both the Census Bureau and DMIS dictionaries, the "broad" variant yields more false positives and fewer false negatives than its "narrow" counterpart; that is by design.

The data in Tables 2 and 3 show the raw numbers used to compute the analytic measures appearing in Table 4. These analytic measures show how completely the "broad" and "narrow" versions of each dictionary detect decedents within the target universe (i.e., those reported as Asian on the death certificates):

- *Detection.* Each surname list detects a large proportion of this target universe. The Census Bureau's "broad" pre-identification surname dictionary, for example, detects 80.5 percent of the target universe (i.e., 5,212 of the 6,478 Asians in Table 2). The DMIS "broad" estimation-oriented dictionary, by comparison, detects 76.2 percent.
- *False Positives.* The percentage of all "positives" that are false indicates the "false positive penalty" accompanying these rates of detection. For the Census Bureau's "broad" dictionary, the data show that 36.8 percent (3,031 of 8,243) of Asian-surnamed decedents are outside the target universe. For the DMIS "broad" dictionary, only 17.0 percent (1,010 of 5,946) of Asian-surnamed decedents are outside the target universe.
- *False Negatives.* Each surname list also fails to detect some proportion of the target universe. The percentage of all "negatives" that are false indicates the "false negative penalty" associated with these rates of detection. For the Census Bureau's "broad" dictionary, the data show that 0.6 percent (1,266 of 210,939) persons with non-Asian surnames

Table 2

Relationship Between Decedent's Recorded Ethnicity and Surname Classification

(California Death Certificates, 1989)

A. CENSUS DICTIONARY (Broad)

Decedent Classification	Surname classification		
	Asian	Not Asian	Total
Asian	5,212	1,266	6,478
Other	3,031	209,673	212,704
TOTALS	8,243	210,939	219,182

B. DMIS DICTIONARY (Broad)

Decedent Classification	Surname classification		
	Asian	Not Asian	Total
Asian	4,936	1,542	6,478
Other	1,010	211,694	212,704
TOTALS	5,946	213,236	219,182

Table 3

Relationship Between Decedent's Recorded Ethnicity and Surname Classification

(California Death Certificates, 1989)

A. CENSUS DICTIONARY (Narrow)

Decedent Classification	Surname classification		
	Asian	Not Asian	Total
Asian	3,146	3,332	6,478
Other	792	211,912	212,704
TOTALS	3,938	215,244	219,192

B. DMIS DICTIONARY (Narrow)

Decedent Classification	Surname classification		
	Asian	Not Asian	Total
Asian	4,672	1,806	6,478
Other	531	212,173	212,704
TOTALS	5,203	213,979	219,182

Table 4

Comparative Analytic Measures

Measure[a]	Census Dictionary		DMIS Dictionary	
	Broad	Narrow	Broad	Narrow
Detection Rate	80.5%	48.6%	76.2%	72.1%
False Positive Rate	36.8%	20.1%	20.5%	10.2%
False Negative Rate	0.6%	1.5%	0.7%	0.8%
Estimation Ratio	127	60.7	91.8	80.3

[a]See text for definitions.

Table 5

Percent of Asian Surname Population Accounted for by Most Frequently Occurring Names: 1989 California Death Certificates

No. of names (most frequently occurring)	Cumulative Percent of Asian Surname Population			
	Census Dictionary		DMIS Dictionary	
	Broad	Narrow	Broad	Narrow
100 names	47.8%	44.2%	54.8%	50.2%
200 names	59.0%	57.5%	67.0%	63.5%

were, nonetheless, Asian. For the DMIS "broad" dictionary, the data show that 0.7 percent (1,542 of 213,236) persons with non-Asian surnames were, nonetheless, Asian.

Table 4 also shows what we term the "estimation ratio," defined as the ratio of all Asian-surnamed persons to all Asian-recorded persons on a per-hundred basis. This ratio can be used to estimate the actual number of Asians in a population per hundred Asian-surnamed members of that population. For example, the DMIS "narrow" dictionary would count 5,203 Asians in a population containing 6,478 Asians (for an estimation ratio of 80.3).

The following points are noteworthy about these dictionaries. First, we note that the Census dictionary is more skewed than is the DMIS one toward false positives. That is as expected, since the Census dictionary was designed to pre-identify Asians, not count their numbers accurately. Accordingly, this dictionary casts a very wide net so as to reduce false negatives.

Second, the DMIS "narrow" variant registers the lowest false positive rate and close to the lowest false negative rate. These detection characteristics may be a function of Donnelley's "Asian-neighborhood" and "Asian-born" approach to constructing its dictionary (see Appendix for technical details.)

Third, our analysis of the sources of false positives revealed that the 322 occurrences of the surname "Lee" account for 11 percent of 3,031 false positives generated by the Census "broad" variant and 32 percent of 1,010 false positives generated by the DMIS "broad" variant. More generally, a select few names ("Lee," "Young," and "George") account for 26 percent of Census "broad" false positives.

Table 5 compares the extent to which each dictionary's detection is concentrated among that dictionary's most frequently occurring surnames. The 100 most frequently occurring Asian surnames on the Census lists detect 47.8 percent (broad) and 44.2 percent (narrow) of all Asian-surnamed decedents. The DMIS lists detect 54.8 percent and 50.2 percent, respectively. Here, again, it appears that Donnelley's

approach to building its lists has realized more of the intrinsic potential for common Asian surnames to detect Asian decedents.

IV. CONCLUSIONS

Death certificates provide a snapshot of older generations, so our results cannot be generalized fully to any contemporary population (e.g., all Californians today). What our findings elucidate are the particular strengths and weaknesses of two underlying approaches to building surname dictionaries and the detection characteristics of each list in its "broad" and "narrow" variants:

1. The Census Bureau's dictionary for pre-identifying Asians achieves that purpose best: The "broad" variant of this dictionary pre-identified the largest fraction (80.5 percent) of the target universe.
2. The DMIS dictionary's "narrow" variant has detection characteristics well suited to the applications-oriented objectives that motivated its design. It identified a large fraction (72.1 percent) of the target universe and incurred comparatively few false positives.
3. The underlying approach DMIS used to build its Asian surname dictionary is well conceived and holds considerable future promise. This approach uses knowledge of neighborhood context to condition inferences and form an expanded list of potential Asian surnames.

A "context-sensitive" approach is especially promising, given the tendency for ethnic groups to cluster residentially. Its logic might be extended in several directions. One is to tailor dictionaries to particular regions or localities with distinctive concentrations of ethnic minorities (e.g., Vietnamese in Southern California). Another is to build surname dictionaries that are ethnically more specific, capitalizing on the known residential concentrations of certain groups (e.g., Salvadorans, Cambodians, Vietnamese). DMIS has taken initial steps in this direction, and further validation studies like ours will be needed to extend this direction of development.

Looking to the future, applications of surname analysis will likely expand and broaden as 1990 Census data grow stale and as interest in ethnic identity and availability of computerized surname files increase. In the public sector, the salience of ethnic identity will be driven largely by concerns of equity among groups:

- Do jury pools contain a fair cross-section of the community?
- Do paramedics or firefighters take longer, on average, to respond to emergencies for one group than for another?
- Do childhood immunization programs reach their intended targets equally for each group?
- Do minority-owned businesses capture a fair share of local government contracts?

In the private sector, ethnic identity is important to delineating and targeting particular markets within metropolitan areas and regions. Supermarket chains cater not to "Asians and Pacific Islanders" but rather in one neighborhood to Koreans (shopping for kimchi) and in another to Thais (buying lemon grass). Airlines targeting ethnic travel markets on particular routes need to determine their share of Koreans flying first class on one route, of Japanese flying first class on another route, and which cuisine to emphasize where.

APPENDIX

REVIEW OF EXISTING SURNAME DICTIONARIES

INTRODUCTION

Surname dictionaries have been devised for various purposes but not always evaluated to check their performance. Certain surname dictionaries do meet minimum scientific standards of construction and validation, and those are the ones we discuss here.

The Census Bureau has developed a dictionary for identifying persons of probable Hispanic origin. The Bureau's List of Spanish Surnames has been carefully developed and undergone thorough evaluation (Passel and Word, 1980; Perkins, 1993). It has been used to gauge the makeup of local electorates, especially in California. The detection characteristics of this list are reasonably well understood, and its limitations have been carefully documented.

With respect to the Asian population, no comparably established surname dictionary exists, although the Census Bureau conducted exploratory work aimed at developing such a list (Passel, et al., 1982) and subsequently developed and evaluated a pre-identification dictionary (U.S. Bureau of the Census, 1988). The Bureau's initial dictionary and methodology have been elaborated by Donnelley Marketing Information Systems (DMIS).

HOW SURNAME DICTIONARIES ARE DEVELOPED[2]

There is no single "best" way to develop a surname dictionary, but researchers have devised several logically distinct approaches for constructing them.

One approach exploits data showing the origin or country of birth of persons with specific surnames. For example, census data may show that 95 percent of all persons named "XXX" report themselves to be of Spanish origin. If we infer that all persons named "XXX" are Hispanic, that percentage implies that we would be correct 95 percent of the time.

---

[2] This section draws from and elaborates points set forth originally in Passel et al., 1982.

This approach can be elaborated by developing statistical profiles of each surname by places of birth. The INS Alien Registration System, for example, contains names of legal residents whose countries of birth are known. Such data might show precisely what separate proportion of persons named, say, "CHUM" are Cambodian, Chinese, or Korean by birth. Such data provide the basis for calculating the probability that a person named CHUM was born in a particular country.

A second approach exploits knowledge of the residential location of persons with Asian-like names. For example, the surname "PARK" generally would reveal little about an individual's race, since there are many non-Korean "Parks". However, any person named Park living in a heavily Korean neighborhood may be a likely Korean. Neighborhood context, then, may condition inference.

A third approach exploits the common beginnings or endings of surnames. Surnames starting with "YAMA...," for example, are potentially Japanese, and any such surnames found among residents of predominantly Japanese neighborhoods would be strong candidates for inclusion on a Japanese surname dictionary.

Developing a surname dictionary confronts certain practical limitations. In theory, one could identify all possible surnames of particular racial and ethnic groups, but doing so would be infeasible. Were such a list developed, it would generate too many false positives (i.e., wrongly designating a person as a member of the group) to be of any use. The practical approach is to build a sufficiently representative, but less than exhaustive, surname dictionary so as not to incur too many false positives.

False positives tend to cause an overestimate of persons identified as group members. False negatives (i.e., failing to designate a person who in fact is a group member) tend to cause an underestimate. It is the *difference* between these two error rates that determines whether one ends up with a net over-estimate or an under-estimate. That is, the two rates could be roughly similar, making group estimates in a large sample quite accurate despite sizable individual error rates. The Census Bureau's Spanish surname list, for example, has a 15.0 percent false positive rate and an offsetting 20.7 percent false

negative rate (Passel and Word, 1980: Table 5). The difference (-5.7 percent) means that the list would identify 94.3 percent of the number of Hispanics in the population.

SPANISH SURNAME DICTIONARY[3]

The Census Bureau Spanish Surname List is a machine-readable reference file developed from approximately 85 million 1977 Federal tax returns. The file consists of 12,497 Spanish surnames condensed from 1.4 million distinct surnames and tabulated for 858 geographic areas. The list has been formatted for ease of use in computer-aided matching:

- All names are listed alphabetically without any blanks or spaces, in upper-case letters. EXAMPLE: "DELEON", not "De Leon".
- Accent marks and tildes have been omitted.

The Bureau furnishes the following guidelines for prospective users:

1. If a surname consists of two names separated by a dash, the person should be coded as Spanish if *either* name appears on the list. EXAMPLE: COLLINS-GARCIA.
2. If the surname consists of two surnames separated by "de", look for the name written first. If it does not appear on the list, look for the name with and without the word "de". EXAMPLE: Perez de Seda (PEREZ; DESEDA; SEDA).
3. If the surname is followed by an initial, ignore the initial and look up only the name. EXAMPLE: "Lopez", not "Lopez R.".
4. Surnames that begin with "de" or "de la" should be looked for with and without prefixes. If any of the following combinations are listed, the surname should be considered Spanish. EXAMPLE: de la Cruz (CRUZ, LACRUZ, DELACRUZ).

[3] This section is drawn from U.S. Bureau of the Census 1980a and 1980b.

ASIAN SURNAME DICTIONARIES[4]

We know of two Asian surname dictionaries: (1) one that is part of a larger special-purpose dictionary for pre-identifying potential minority business owners, developed by the Census Bureau in connection with its Survey of Minority-Owned Business Enterprises; and (2) one developed by Donnelley Marketing Information Systems (DMIS) for general-purpose use in identifying persons likely to be Asian. Our focus below is on the DMIS general-purpose dictionary; the Census Bureau's special-purpose dictionary, and evaluations of it, are provided in various unpublished documents (see U.S. Bureau of the Census, 1988).

The genesis of the DMIS Asian Surname File (ASF) was the preliminary list of 1,010 Asian surnames developed by the Census Bureau in the early 1980s (Passel et al, 1982). DMIS refined and expanded that list from its original research-oriented focus to its current applications-oriented form. The dictionary now encompasses a national universe of Asians whose actual or presumptive identification derives from an intersection of (1) the Donnelley Marketing Residential Name and Address List (RNAL) and (2) Alien Registration System data from the Immigration and Naturalization Service (INS) in 1979.

The Census Bureau's original list of 1,010 surnames was refined and expanded using complementary information about (1) Asian-like surnames of the residents of census block groups known to be inhabited predominantly by Asian households (from RNAL data); and the association between surnames and Asian country of birth (from INS data). This combination of "Asian-neighbored" and "Asian-born" information provided the basis for enlarging the original list of 1,010 surnames to the current list of 9,814 surnames, many of which are further associated with one or more specific Asian nationalities (listed below).

The operational steps followed are summarized below:

1. DMIS first analyzed the preliminary Census Bureau surname list in relation to the Donnelley Marketing list. DMIS extracted a list of 1980 census block groups where 50 percent or more of all households were Asian-headed. Within these block groups,

[4] This section is based on Donnelley (n.d.).

all surnames on a 1980 version of the DM list were tabulated to generate a frequency distribution of surnames on the 1980 DM list for all block groups where the 1980 Census showed a majority of households to be Asian. These "Asian neighboring" surnames found on the DM list were then matched to the Census Bureau list, and matches and non-matches were counted for further analysis of (a) the frequency of non-matched names in these majority Asian block groups and (b) the common beginnings and endings that characterize surnames on the Census Bureau list.

2. Many non-matched surnames were found to contain common beginnings or endings. The analysis identified over 4,000 such "Asian-like" surnames in majority-Asian block groups. These surnames with common beginnings or endings were added to the Census Bureau's original list of 1,010 surnames to form an expanded list of potentially Asian surnames.
3. DMIS subsequently obtained the INS Alien Registration System File for 1979. From that file, they selected all surnames occurring 10 or more times for which 80 percent or more of the occurrences were aliens from Asian countries of birth (shown in the list below). This procedure defines a broad universe of *Asian-born* surnames on the INS file, with additional detail on country of birth.
4. DMIS identified as many as three nationality indicators according to the following logic. First, potential country of origin was indicated only where at least 20 percent of the Asian-born surnames on the INS file reflected that specific country. Some names (e.g., "CAVAN") have up to three nationality indicators. That is, no less than 20 percent of persons named "CAVAN" were listed on the INS file as Philippine-born; another 20 percent or more were born in Viet Nam; and yet another 20 percent or more were born in Laos. The majority of Asian surnames, however, have a single nationality indicator, and many have none at all. Most of the

surnames with no nationality indicator derived from the Donnelley Marketing list.

5. The surnames derived in step 3 (along with the nationality indicators from step 4) were added to the ASF.
6. DMIS next compared the original ASF surnames with the full INS file to identify obvious anomalies. Several of the original ASF surnames found to have predominantly non-Asian origins were removed from the ASF.
7. As a final check, DMIS applied human judgment, screening all ASF surnames to identify those that could be potentially misleading in certain contexts. In about two dozen instances where DMIS was convinced that the surname could easily mislead, they placed a caution indicator on the ASF. "Lee" and "Ohara" are two examples where neighborhood context would suggest whether or not the individual might be Asian.
8. The final ASF file was placed in a format compatible for list matching. This involved removing non-alphabetic characters and internal spaces in surnames. Both the original surname spelling and the "list matching" spelling are included in the ASF. As an addition, the ASF was matched to the Census Bureau Hispanic Surname File and where a match was found, an indicator was placed on the ASF. These matches appear to have occurred only with surnames on the ASF with a predominant country code for the Philippines.

The final ASF contains 9,814 surnames, each with the following information for each surname:

Surname: 15-character positions

Spanish flag: blank or "SPNSH" if matches with Census Bureau Spanish-surname file

Caution flag: blank or "CAUTN" if the surname could be misleading

Country codes: a primary, secondary, and tertiary code if in INS

COUNTRY CODES::

BANGL - Bangladesh
BURMA - Burma
CAMBO - Cambodia
CHINA - China (includes Hong Kong and Taiwan)
INDIA - India (includes Bhutan and Nepal)
INDON - Indonesia
JAPAN - Japan
KOREA - Korea
LAOS - Laos
MALAY - Malaysia
PAKIS - Pakistan
PHILI - Philippines
SINGA - Singapore
SRILA - Sri Lanka
THAIL - Thailand
VIETN - Viet Nam
__N__ - Not Identified

PROCESSING STEPS INVOLVED

The easiest technique for associating an ethnic or racial indicator with each surname in a computerized list using a surname dictionary is called *merge/match*. This approach involves recoding all the surnames on the list so they match the format of the dictionary. For the Census Bureau and the DMIS dictionaries discussed here, this means removing any character other than a letter (e.g., a hyphen[5]) from

[5] Hyphenated names are handled using a table lookup technique. Consider the surname MONROE-GARCIA (not on the Census Bureau's dictionary of Spanish surnames, since none of those surnames contains a hyphen). Removing the hyphen produces "MONROEGARCIA", but the dictionary does not contain that surname either. The dictionary does contain "GARCIA", though, and Census Bureau guidelines suggest counting a hyphenated name composed of a Spanish surname as a Spanish surname.

The table lookup technique entails testing each surname in three steps: (1) decompose the name into components called factors (e.g., decompose MONROE-GARCIA into MONROE and GARCIA); (2) expand each factor into one or more potentials (e.g., expand PEREZ DE SEDA into three potentials: SEDA, PEREZ, and DESEDA); search the

the name and converting all letters to upper case. The recoded list of surnames from the file to be processed is then sorted alphabetically and merged with the dictionary. Wherever two names match, the name is scored as a "hit".

The surnames on the list must be preprocessed to identify potential errors. For example, an analysis of a sample of 200,000 surnames of New York City voters revealed that about one percent of them contain one or another error needing preprocessing (e.g., converting "MC CARTHY" to "MCCARTHY"; "PEAR5SON" to "PEARSON").

---

dictionary for each potential in each factor. If any one potential is found in the dictionary, code the entire surname as a Spanish surname.

REFERENCES

Donnelley Marketing, internal memo by Tom Hryniewicz, n.d.

Hryniewicz, Thomas and Kenneth Hodges, "A Longitudinal Surname Match Method for the Production of Small Area Estimates of the Hispanic Population," presented at the 1989 Southern Demographic Association Meetings, Durham, NC.

Passel, Jeffrey S., David L. Word, Nampeo D. McKenney, and Yun Kim, "Postcensal Estimates of Asian Populations in the United States: A Description of Methods Using Surnames and Administrative Records," presented at the 1982 Population Association of America meetings, San Diego, Calif., April 1982.

Passel, Jeffrey S. and David L. Word, "Constructing the List of Spanish Surnames for the 1980 Census: An Application of Bayes' Theorem," presented at the 1980 Population Association of America meetings, Washington, D.C.

Perkins, R. Colby, "Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results," Technical Working Paper No. 4, Population Division, U.S. Bureau of the Census, August 1993.

Shin, Eui-Hang and Eui-Young Yu, "Use of Surnames in Ethnic Research: The Case of Kims in the Korean-American Population," *Demography*, Vol. 21(3), August 1984, pp. 347-359.

U.S. Bureau of the Census, Census of Population of Housing, 1980a. Spanish Surname List (machine-readable data file), prepared by the Bureau of the Census. Washington, D.C.: The Bureau (producer and distributor), 1980.

_____, Census of Population and Housing, 1980b. Spanish Surname List Technical Documentation, prepared by Data Access and Use Staff, Data User Services Division, Bureau of the Census, Washington: The Bureau, 1980.

_____, unpublished memoranda on surname matching and surname and first name research for the Survey of Minority-Owned Business Enterprises (SMOBE), prepared by the Assistant Chief, Research and Methodology, 1988.