THE ASIAN AND PACIFIC ISLANDER SURNAME LIST: AS DEVELOPED FROM CENSUS 2000

Matthew R. Falkenstein, U.S. Census Bureau[1]

Matthew R. Falkenstein, U.S. Census Bureau, 4700 Silver Hill Rd,
Washington, D.C. 20233-1912

Key Words: Race, Ethnicity

1. Introduction

The Asian and Pacific Islander (API) population has rapidly grown from a small minority a few decades ago to large and diverse population groups. Since 1980, the Census Bureau has made more of an effort to allow for more accurate self-identification of race, including the introduction of eleven race categories for the API race. Thanks partially to better API race reporting stemming from the new category scheme, interest in detailed

race data has spurred a demand for a surname list tied to the Census 2000 Asian and Pacific Islander race categories.

1.1 Purpose of the API Surname List

The main purpose for the API Surname List was to improve race models created by Planning, Research and Evaluation Division (PRED) Administrative Records branch (AR). PRED has developed these models for use in supporting various Population Division migration and estimates programs.

Although an API race category was on Census 1990, there was no API race category on Census 2000.We assembled a race category called API from Census 2000 data specifically for internal race models. However, the API Surname list was designed so that individual Asian and Pacific Islander races can be readily disaggregated: Chinese, Korean, Vietnamese, etc surname lists can be produced easily. For the researcher interested in a particular race group, a separate surname list can be produced.

We think that the API Surname list and its potential component surname lists may provide the researcher with a viable alternative to a pervasive problem: when using demographic survey or administrative data for survey or research, race data are often inaccurate or incomplete. The association of a surname with the API race groups will allow researchers to fill missing data, or at very least make reasonable assumptions about the race of the respondent based on their surname.

1.2 Previous Surname List Research

The possibility of using surname to enhance Census operations has been explored for many years at the Census Bureau. For example, various Spanish surname files were produced from the censuses of 1950 through 1990. For example, the 1950 Spanish surname list helped identify Hispanic population found in Arizona, California, Colorado, New Mexico, and Texas. More comprehensive lists were developed as additional data became available.

Bye (1998) compiled an API surname file based on four existing files:

- From the Census Bureau, the 1990 Post Enumeration Survey (PES)
- From the Social Security Administration (SSA) NUMIDENT file, a list of Hawaiian-born persons obtaining social security numbers (SSN)
- From SSA NUMIDENT, a list of over-50 persons born in 19 Asian countries who had an SSN in 1995 or earlier
- From an Immigration and Naturalization Service file, a list of naturalized citizens born in 19 Asian countries

The resulting list contributed to Census Bureau race regression models produced by PRED. Note that the universe was somewhat limited due to a small sample size of API respondents.

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Lauderdale and Kestenbaum (2000), in an attempt to improve Asian race category identification, used Social Security Administration (SSA) data to compile several Asian surname lists based on six major Asian race categories.

The 1990 Spanish Surname list (see Word and Perkins 1996) was based on the Census Spanish Origin File (SOR). The SOR uses race self-identification, with a direct connection to a respondent's Hispanic origin and first and last names.

We considered the 1990 Spanish Surname list as the best model for our API Surname list. The methodology used to produce the API Surname list is similar to the 1990 Spanish Surname List in that it relies on a Census 2000 respondent's actual race self-identification. The size of the Census 2000 (over 282 million records, with nearly 12 million self identified as API) also gives great weight to its validity: sampling error is nonexistent as the population is used rather than a sample.

1.3 Multiple Race and Write In Race

About 2.4 percent of Census 2000 respondents chose more than one race. In our methodology, if a person chose more than one race, we tallied each one as a fraction of the total. For example, if a person chose Chinese and Vietnamese, that Census record was calculated at 0.500 for each, and the total equaled 1.000. If a person chose three races, each was calculated at about 0.333, for a total of 1.000.

It should also be noted that we made no attempt to tally any write in race. This was mainly due to the small number of write ins, the complications involved with interpretation and translation to an electronic file, and to keep the general processing and editing procedures clear. Records with write in races were edited out of the calculation.

1.4 Preservation of Privacy and Confidentiality

Since the API Surname List is expected to be available to other Federal agencies, academia, and the public, certain limitations protecting the individual's privacy were required. To protect the privacy of the individual respondent, we set a level of 50 or greater occurrences on Census 2000 as the limit for inclusion of a unique surname to the list. There is no link on the list to a respondent's geographic data, age, or first name. Therefore, no identification of an individual respondent is possible. Given this condition, we believe the API Surname List does not violate Title XIII or other U.S. privacy laws.

2. Methodology

Each Census 2000 respondent was asked to self-identify race. There were fifteen general categories including eleven Asian and Pacific Islander race groupings. Based on that data from Census 2000 (from the Hundred Percent Unedited File (HCUF), Version Double Prime), we tallied the surname count for each Census 2000 API race.

2.1 Pre-processing Edits

A few pre-edits were done in preparation for calculation of the API strength of association. About 20 million records with surnames or races with the following conditions were deleted:

- Blank, "X" or "A" filled or partially filled
- Single alpha-numeric characters (except "O")
- Write in race
- No race chosen (no imputed races were used)

After editing, about 262 million records were processed. About 10.9 million were self-identified as one or more of the Asian or Pacific Islander races.

We allowed the single character surname "O" because of the possibility of some Asians spelling their surname this way. This may be another spelling of the surname also spelled "Oh". The surname "O" appears in Lauderdale and Kestenbaum (2000).

2.1 API Strength of Association by Grouping

We calculated the proportion of respondents in each race by dividing the count of a surname for each race by the total count for each surname.

The next step was to sum all of the API proportions for each surname. The calculation below and Table 1 describe the summation. We use the top-ranked surname NGUYEN as an example:

API for Nguyen = (Chinese + Japanese + Vietnamese + Korean + Asian Indian + Filipino + Other Asian + Native Hawaiian + Guamanian/Chamorro + Samoan + Other Pacific Islander)

Table 1. Example of API Proportions for Surname NGUYEN

API Surname	API Race	Proportion
NGUYEN	Chinese	0.0072
	Japanese	0.0005
	Vietnamese	0.9411
	Korean	0.0013
	Asian Indian	0.0049
	Filipino	0.0015
	Other Asian	0.0203
	Native Hawaiian	0.0000
	Guamanian /Chamorro	0.0000
	Samoan	0.0001
	Other Pacific Islander	0.0004
	Total API %	**0.9773**
	Other Races	0.0277
	TOTAL by Surname	***1.0050**

*Shown with rounding error

Since our goal was to identify the surnames most strongly associated with Asians and Pacific Islanders, we broke out the list into a grouping scheme to describe what we term the API "Strength of Association" (API SOA) for each surname relative to the API race category. This was similar to the labeling plan described by Word and Perkins (1996) for the 1990 Spanish Surname list. Table 2 illustrates.

Table 2. API SOA Groupings

API SOA Grouping	Proportion Range
Heavily API	0.75 or greater
Generally API	0.50 – 0.74
Moderately API	0.25 – 0.49
Occasionally API	0.05 – 0. 24
Rarely API	Less than 0.05

After adding the race proportions and calculating the API SOA, we resorted the list by descending Census surname count within each API grouping. This gave us an API list sorted by the most strongly associated API surnames. For example, the name "CHAN" had an API SOA of 0.94, and a total Census 2000 frequency of occurrence of 65,956. Since it's API SOA was greater than 0.75, CHAN was placed in the "Heavily API" grouping. Using this methodology, the name Chan was ranked tenth and was effectively the tenth most common API surname within this population.

This approach was adopted because there are API surnames that had a very high API SOA, but a low overall count of Census respondents. The inverse was also true: surnames with a lower API SOA but a very high Census 2000 count occurred. In the case of "LEE", the API strength of association was about 0.40, but the count was about 600,000. Numerically, LEE was the most common name on our list, but on Census 2000, 60% of persons with the surname LEE self identify race as something other than API.

3. RESULTS

The total surname record count after editing is 8,436,339. Note, however, that the figure comprised *all* of the surnames on the API Surname list, including those with a small API SOA or a low Census 2000 count. Many surnames occurring only a few times contained errors in spelling, which is reflected in the high total surname count. Table 3 describes the distribution of occurrence for all surnames on the API Surname file.

Table 3. Frequency Distribution of Census 2000 Surname Count

Census Frequency	Surname Count	Percentage of Total (%)
1 - 9	7,621,482	90.34
10 - 49	552,964	6.56
50 +	261,891	3.10
TOTAL	**8,435,198**	**100.00**

As Table 3 shows, a vast majority of API surnames (90%) occur nine times or fewer. Only about 262,000 surnames occurred 50 or more times for what was essentially the entire U.S. population. About 84.1% of the Asian and Pacific Islander population has a surname that occurred 50 or more times.

Table 4 depicts the top 25 names in the "Heavily API" group based on the strength of API association. Each table of names was sorted in order of frequency within each group, as discussed earlier.

Table 4. Top 25 API Surnames – API Proportion of 0.75 or Greater

Surname	Rank	Census 2000 Count
NGUYEN	1	290,101
KIM	2	191,623
PATEL	3	143,325
TRAN	4	129,138
CHEN	5	99,871
WONG	6	98,064
LE	7	74,646
SINGH	8	67,651
WANG	9	66,645
CHAN	10	65,956
CHANG	11	65,202
YANG	12	64,345
PHAM	13	54,512
LI	14	54,119
LIN	15	49,645
LIU	16	48,458
WU	17	43,958
LAM	18	43,150
HUANG	19	42,432
HO	20	39,073
HUYNH	21	37,479
SHAH	22	36,801
YU	23	35,672
CHUNG	24	35,450
CHOI	25	34,856
	TOTAL	1,912,172

3.1 Most Common Race Groups

As one would expect, the strongest associated API surnames fall within the "Heavily API" or Generally API" groups (11,446 names). For example, the surnames in Table 4 (Heavily API) appear to be strongly Chinese, Vietnamese, Asian Indian, and Korean. Indications are that certain surnames are very commonly associated with these race categories. Names like Wong, Nguyen, Kim, and Patel are common in the countries of China, Vietnam, Korea, and India, respectively. About 1.91 million persons have surnames in the top 25 list.

Not shown are the surnames in Generally API and Moderately API groups, where more Islamic names (numerous Asian countries) and the Hispanic names (some are Filipino) begin to appear. In addition, the surnames appear decidedly less associated with eastern Asian countries in the Occasionally API and Rarely API groups.

3.2 Less Populous Race Groups

Races with a small representation in terms of U.S. population were generally overwhelmed by the surnames of the more populous Asian race groups. For example, Asian Indian, Korean, Chinese and Vietnamese names dominated the top of the "Heavily API" and "Generally API" categories. Race groups such as Native Hawaiian and Samoan had small U.S. populations, and so failed to make the top rankings of surnames for any of the categories.

3.3 Consistency of Methodology

To evaluate the API Surname List, we did a simple list-to-list comparison against the first 50 surnames on three Asian race lists produced by Lauderdale and Kestenbaum (2000).

The Lauderdale-Kestenbaum lists were based on SSA data for Asian Americans born in Asia before 1941. The intent was to develop a proxy of racial ethnicity based on elderly Asian Americans. Records in the file tallied to about 1.8 million persons from 16 Asian countries.
In an attempt to evaluate how well we tallied and sorted API surnames, we attempted to compare our tallies with three of the Lauderdale-Kestenbaum Asian Ethnicity lists. We took three subsets of the API Surname List and resorted into the Chinese, Japanese and Asian Indian race groups. We considered a surname to be a successful match if it fell within the first 50 on both our list and the Lauderdale-Kestenbaum list.

We successfully matched 30 Chinese surnames from our list with the first 50 from the Lauderdale-Kestenbaum list. On the race Japanese list, 31 out of first 50 names matched. On the Asian Indian list, only about 20 out of the first 50 matched. However, when we checked the Lauderdale-Kestenbaum list against all of the Asian Indian surnames that ranked in either the Heavily API or Generally API SOA category, we matched 36 out of 50.

Our evaluation was far from conclusive, and the matching rate was not high. Most likely this was due to the differing universes (Census 2000 in this paper and SSA data in Lauderdale-Kestenbaum) and the different time periods. However, given the two completely distinct sources, a 60% match rate of the first 50 names does not seem unreasonable.

4. CONCLUSION AND FURTHER RESEARCH

The Census 2000 file was used as a basis for a tally of API surnames. The API proportion was calculated for each surname and summaries presented. Surnames were grouped by strength of API association and ranked by frequency of Census 2000 occurrence by group.

The initial API Surname list consists of 8.43 million surnames, but only about 262,000 occur 50 or more times. In order to preserve the confidentiality of the Census 2000 respondent, the final API Surname list was limited to those names that occur 50 times or more.

It appeared that surnames of Chinese, Asian Indian, Vietnamese and Korean race ranked highest in the two top API SOA categories. Japanese names ranked somewhat below those race groups, and Pacific Islander ranked even lower. These groups have lower U.S. populations. Western European and Hispanic surnames were the most common surnames in the "Rarely API" SOA API category.

Further research is warranted into this rather interesting subject. Additional applications of the API Surname list may include imputation of missing race, planning for special census enumerations, and targeting of the API population for surveys.

The question on how to identify Filipino surnames remained unanswered with this surname list. There is a meshing of Hispanic based Filipino surnames and other Hispanic surnames. How to unravel that is not addressed here, as the question of Hispanic origin was never approached.

With the methodology we used, individual race groups can be produced. A surname list is possible for all Asian or Pacific Islander race groups, though, as the list stands, success may be limited with certain race groups like Filipinos.

Although additional research is required, the API Surname List should prove useful to a variety of researchers with diverse interests.

5. REFERENCES

Bureau of the Census. 1995. Name Files. http://www.census.gov/genealogy/names/. Accessed August 15, 2001.

Bureau of the Census. March 2001. *2000 Census NUMIDENT Program Specification. (CNUM0101-00).* Planning, Research and Evaluation Division.

Bye, Barry V. December 1998. *Race and Ethnicity Modeling with SSA NUMIDENT Data: Individual-Level Regression Model – Version 2.* http://cww.census.gov/msdir/pred/prednet/memos/memo19.pdf. Accessed November 12, 2001.

Lauderdale, Diane S. and Bert Kestenbaum. 2000. *Asian American Race Identification by Surname.* Population Research and Policy Review. 19: 283-300.

Philipp, Dan. February 2001. *Hundred percent Census Unedited File -Version 2.* Bureau of the Census, Decennial Systems and Contracts Management Office, Data Collection Control Staff.

Word, David L. and Perkins. March 1996. *Building a Spanish Surname List for the 1990's- A New Approach to an Old Problem.* Bureau of the Census, Population Division Technical Working Paper No. 13. http://www.census.gov/population/documentation/twpno13.pdf. Accessed August 19, 2001.