Name Method: tables for the manuscript

Cal Chengqi Fang

2024-07-23

In this document, I made some tables for our manuscript to illustrate our analysis process and results.

Table 1. Illustrative surnames from the Probabilistic List and the estimated probabilities of origins from each area conditional on having the corresponding surnames

	Country/Area of origin*																		
Illustrative surnames	Afghanistan	Bangladesh, India, Pakistan, Sikkim, British India	Bhutan	Brunei, Indonesia	Cambodia	Japan, Southern Ryukyu Islands	Korea, Democratic People's Republic of Korea, Republic of Korea	Laos	Malaysia	Maldives	Mongolia	Myanmar	Nepal	Peoples Republic of China, Hong Kong, Macau (Macao), Taiwan	Philippines	Singapore	Sri Lanka	Thailand	Vietnam, Democratic Republic of Vietnam, Republic of Vietnam	All the other countries and regions
HUANG	0.00	0.09	0.00	0.13	0.07	0.19	0.13	0.08	0.12	0.00	0.00	0.14	0.00	96.28	0.32	0.22	0.00	0.06	1.31	0.86
KIM	0.00	0.31	0.00	0.02	0.74	0.81	94.74	0.01	0.05	0.00	0.00	0.15	0.00	0.34	0.10	0.02	0.00	0.11	0.59	2.02
SUZUKI	0.00	0.00	0.00	0.03	0.03	95.73	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.33	0.04	0.00	0.09	0.03	2.79
NGUYEN	0.00	0.02	0.00	0.14	0.32	0.07	0.07	0.31	0.17	0.01	0.00	0.01	0.00	0.25	0.41	0.01	0.00	0.20	95.14	2.88
CHANG	0.00	0.24	0.00	0.23	0.37	0.62	21.64	3.02	0.88	0.00	0.00	0.50	0.00	62.10	0.56	0.35	0.01	2.21	1.52	5.75
INTHAVONG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	90.75	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	7.88	0.45	0.00
VICENCIO	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	1.07	55.85	0.02	0.02	0.02	0.02	42.76
SAMEER	0.00	52.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	47.87
PATAN	17.74	33.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.45	0.00	41.94

* Note: This table exemplifies the Probabilistic List with some illustrative surnames. Each column represents one illustrative surname. Numbers in the row denote the probabilities of someone originating from the corresponding area conditional on having corresponding surnames. Probabilities in each row may not add up to 1 due to rounding error.

Table 2. Numbers of surnames in the Deterministic List that are predictive for each area of origin, with the number of US-born and foreign-born SSN records having that surname

Country/Area of origin	Number of predictive surnames*	Number of US-born records with corresponding predictive surnames	Number of foreign-born records with corresponding predictive surnames
Afghanistan	215	12,946	18,672
Bangladesh, India, Pakistan, Sikkim, British India	6,738	1,281,803	2,506,632
Bhutan	58	4,628	22,019
Brunei, Indonesia	368	14,793	41,692
Cambodia	537	96,110	92,847
Japan, Southern Ryukyu Islands	2,995	523,987	758,857
Korea, Democratic People's Republic of Korea, Republic of Korea	277	998,728	1,289,503
Laos	574	158,842	127,920
Malaysia	119	7,105	19,192
Maldives	0	0	0
Mongolia	48	1,060	3,236
Myanmar	187	19,991	71,161
Nepal	138	12,000	63,997
Peoples Republic of China, Hong Kong, Macau (Macao), Taiwan	572	1,078,677	2,813,805
Philippines	7,794	940,978	923,734
Singapore	11	466	1,852
Sri Lanka	115	10,020	15,979
Thailand	41	1,088	2,524
Vietnam, Democratic Republic of Vietnam, Republic of Vietnam	257	594,284	1,221,744

* The numbers do not add up to 22,621 because there are surnames that are equally predictive of more than one country/area.

† Note: The second column of this table presents the numbers of surnames that are predictive of one country/area with a 50% or higher probability. The third and fourth columns are the number of Social Security Number (SSN) application records that have the corresponding surnames.

Intext. Validation of the Probabilistic List's ability to predict the subset of names on the Frequently Occurring Surnames from the 2010 Census that have their most frequent self-reported race/ethnicity as "Asian and Pacific Islander"

This was originally a table but we decided to only report the first row in the text and ignore the other rows as the validation done on KP cohort is more robust and complete.

There are 3170 surnames that appeared in both the Probabilistic List and the list of Frequently Occurring Surnames from the 2010 Census. The sensitivity and specificity of our list are 87.02 and 97.99, respectively. And the PPV of our list predicting the surnames in the list of Frequently Occurring Surnames from the 2010 Census is 77.41.

Table 3. Validation of the Deterministic List's ability to predict foreign-born population's ethnicity through cross validation

Country/Area of origin	PPV (%)	Sensitivity (%)	Specificity (%)
Afghanistan	75.95	51.44	99.96
Bangladesh, India, Pakistan, Sikkim, British India	87.40	98.84	95.94
Bhutan	62.42	47.71	99.92
Brunei, Indonesia	82.72	68.26	99.93
Cambodia	75.96	62.79	99.78
Japan, Southern Ryukyu Islands	94.56	96.65	99.55
Korea, Democratic People's Republic of Korea, Republic of Korea	80.52	93.51	97.17
Laos	71.86	71.86	99.63
Malaysia	73.01	22.44	99.95
Maldives	33.33	1.16	100.00
Mongolia	97.00	93.01	100.00
Myanmar	72.89	66.73	99.81
Nepal	72.61	72.41	99.82
Peoples Republic of China, Hong Kong, Macau (Macao), Taiwan	87.96	93.62	95.39
Philippines	86.74	95.03	98.67
Singapore	65.65	5.18	99.99
Sri Lanka	79.63	75.10	99.97
Thailand	86.64	2.84	100.00
Vietnam, Democratic Republic of Vietnam, Republic of Vietnam	91.49	92.67	98.82

Note: The table presents the result of validating the Deterministic List through four-fold cross-validation. In each iteration, 75% of the sample is used to estimate a Probabilistic List. The country/area that each surname predicts with a probability of 50% or higher is then compared with the country/area of birth of samples in the set-aside 25% test set. The results are summarized into a confusion matrix, and Positive Predictive Value (PPV), sensitivity, and specificity are calculated. The table shows the mean results from the four test-set evaluations.

Table 4. Sensitivity and specificity of the Deterministic List when used to predict ethnicity with Kaiser Permanente cohort

Country/Area of origin	Number of populations in Kaiser Permanente Asian Cohort	Sensitivity (%)	Specificity (%)
Asia	1,085,429	71.51	97.32
Afghanistan	1,458	22.02	99.95
Bangladesh, India, Pakistan, Sikkim, British India	110,563	69.93	99.27
Indonesia	1,928	20.49	99.99
Cambodia	5,425	42.18	99.94
Japan	24,844	58.45	99.80
Korea, Democratic People's Republic of Korea, Republic of Korea	20,492	78.49	99.40
Laos	5,064	38.19	99.94
Malaysia	702	11.40	99.95
Myanmar	1,893	45.85	99.97
Nepal	2,310	59.91	99.97
Peoples Republic of China, Hong Kong, Macau (Macao), Taiwan	206,370	76.53	99.43
Philippines	192,565	38.93	99.26
Singapore	245	4.08	99.99
Sri Lanka	562	25.09	99.99
Thailand	3,846	0.52	100.00
Vietnam, Democratic Republic of Vietnam, Republic of Vietnam	56,561	85.97	99.64

spaceholder

Table 5. Positive predictive value (PPV) of the Deterministic List's ability to predict Asian ethnicity by geography

Country/Area of origin	Nationwide			San Francisco metropolitan area		
	Prevalence (%)	PPV (%)	PPV conditional on Asian (%)	Prevalence (%)	PPV (%)	PPV conditional on Asian (%)
Asia	5.77	62.05	NA	27.27	90.91	NA
Bangladesh, India, Pakistan, Sikkim, British India	1.49	59.23	98.25	5.43	84.62	97.43
Cambodia	0.08	36.65	74.92	0.18	56.12	57.10
Indonesia	0.03	34.27	61.95	0.10	67.67	57.05
Japan	0.23	40.27	90.98	0.84	71.32	88.15
Korea, Democratic People's Republic of Korea, Republic of Korea	0.45	37.15	71.04	1.11	59.45	54.06
Laos	0.05	25.85	55.97	0.15	48.59	41.09
Malaysia	0.01	1.49	4.51	0.02	5.30	3.42
Myanmar	0.06	47.80	79.94	0.12	65.38	62.41
Nepal	0.06	54.88	86.22	0.14	74.05	74.79
Peoples Republic of China, Hong Kong, Macau (Macao), Taiwan	1.35	64.82	91.18	11.23	94.44	95.79
Philippines	0.89	32.19	93.30	5.38	74.95	94.70
Sri Lanka	0.02	30.53	73.27	0.04	51.02	56.89
Thailand	0.06	100.00	100.00	0.13	100.00	100.00
Vietnam, Democratic Republic of Vietnam, Republic of Vietnam	0.57	57.80	81.58	1.45	77.84	68.43

spaceholder

Sample Table. Numbers of foreign-born records from corresponding country/area - Leftout

Country/Area of birth	Number of foreign-born records
Afghanistan	32,139
Bangladesh, India, Pakistan, Sikkim, British India	2,273,636
Bhutan	33,841
Brunei, Indonesia	58,115
Cambodia	136,556
Japan, Southern Ryukyu Islands	749,436
Korea, Democratic People's Republic of Korea, Republic of Korea	1,164,951
Laos	132,118
Malaysia	105,638
Maldives	150
Mongolia	3,502
Myanmar	98,144
Nepal	73,658
Peoples Republic of China, Hong Kong, Macau (Macao), Taiwan	2,768,925
Philippines	890,207
Singapore	46,230
Sri Lanka	18,651
Thailand	85,474
Vietnam, Democratic Republic of Vietnam, Republic of Vietnam	1,263,853
All the other countries and regions	668,407