Imputing the Legal Status of the Foreign Born Persons on Surveys: Two Approaches

D. H. Judson, Decision Analytics
Sharon K. Long, University of Minnesota

This work was funded by SHADAC, a program of the Robert Wood Johnson Foundation.

Motivation for Study

- States need estimates of size and characteristics of undocumented immigrants to implement federal health reform legislation
- Existing data sources on legal status are “thin”
- Question of how best to impute the legal status of foreign born persons in national surveys
  - Study compares alternate methods

The Foreign Born by "Status"

- Definitions are not necessarily common across entities
- (Authorized) Legal immigrants
  - Naturalized (obtained citizenship)
  - Lawful Permanent Resident (LPR; formally admitted status)
- Legal Temporary (a/k/a nonimmigrant)
  - Application accepted; must have no violations
- Refugee/Asylee/Temporary Protected Status
  - Granted this status but not yet converted to other status
- Unauthorized or Other
  - Has or has not applied for any other status or status has not been granted yet

A Simple Residual Method

- FB = [L – (M + E)] + T + R
  - Where:
    - FB = Total Foreign Born Population,
    - L = Legal Immigrants,
    - M = Legal Immigrant Mortality,
    - E = Emigration of Legal Immigrants,
    - T = Temporary Migrants; and ,
    - R = Residual Foreign Born (Unauthorized or Quasi-Legal)
- Solve for R (the residual) after all other statuses have been identified
- Voila: We have a number for the unauthorized – the population of interest that does not exist in any other record

Complaints About the Residual Method

- Using survey data for estimates can result in biases as the unauthorized do not necessarily want to respond to surveys (coverage)
- Administrative records are generated for other purposes thus may contain missing information or errors (e.g., LPR file and emigration)
- Other characteristics of interest do not necessarily “fall out” of residual methods
- Several assumptions (e.g. migration) may not hold
- Negative residuals

A Model Based Approach

- The Survey of Income and Program Participation (SIPP) is a longitudinal survey primarily used to determine economic well-being
- Migration history is collected in topical module 2 including legal status at entry (“permanent”, “refugee”, “other”) and if a status change occurred
- However, SIPP sample size is small relative to other surveys such as the American Community Survey (ACS)
- A technique to enhance the larger survey:
  - Use logistic regression to predict legal status; estimate parameters for various demographic characteristics also found on other surveys
  - Apply those parameter estimates to the larger survey
  - *Predict* the foreign born person’s legal status
- Target: “Not LPR or likely misreporting of citizenship status”

Pros of the Model-Based Approach

- Take advantage of *direct* information -- No other Federal survey asks legal status of the foreign born
  - Immigration status upon entry to the U.S.
  - Change in status to permanent resident
- Then, take advantage of a much larger survey (American Community Survey) to impute the legal statuses estimated from the SIPP
- The model can be run in future SIPP panels to update the parameter estimates

Cons of the Model Based Approach

- Public-use files collapse legal status categories
- Because the SIPP sample size is small, the number of cases of immigrants will be relatively small
- There is likely to be (downward) response bias in the migration questions

Latent Class Analysis

- Latent Class Analysis (a/k/a finite mixture modeling) improves on simple “classical” clustering
- People in the same “cluster” share a common joint probability distribution among the observed variables estimated by maximum likelihood methods

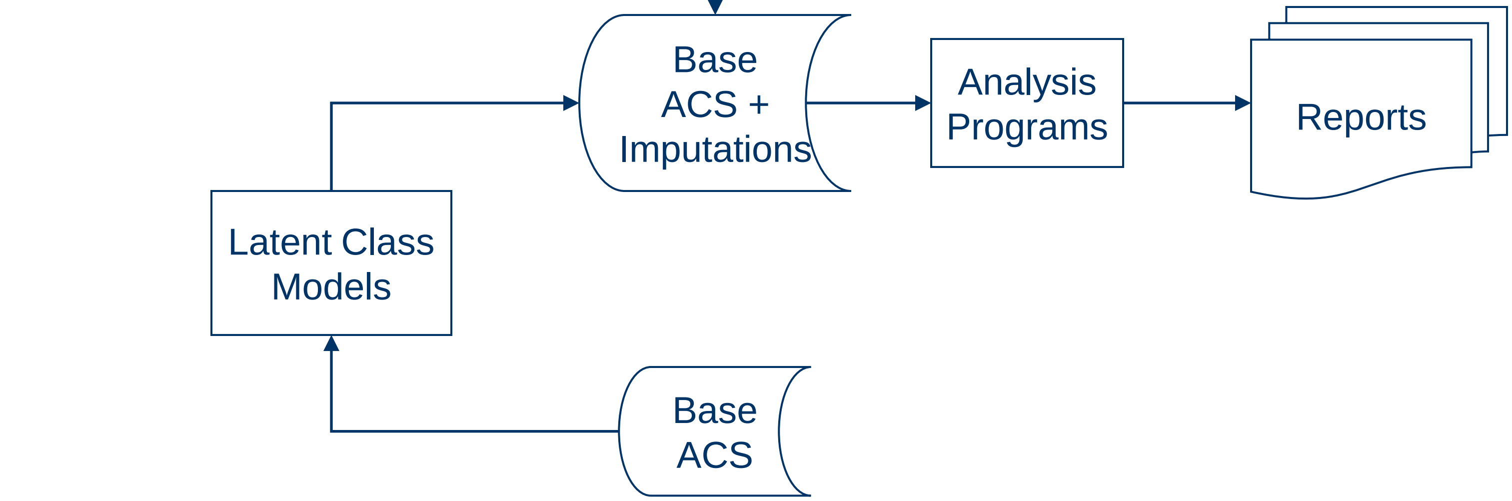
Pros of Latent Class Modeling

- Various diagnostics are available
- It allows the inclusion of exogenous variables
- It can be performed on variables that are measured on different scales
- Output is a probability (based on a probability model) rather than a fixed class assignment
- The clustering is testable against other methods
  - E.g. multinomial logit analysis
  - The results should be very similar

Cons of Latent Class Modeling

- Conditional independence
- Choice of explanatory variables
- Model identification
- The data are allowed to do “all” the talking

Estimation System Diagram



Problems to be Addressed

- Coverage of the foreign born on surveys
  - Foreign born in general, and unauthorized foreign born in particular, are widely believed to be undercovered in surveys
  - Resolved via ‘control totals’:
    - Simple Rake Factor: Control only to total FB population
    - Complicated Rake Factor: Control to broad age/sex categories
- Model specification, three approaches
  - Analyst-driven RHS (SIPP)
  - “Boosted” automatic interaction detection (Boosted SIPP)
  - A variety of latent class models attempted
- Variance Estimation
  - Not primary focus today, but if imputation independent of sample design, $V(T)=V(I)*V(S)$, where sample design variance is calculable, and model mean squared error can be taken as imputation variance

An Overview of the Imputation Models

- SIPP: Left hand side “not LPR or likely misreport”, Right hand side a collection of variables (year of entry, age, age-squared, ‘hard-to-count’ variables) likely to be predictive
- Boosted SIPP: Right hand side variables entered into automatic boosting program, unattended
- LC: Due to identifiability concerns, a limited set of hard-to-count variables

Comparison of Results

- After probabilities are placed on individual records, totals by state/domain are calculated by summing *probabilities*
- If the approach has external validity, the results of the (various) modeling strategies should look reasonably similar on an aggregate basis
- If the results are similar to those generated by residual methods, the validity of each is enhanced
- Small-domain results should make demographic sense

Comparison of Total Estimates to Each Other (1 of 2)

Table 1: Survey Total Estimates by state (Model 1-SRF,1-CRF,2-CRF,3-CRF,4-CRF, SIPP, and Boosted SIPP), based on 2009 ACS

	(LC1-SRF) Total	(LC1-CRF) Total	(LC2-CRF) Total	(LC3-CRF) Total	(LC4-CRF) Total	(SIPP) Total	(Boosted SIPP) Total
State							
Alabama	52,766	55,900	54,391	54,391	54,391	59,082	55,957
Alaska	10,823	10,292	10,285	10,285	10,285	9,190	8,046
Arizona	308,182	310,990	317,093	317,096	317,096	318,739	329,204
Arkansas	39,644	41,376	42,195	42,196	42,196	43,957	44,683
California	2,626,233	2,579,563	2,622,181	2,622,203	2,622,203	2,566,899	2,737,879
Colorado	164,515	170,177	169,840	169,840	169,840	171,813	176,573
Connecticut	120,885	121,192	121,816	121,816	121,816	116,817	111,107
Delaware	22,872	23,884	22,953	22,952	22,952	26,168	24,171
District_of_Columbia	22,371	22,261	22,229	22,228	22,228	23,648	22,130
Florida	901,494	852,226	836,855	836,851	836,850	877,530	833,313
Georgia	304,404	321,296	318,523	318,521	318,520	323,882	318,056
Hawaii	45,075	39,122	39,417	39,417	39,417	33,986	31,511
Idaho	30,287	31,199	31,671	31,671	31,671	30,528	30,881
Illinois	470,551	471,308	476,907	476,908	476,908	450,244	468,617
Indiana	89,362	94,461	93,177	93,176	93,176	98,784	93,502
Iowa	36,974	39,699	38,710	38,710	38,710	38,384	38,119
Kansas	62,707	66,371	65,648	65,648	65,648	68,848	65,005
Kentucky	44,564	47,428	46,137	46,137	46,137	51,071	46,704
Louisiana	43,472	44,963	44,133	44,132	44,132	48,234	42,690
Maine	9,212	8,986	9,566	9,566	9,566	6,757	6,423
Maryland	198,727	200,827	199,325	199,324	199,324	194,428	178,759
Massachusetts	242,264	241,781	239,154	239,153	239,153	227,124	207,128
Michigan	149,844	148,361	147,947	147,947	147,947	143,990	129,942
Minnesota	99,245	105,356	104,678	104,677	104,677	100,107	96,652
Mississippi	20,396	21,981	21,606	21,605	21,605	23,226	21,391

Comparison of Total Estimates to Each Other (2 of 2)

Missouri	59,687	61,159	60,568	60,567	60,567	61,352	56,820
Montana	4,526	4,501	4,595	4,595	4,595	4,433	3,847
Nebraska	36,809	39,422	38,487	38,487	38,487	40,216	39,723
Nevada	152,006	153,532	153,075	153,075	153,075	164,842	160,705
New_Hampshire	18,231	17,851	18,045	18,045	18,045	16,463	15,259
New_Jersey	441,543	444,570	436,008	436,004	436,004	444,634	427,844
New_Mexico	66,708	67,345	68,605	68,606	68,606	69,134	68,706
New_York	1,006,584	968,245	950,415	950,412	950,412	926,298	914,987
North_Carolina	238,912	255,069	249,886	249,883	249,883	266,484	262,477
North_Dakota	5,786	5,905	5,732	5,732	5,732	6,492	5,331
Ohio	110,866	115,010	114,059	114,058	114,058	109,791	100,255
Oklahoma	63,481	67,420	67,415	67,415	67,415	72,102	71,137
Oregon	117,998	119,787	121,503	121,504	121,504	120,239	124,764
Pennsylvania	163,302	163,497	163,780	163,779	163,779	150,333	140,531
Rhode_Island	35,923	34,898	34,179	34,179	34,179	35,580	35,383
South_Carolina	71,773	75,644	73,696	73,695	73,695	84,884	78,565
South_Dakota	4,789	5,119	4,973	4,973	4,973	5,173	4,469
Tennessee	88,361	94,774	94,105	94,105	94,105	98,138	93,770
Texas	1,347,441	1,370,619	1,379,599	1,379,603	1,379,603	1,418,183	1,450,644
Utah	68,551	72,520	74,017	74,017	74,017	72,127	73,349
Vermont	4,713	4,305	4,481	4,481	4,481	2,861	2,891
Virginia	220,842	225,510	224,088	224,087	224,087	226,234	206,909
Washington	220,420	224,261	224,469	224,469	224,469	212,192	207,814
West_Virginia	5,111	5,275	5,505	5,505	5,505	4,639	4,314
Wisconsin	73,594	77,210	76,734	76,733	76,733	78,528	75,997
Wyoming	5,173	5,548	5,544	5,544	5,544	5,213	5,067
Observations	171305	171305	171305	171305	171305	171305	171305

Comparison of Total Estimates with Office of Immigration Statistics Residual Estimates

State of Residence of the Unauthorized Immigrant Population

OIS residual estimates			SIPP model-based estimates		Latent Class (model 1)-based estimates		Boosted SIPP model-based estimates	
					ACS 2009			
State of residence	**January 2009**	**Percent of total**	**Total Estimate**	**Percent of total**	**Total Estimate**	**Percent of total**	**Total Estimate**	**Percent of total**
Total	**10,750,000**		**10,750,000**		**10,750,000**		**10,750,000**	
California	**2,600,000**	**24%**	**2,566,899**	**24%**	**2,579,640**	**24%**	**2,741,810**	**26%**
Texas	**1,680,000**	**16%**	**1,418,183**	**13%**	**1,370,620**	**13%**	**1,437,921**	**13%**
Florida	**720,000**	**7%**	**877,530**	**8%**	**852,209**	**8%**	**829,632**	**8%**
New York	**550,000**	**5%**	**926,298**	**9%**	**968,236**	**9%**	**919,174**	**9%**
Illinois	**540,000**	**5%**	**450,244**	**4%**	**471,309**	**4%**	**466,011**	**4%**
Georgia	**480,000**	**4%**	**323,882**	**3%**	**321,289**	**3%**	**317,099**	**3%**
Arizona	**460,000**	**4%**	**318,739**	**3%**	**311,002**	**3%**	**326,542**	**3%**
North Carolina	**370,000**	**3%**	**266,484**	**2%**	**255,061**	**2%**	**261,586**	**2%**
New Jersey	**360,000**	**3%**	**444,634**	**4%**	**444,557**	**4%**	**430,723**	**4%**
Nevada	**260,000**	**2%**	**164,842**	**2%**	**153,532**	**1%**	**160,574**	**1%**
Other states	**2,730,000**	**25%**	**2,992,264**	**28%**	**3,022,544**	**28%**	**2,858,927**	**27%**

Detail may not sum to totals because of rounding.

Source: U.S. Department of Homeland Security.

Table 4.

Comparison of Total Estimates with Office of Immigration Statistics Residual Estimates

Country of Birth of the Unauthorized Immigrant Population

OIS residual estimates			SIPP model-based estimates		Latent Class (model-1) based estimates		Boosted SIPP-based estimates	
			ACS 2009		ACS 2009		ACS 2009	
Country of birth	**January 2009**	**Percent of total**	**Total estimate**	**Percent of total**	**Total estimate**	**Percent of total**	**Total estimate**	**Percent of total**
Total	**10,750,000**		**10,750,000**		**10,750,000**		**10,750,000**	
Mexico	**6,650,000**	**62%**	**4,865,822**	**45%**	**4,583,566**	**43%**	**5,229,107**	**49%**
El Salvador	**530,000**	**5%**	**478,028**	**4%**	**438,653**	**4%**	**497,099**	**5%**
Guatemala	**480,000**	**4%**	**413,356**	**4%**	**347,778**	**3%**	**421,152**	**4%**
Honduras	**320,000**	**3%**	**243,045**	**2%**	**205,557**	**2%**	**240,414**	**2%**
Philippines	**270,000**	**2%**	**228,521**	**2%**	**269,011**	**3%**	**204,538**	**2%**
India	**200,000**	**2%**	**465,762**	**4%**	**493,209**	**5%**	**403,889**	**4%**
Korea	**200,000**	**2%**	**192,292**	**2%**	**220,526**	**2%**	**181,195**	**2%**
Ecuador	**170,000**	**2%**	**155,081**	**1%**	**135,270**	**1%**	**157,668**	**1%**
Brazil	**150,000**	**1%**	**133,521**	**1%**	**145,677**	**1%**	**118,446**	**1%**
China	**120,000**	**1%**	**290,833**	**3%**	**321,372**	**3%**	**266,022**	**2%**
Other Countrie	**1,650,000**	**15%**	**3,283,740**	**31%**	**3,589,380**	**33%**	**3,030,470**	**28%**

Detail may not sum to totals because of rounding.

Source: U.S. Department of Homeland Security.

Table 3.

Comparison of Total Estimates with Office of Immigration Statistics Residual Estimates

Period of Entry of the Unauthorized Immigrant Population

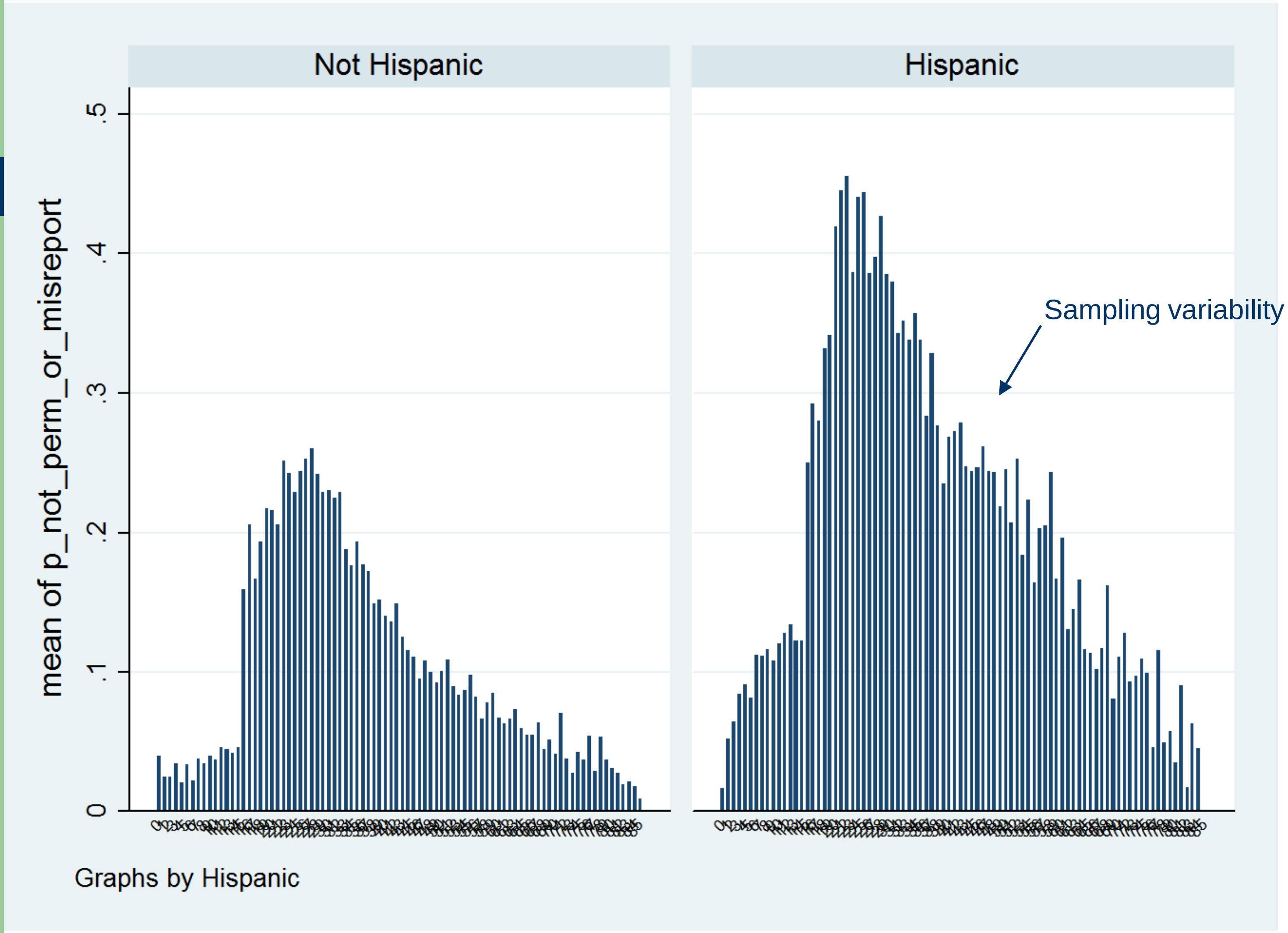
OIS residual estimates				SIPP model-based estimates		Latent Class (model 1)-based estimates		Boosted SIPP-based estimates	
				ACS 2009		ACS 2009		ACS 2009	
Period of entry	**January 2009**	**Percent of Total**		**Total estimate**	**Percent of total**	**Total estimate**	**Percent of total**	**Total estimate**	**Percent of total**
All years	**10,750,000**		All years	**10,750,000**		**10,750,000**		**10,750,000**	
2005-2008	**910,000**	**8%**	2005-2009	**4,566,277**	**42%**	**3,076,479**	**29%**	**3,522,279**	**33%**
2000-2004	**3,040,000**	**28%**	2000-2004	**2,913,867**	**27%**	**3,242,733**	**30%**	**2,934,170**	**27%**
1995-1999	**3,080,000**	**29%**	1995-1999	**1,310,408**	**12%**	**1,911,875**	**18%**	**1,745,499**	**16%**
1990-1994	**1,670,000**	**16%**	1990-1994	**821,096**	**8%**	**1,076,482**	**10%**	**1,111,204**	**10%**
1985-1989	**1,190,000**	**11%**	1985-1989	**572,967**	**5%**	**714,651**	**7%**	**735,983**	**7%**
1980-1984	**860,000**	**8%**	1980-1984	**278,452**	**3%**	**352,132**	**3%**	**344,991**	**3%**
			Before 1980	**286,933**	**3%**	**375,648**	**3%**	**355,873**	**3%**

Detail may not sum to totals because of rounding.

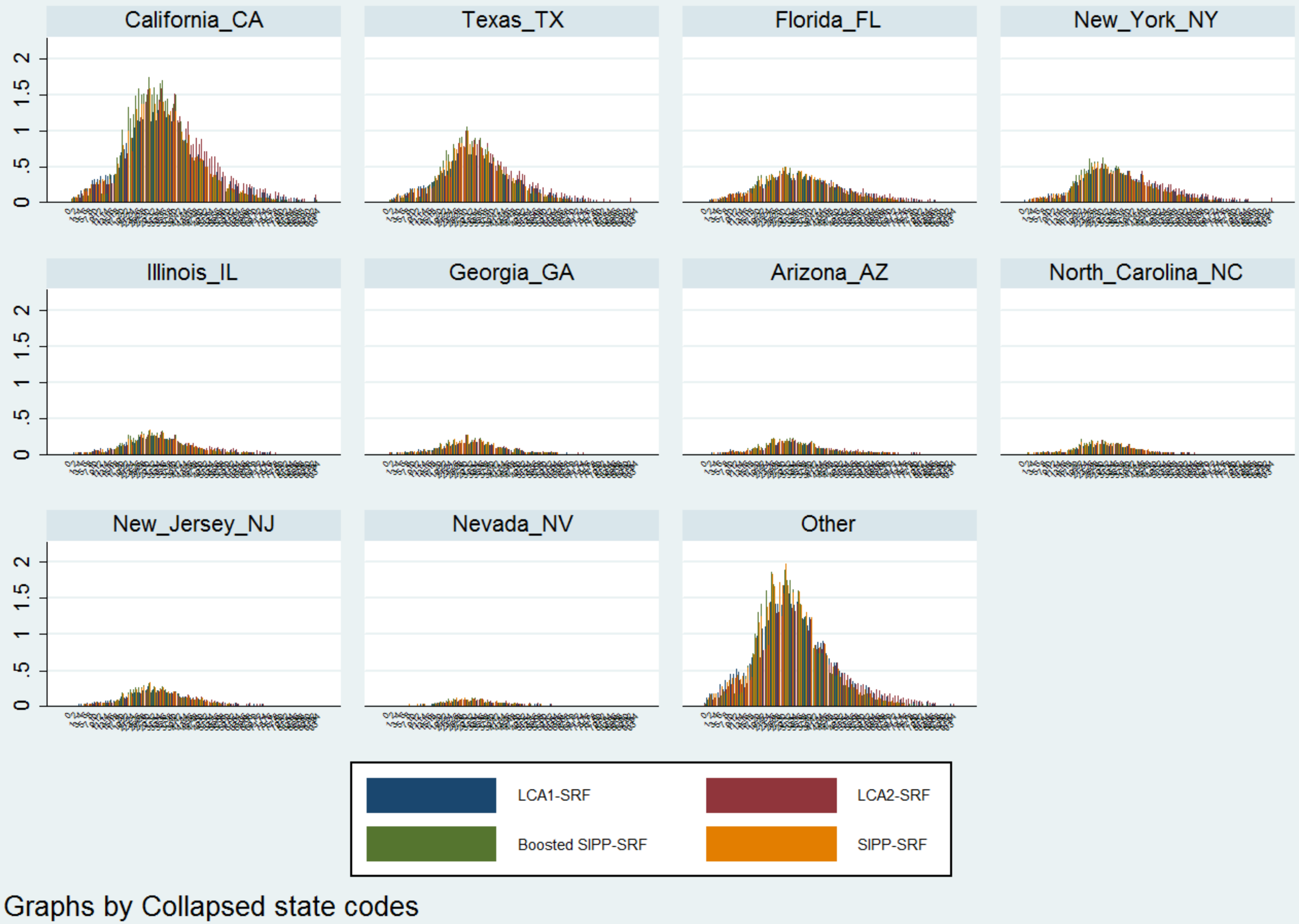
Source: U.S. Department of Homeland Security.

Table 1.

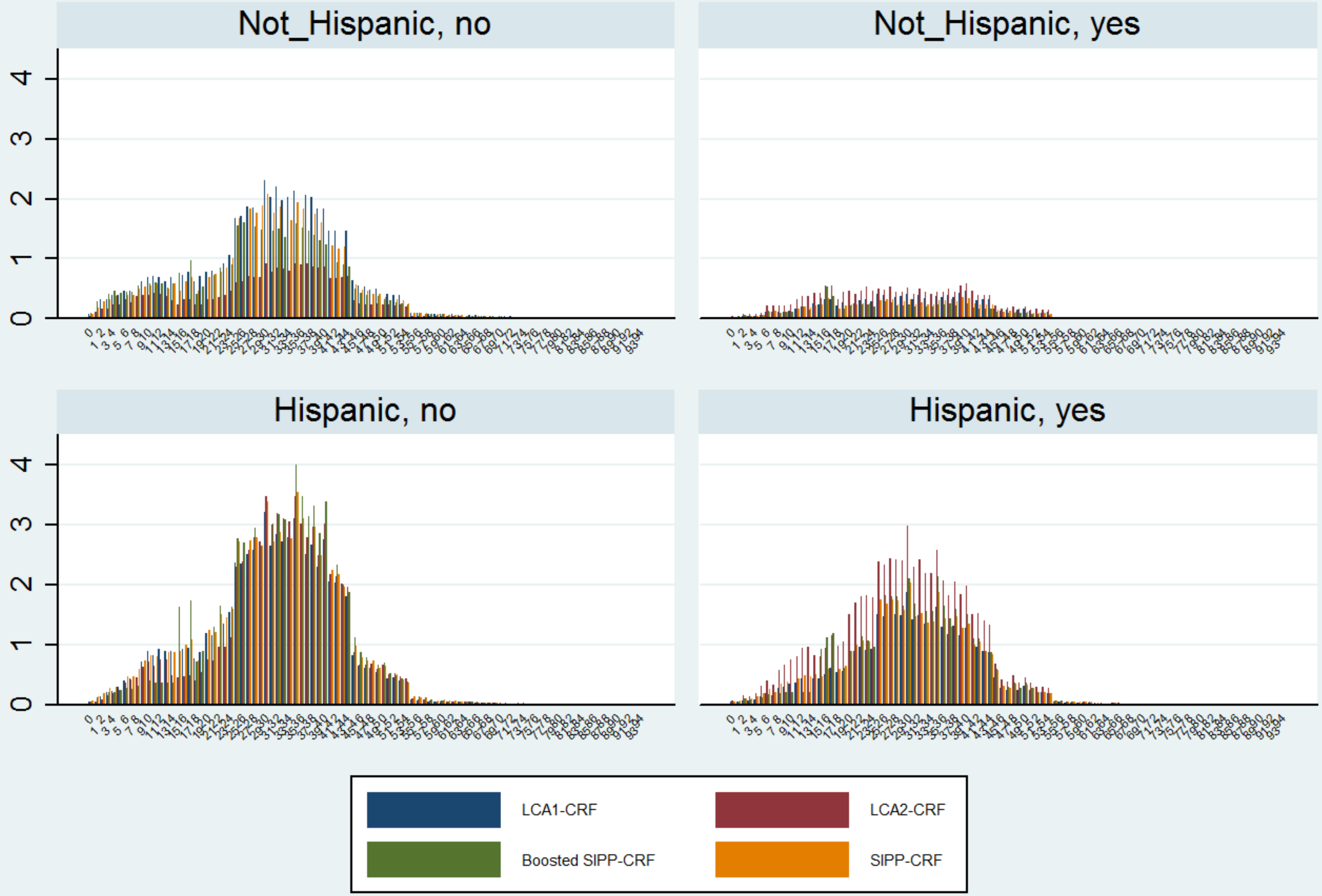
Prob[Nonpermanent] by age and Hispanic Ethnicity



Note: Y-scale in 100,000’s



Note: Y-scale in 100,000's



Graphs by Hispanic and non-Husband/Wife household

Note: Y-scale in 100,000's

Graphs by Numeric sex code (1=Male, 2=Female) and 1-Digit Occupation

Note: Y-scale in 100,000's

Conclusions and Next Steps

- Model coefficients appear consistent with *a priori* beliefs about prediction of legal status (not presented today)
- *Levels* may not match exterior control totals; thus, an external rake may be needed; but once raked, results are consistent across models and with residual results (with some interesting differences, as well)
- Demographic outputs look promising
- A Bayesian approach (incorporating priors about levels) may be a good next step for modeling
- Finalize variance estimation
- Try the model-based method on other surveys
- Submit results to scientific scrutiny