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

Dean H. Judson
Decision Analytics
15000 Fort Trail, Accokeek, MD 20607 (email: dhjudson@comcast.net)

Legislation and policymaking often run ahead of our ability to produce the data needed for implementation and evaluation. In the case of the 2010 Patient Protection and Affordable Care Act (ACA), this has occurred with respect to state-level estimates of the legal status of the immigrant population. Several provisions that affect unauthorized immigrants and are directly relevant to state planning are included in the bill. For example, unauthorized immigrants are exempted from the 'individual mandate' to obtain coverage, are excluded from the Medicaid expansion, and are prohibited from purchasing health insurance coverage in federal or state health insurance exchanges. Because of provisions such as these, unauthorized immigrants are expected to comprise a substantial portion of the remaining uninsured population after the ACA is in full effect. Accordingly, accurate state-level estimates of the size of the unauthorized immigrant population are needed as states estimate the size and characteristics of the populations eligible for the ACA and determine the appropriate safety net capacity for the remaining uninsured population after 2014. However, no Federal agency makes a state-by-state estimate of persons by legal status. Similarly, no specifically-state-representative Federal survey asks "legal status questions" of the foreign-born. This paper focuses on two experimental methods for imputing legal status of foreign-born persons. The first of these involves developing a regression-based imputation model from the Survey of Income and Program Participation and using that model on other target surveys; the other involves developing a latent class model to classify foreign-born persons. Some caveats and other considerations will be discussed.

*This work was funded by the State Health Access Data Assistance Center (SHADAC) at the University of Minnesota, a program of the Robert Wood Johnson Foundation.

Introduction: The Problem of Estimating the Size and Composition of the Unauthorized Population in the U.S.

The 2010 Patient Protection and Affordable Care Act (ACA) will introduce the most significant changes to the US health care system since the introduction of Medicare and Medicaid in 1965. Preparing for those changes has created many challenges at the federal, state, and community levels, as estimates of the size, characteristics, and location of the populations affected by the many different provisions of the ACA are needed. Key among the needed estimates are state-level estimates of the legal status of the immigrant population as several provisions exclude unauthorized immigrants. For example, unauthorized immigrants are exempted from the 'individual mandate' to obtain coverage, are excluded from the Medicaid expansion, and are prohibited from purchasing health insurance coverage in federal or state health insurance exchanges. Because of these provisions, unauthorized immigrants are expected to comprise a substantial portion of the remaining uninsured population after the ACA is in full effect. Accordingly, accurate state-level estimates of the size and composition of the unauthorized immigrant population are needed as states estimate the size and characteristics of the populations eligible for the ACA and determine the appropriate safety net capacity for the remaining uninsured population after 2014.

National estimates show the numbers of foreign-born residents increasing over time, reaching 31.1 million, or 11.1 percent of the US population by 2000 (Malone, Beluja, Costanzo, and Davis, 2003), and 40.2 million by 2010 (Passel and Cohn, 2011). Estimates of the share of that population that is not "authorized" to reside in the United States have varied widely, often with little information on the data and methods used to generate the estimates (Walsh, 2007). Among the studies with documented methods, however, there is more consistency, with the most recent estimates showing 11.2 million unauthorized people, or about one-third of the foreign-born population of the US (Passel and Cohn, 2011).

Obtaining state-specific estimates of the unauthorized population is challenging since no Federal agency makes a state-by-state estimate of persons by legal status and Federal surveys do not typically ask "legal status questions" of the foreign-born. This paper focuses on two experimental methods for imputing legal status of foreign-born persons. The first of these involves developing a regression-based imputation model from the Survey of Income and Program Participation and using that model on other target surveys; the other involves developing a latent class model to classify foreign-born persons.
The **primary** purpose of this imputation exercise is to develop state-level estimates of the unauthorized population. A **secondary**, but potentially still important output, is to generate small-domain estimates of the number of unauthorized immigrants by key subgroups of interest, such as by demographic characteristics, employment, and health status . It is only a **tertiary** goal to place probability values on individual persons; we will examine record-level imputations as a way of building confidence in the models, but record-level imputation is not the focus here. .

The Ambiguous Definition of 'Legal Status'

There is substantial debate as to the exact legal categories that migrants fall into, and what language to use to describe these categories. More importantly, reports from different agencies (and sometimes from different parts of the same agency[1]) use terminology that is either undefined or inconsistent.

A major area of potential confusion in defining immigration status categories involves foreign-born residents who have pending applications. The reason is that (1) these pending applications have different objectives and potential immigration status outcomes, and (2) in some cases, an immigration status may be defined in terms of the reason(s) for estimating immigration status, rather than in strict legal terms.

Regarding (1), for example, a pending application may involve petitioning for:

- a legal immigrant status (such as lawful permanent resident, or LPR),

[1] See, e.g., Costanzo, et. al. (2001), footnote 4.

- a change of temporary or nonimmigrant status (such as a student applying to be a temporary worker), or
- an extension of nonimmigrant status (seeking permission to remain in the United States for a longer period of admission than was originally granted).

Regarding (2), there are legitimate reasons for interpreting the immigration status of the resident foreign-born population in multiple ways. For example, an agency may need to render a legal classification for an enforcement, budget, or program purpose, while policymakers may want to estimate how many foreign-born residents may qualify for or be affected by proposed legislation. Thus, there may be no single, uniquely correct way to estimate numbers of foreign-born residents with pending applications by immigration status.

Classifying foreign-born residents with pending applications is also complicated because many of them:

(1) have also been issued employment authorization documents, also called EADs, and

(2) may or may not be legally present in the United States, and may or may not become legal immigrants (i.e., LPRs with a "green card") or granted some other immigration status.[2]

Having an EAD and/or being illegally present in the United States may complicate identifying a "correct" immigration status for some foreign-born residents who are asked to identify their immigration status in a survey. For example, those who have been issued EADs (1) have some basis for thinking of or identifying themselves as "documented," but (2) may not know or identify themselves according to their correct immigrant statuses. This applies, of course, to direct survey questions rather than aggregate demographic estimates.

There is no apparent consensus by immigration researchers or federal agencies on whether or not (or how) to define some foreign-born residents with pending applications as "unauthorized" (that is, illegally residing in the United States), because their immigration status may be ambiguous. The Department of Homeland Security (DHS), for example, uses the term "unauthorized" to describe foreign-born persons residing illegally in the United States, but includes some "unauthorized immigrants" who have pending applications in the legally resident foreign-born population.[3]

---

[2] Foreign-born residents who encompass a wide range of immigration statuses are eligible to obtain EADs, including those who may be illegally present in the United States. As background, certain aliens who are temporarily in the United States may file a Form I-765, Application for Employment Authorization, to request an Employment Authorization Document (EAD). There are more than 40 eligible categories under which aliens may apply for an EAD, including Temporary Protected Status, Spouse/Dependent of foreign government official, spouse of Class E (treaty trader or treaty investor) nonimmigrant, fiance(e) of a U.S. citizen, Public Interest Parolee, and certain legalization applicants. Some aliens who are issued EADs separately apply for, and can be expected to receive, legal permanent resident (LPR, or "green card") or other status allowing them to live permanently in the United States. Others, such as the spouse of a Class L (intracompany transferee) temporary worker, may leave the United States after their lawful period of admission expires, or become overstays. An overstay is an illegal alien who was legally admitted to the United States for a specific authorized period but remained here after that period expired, without obtaining an extension or a change of status or meeting other specific conditions. Under certain circumstances, an application for extension or change of status can temporarily prevent a foreign visitor's status from being categorized as illegal.

[3] DHS defines "unauthorized immigrants" as "foreign-born persons who entered without inspection or who violated the terms of a temporary admission and who have not acquired LPR status or gained temporary protection against removal by applying for an immigration benefit" (2002 Yearbook of Immigration Statistics. Department of Homeland Security. Washington, D.C.: U.S. Government Printing Office, 2003, p. 213). These definitions are taken from a separate report, "Estimates of the Unauthorized Immigrant Population Residing in the United States: 1990 to 2000." (Office of Policy and Planning. U.S. Immigration and Naturalization Service, January 2003), which includes "unauthorized immigrants with pending I-485 forms—LPR status not yet official by January 1, 2000," as part of the "legally resident foreign-born population—entered 1990-1999 (see Table 3, p. 18). This report is also available at

Since there is no apparent consensus, it is incumbent upon us to be as clear as possible in our use of terminology. Therefore, in order to avoid confusion, we propose the following terminology and definitions for describing each legal status. The major categories are in bold below. The universe of discussion is all foreign-born persons whose place of residence is in the United States on the estimates (July 1) or enumeration (April 1) day[4].

http://www.uscis.gov/graphics/shared/statistics/publications/Ill_Report_1211.pdf (downloaded June 11, 2006).

[4] Depending upon the data source, this will include or exclude the Group Quarters population.

Table 1: Legal Status Terminology

Term	Legal and procedural definition	
(Authorized) Legal Immigrant		
	Naturalized	Has obtained U.S. citizenship.
	Lawful Permanent Resident (LPR)	Has applied for LPR status, and has been formally admitted.
(Authorized) Legal Temporary ("Legal Nonimmigrant")		• Lawful temporary application has been accepted; AND • Terms of admission have not been violated; AND • Neither naturalized nor LPR status has been granted, even if application exists.
Refugee/Asylee		• Has applied for refugee or asylee status and been granted same; OR • Present in the U.S., citizen of Temporary Protected Status-recognized country; AND • Has not converted status to Legal Temporary, Lawful Permanent, or Naturalized.
Residual, Unauthorized* or Other		All other: • Application in process but not yet granted; OR • Entered without inspection; OR • Violated terms of residence.
	Within residual:	
	Quasi-legal	• Has applied for Lawful Permanent Resident, Legal Temporary status, Refugee, Asylee, or Temporary Protected Status; AND • Status not yet granted.

Table notes:

* We believe that this category definition is equivalent to the Office of Immigration Statistics (OIS) definition.

We note for the record that table 1 is primarily a conceptual exercise. Few, if any, data sources literally correspond to these categories: Often questions sufficient to classify cases are not asked, or the data source is a transaction database instead of a person database, or the lag or slippage between the data source and the population of interest is sufficiently great so that the database cannot properly represent the population of interest. This lack of sufficient data is one of the greatest difficulties in estimating the foreign-born population by legal status.

Traditional Approach to Estimating Legal Status: A (Brief) History of the Residual Method

Prior efforts to estimate the size of the foreign-born population by legal status have largely relied on residual methods (Judson and Swanson 2011). We provide a brief review of those methods here, with a focus on the limitations of that approach for the objectives of this paper.

The U. S. Census Bureau defines the foreign-born as people who are not U.S. citizens at birth--a definition that is used in this paper. This population consists of legal immigrants, temporary migrants, and unauthorized migrants as shown in the following equation (Costanzo et al., 2001):

$FB = [L – (M + E) + T + R]$,
where,

FB = Foreign-born population;
L = Legal immigrants;
M = Mortality to legal immigrants;
E = Emigration of legal immigrants;
T = Temporary (legal) migrants; and
R = Residual.

The "residual" is then assumed to be foreign-born unauthorized and quasi-legal migrants.

The preceding equation is a useful way to look at the foreign-born as a whole as well as by status.. By estimating the components and rearranging the equation, we obtain:

R = FB – L + M + E – T, an estimate of the residual population.

The use of the a residual estimate for estimating the foreign-born is primarily, but not exclusively, aimed at estimating those who lack legal documents (see, e.g., Passel, Van Hook, and Bean, 2004) There are several variations of the 'residual" method for this purpose. Many of them are members of the "stock method" in that they attempt to estimate the number without legal documents as the difference between the non-citizen population enumerated in a census or a survey (i.e., the Current Population Survey or the American Community Survey) and the legally resident alien population, where enumerated unauthorized resident migrants = enumerated non-citizens - legally resident aliens. Others are "flow-based" residual methods. In general, the residual methods can be done by national origin, period of entry, age, sex, and depending on the level of detail available in estimates of the legal population, by state and metropolitan area. However, finer levels of detail create a potential problem: If the data sources for FB, L, M,E, and T are from different sources (e.g., Censuses, surveys, administrative systems), then for fine levels of detail there is the potential for a 'negative' residual increases, a problem that is demographically 'embarrassing'.

The Problem with Residual Reasoning

Residual reasoning is easily described: Define the known elements by forming an identity equation that should include every member of some population. Assume one has direct estimates or counts of all subgroups except the subgroup for which no data source exists. Solve for the unknown and "viola" one has the amount which cannot be directly detected. However, when the data sources to estimate each of the components vary (some by a survey, some by administrative data, some by assumption or algorithm) and are subject to error, the logic of residual reasoning, which seems so algebraically neat and tidy, begins to break down. The reality is that, in estimates of the undocumented population based on residual methods, the residual group is assumed to capture the undocumented population along with other population groups not captured accurately in the source data.

Residual reasoning is not by itself problematic: The search for dark matter and dark energy by residual reasoning has the same characteristics (Carroll, 2007). The residual method in demography is analogous. Seeing a population of foreign-born persons:

- First eliminate those that are naturalized,
- Second eliminate those that are lawful permanent residents, ,
- Third eliminate refugees, asylees, and legal temporary persons,:
- Finally, conclude the remainder must be unauthorized persons.

However, the limitation of residual reasoning is the same limitations in earlier inferences about dark matter: Given a choice, we would much rather have a *direct* measure of the population rather than an *assumed residual*. Thus, we address this problem by attempting to estimate the undocumented population directly, rather than by residual methods.

Imputation and the American Community Survey

Our base for this study is the American Community Survey (ACS). The American Community Survey is a continuously fielded household survey administered by the U.S. Census Bureau. It is the successor to the Decennial Census Long-Form Survey and the Supplemental Survey and asks questions on a wide range of topics focusing on the demographics, economic situation, well-being, and program participation of U.S. households and their members. Its large sample size of about 3 million persons per year (drawn from a comprehensive frame of both private residences and group quarters) enables it to be used to generate statistics at state and sub-state levels, albeit estimates made at finer levels of geography than state can require multiple years of data. To allow independent research using ACS, the Census Bureau produces and releases Public Use Microdata Sample (PUMS) files.

Because we do not "know" the legal status of foreign-born persons, and we have highlighted above our reservations regarding using residual methods, we propose to test imputation techniques in this context. Our goal in making imputations is to make a statistical "best guess" as to each person's legal status, and then, to construct state- and domain-level estimates, add up those statistical "best guesses". Imputation has a long history in censuses and surveys (Scheuren, 2005), and the use of imputation for census-taking and survey measurement is well established. There are a variety of techniques used for imputation; the specific technique used depends upon the particular type of variable being imputated (e.g., household count imputation is a different animal than item [individual variable] imputation. Before we launch into our proposed methods, we would like to highlight three "problems" that are addressed in our work.

The Problem of Coverage

It is widely believed by many researchers on the foreign born that the foreign born in general, and the unauthorized foreign born in particular, suffer from more extensive undercoverage in censuses and surveys than non-foreign-born persons. Judson (2009) documents some of these beliefs, and examines evidence that is almost uniformly consistent with, but does not definitively prove, such a belief. Censuses and surveys attempt to correct for undercoverage via post-stratification to population controls derived from population and housing estimates, but it is not known whether controls to age, race, sex and ethnicity correctly account for supposed undercoverage of the foreign born.

This point raises a problem for deriving estimates based on imputations. If there is either a general coverage problem associated with the foreign born (again, in particular, the unauthorized foreign born), then population totals could be askew—perhaps only a little, if ACS coverage corrections are robust, but perhaps quite a bit if not so. This is not uniquely true for estimating levels of the foreign born *only*, but it is *certainly true* for this group.

In addition, if there is any kind of 'social desirability bias' associated with foreign-born-related questions, that bias will also affect levels for estimates. For example, Passel and Cohn (2011) negate the reported naturalized citizenship status of some foreign-born respondents because their characteristics (specifically, reported year of entry[5]) are not consistent with being a citizen according to U.S. immigration law. In effect, they believe that too many foreign-born persons are counted as 'naturalized', thus downwardly affecting the levels of non-naturalized (and hence unauthorized) as well.

Both kinds of bias could be present in the data, thus leading to a need for corrections involving control totals.

The Problem of 'Control Totals'

If levels of foreign-born population estimates are potentially biased downward, and non-naturalized but authorized and unauthorized foreign born in particular are potentially downwardly biased (again, we use the word 'potentially' because evidence is somewhat sketchy on these beliefs), then it is natural enough to

[5] The "year of entry" is derived from the question in the CPS that asks "When did you come to live in the United States?" The implicit assumption underlying the question is that there is a single entry into the U.S., which ignores circular migration and temporary residency.

imagine using post-strata that are specifically designed to correct for foreign-born legal statuses. But where should such control totals come from?

Passel and Cohn (2011), and other, earlier, estimates, use control totals that are themselves derived using the residual method; they then use iterative proportional fitting (or 'raking') to bring the algorithmically-derived legal statuses into harmony with external controls. This method is common and is a near-cousin to the methods used for post-stratification in 'typical' surveys. Currently, their estimation models use the Current Population Survey as the 'base' from which LPR and other quantities are subtracted.

The Office of Immigration Statistics (OIS; Hoefer, Rytina and Campbell, 2006; Hoefer, Rytina, and Baker 2009) also uses a residual method to generate totals. Their base, however, is the American Community Survey, from which LPR and other quantities are subtracted. For the record, because the Passel and OIS systems are similar, their final estimates results typically are very comparable.

Because OIS uses the same data base as this paper, are considered the "official" U.S. government estimates, and because OIS publishes national totals by age and sex, for this estimation system we have chosen to use OIS totals. However, we wish to be very 'light handed' in applying control totals, so as to let the imputations do most of the 'talking'. We will use two approaches:

1) The 'simple rake factor' (SRF) approach will control *only* to the total national estimate of unauthorized foreign born, leaving everything else to the imputations; and
2) The 'complicated rake factor' (CRF) approach will control to the total national estimate by broad age/sex categories, leaving everything else to the imputations.

The Problem of Variance Estimation

Traditional residual methods have struggled mightily with the problem of uncertainty or variance estimation. If this year's sum of unauthorized is 11.5 million persons and last year's was 11.2, is that change 'statistically significant'? Further, when data from surveys and administrative systems, combined with assumptions about emigration and mortality, what even can be said about uncertainty? Judson and Cornwell, 2010, made a first attempt, using ACS sampling variances and approximating administrative records "uncertainty" using a time-series approach, but other work in the literature has not addressed the uncertainty of the estimates.

In this paper, there are two sources of uncertainty: The sampling uncertainty associated with the ACS design, and the imputation uncertainty associated with the probability prediction. In principle, if we treat the imputations as approximately independent of the sample design, for a sample total, we have:

$$V(\hat{T}) = V(\hat{I})V(S),$$

where,

$V\left(\hat{T}\right)$ is the total population estimate variance,
$V\left(\hat{I}\right)$ is the imputation variance, and
$V\left(S\right)$ is the sampling variance.

Again in principle the sampling variance is straightforward (although we will complicate matters somewhat in a moment). The Census Bureau provides balanced repeated replication (BRR) replicate weights and a formula for estimating variance for random variable X_0 (X_0 is the point estimate of interest):

$$V(X_0) = \frac{4}{80}\sum_{i=1}^{80}\left(X_i - X_0\right)^2$$

For the imputations, we can treat the imputation as the output of a statistical model, and calculate it directly:

$V(\hat{I}) = MSE_{Model}$,
and merely combine the two multiplicatively.

Let us return momentarily to the replicate weights. The general formula provided for the ACS would work just fine, except for one conundrum: By using the simple and complex rake factors, we have *changed* the original person weight by raking. The replicates are based on the original person weight, and thus would misrepresent the sampling variability in our *raked* estimates. This problem is found in replicate weights generally; as noted by Korn and Graubard (1999:34):
"…theoretically one should recalculate the sample weights for each replicate. However the analyst may not have enough detailed information about these adjustments to perform these calculations…"
This is, of course the situation here. We do not have information to replicate rake factors (either simple or complicated) R_i, for the ith replicate. Therefore, we are going to settle for a reasonable approximation, $R_i = R, \forall i$, and obtain:

$$V(X_0) = \frac{4}{80}\sum_{i=1}^{80}(R_i X_i - RX_0)^2 \cong \frac{4}{80}\sum_{i=1}^{80}(RX_i - RX_0)^2 = \frac{4R^2}{80}\sum_{i=1}^{80}(X_i - X_0)^2$$

for our variance estimate. Based on empirical results in table 2.5-2 of Korn and Gruabard, 1999: 36), it appears that this approximation slightly inflates the variance estimate, making this estimate "conservative". For the purposes of this paper, we will be focusing on point estimates rather than variance estimation; however, we recognize its importance to the overall system and intend to incorporate proper variance estimation in the near future.

Alternate Approach #1: Using the Survey of Income and Program Participation to Develop an Imputation Model

In 2006 the first author tested the use of the Survey of Income and Program Participation to develop an imputation model. This model takes advantage of key migration legal status questions that are asked in wave two, on the 'migration' topical module. The questions asked include:

1) When ... moved to the U.S. to live, what was ...'s immigration status?

A follow-up question included:

2) Has ...'s status been changed to permanent resident?

A person who indicated "other" (i.e., not permanent or refugee) to the first question, and "No" to the second question, could reasonably be considered to still be in the "other" category as of the survey date. Furthermore, as with Passel, certain foreign-born persons report themselves "naturalized" but with a year of entry too recent to likely be an accurate report. Thus the sum total of the first group "other" and not adjusted plus the probable misreport on citizenship, represents our target variable of undocumented immigrants.[6]

Using this classification as a guide, a cross-sectional logistic regression model can be constructed; the resulting right-hand-side variables of this model, then, generate a predicted *probability* that the person is in the "other" category. This model, when applied to the American Community Survey, provides a probabilistic imputation, a form of so-called "cold deck" imputation (Lohr, 1999). Judson, 2006, reports the estimates derived from this assumption; it appeared at the time that a possible "social desirability bias" might generate estimates that were generically too low, relative to competing methods of estimation.

[6] In the SIPP, year of entry is derived from the question: "When did this person come to live in the United States?" Respondents who "came to live" in the U.S. more than once were asked to report their most recent year of entry. It may be that some of those who are assumed to be a probable misreport on naturalized status because of a "too recent" year of entry have had moved between residing in the U.S. and another country.

Alternate Approach #2: Using Latent Class Analysis to Develop an Imputation Model

An intriguing idea was first proffered in Judson and Swanson (2010): Although a foreign-born person's legal status is unknown, if a clustering of lawful and unauthorized classes exist, a natural method to attempt to discover those classes is latent class analysis (the statistically-principled successor to the older cluster analysis technique). It, too, would output a *probability* of belonging to each class. The latent class method has the inherent liability that the *researcher* chooses how to label each class (in this case "unauthorized" versus "not unauthorized"). This labeling, one must admit, is partly arbitrary; if, however, the resulting output is comparable across methods, then the validity of each will be increased.

Results From the Models and Comparisons to Other Systems

As stated in the introduction, the primary goal of these imputation exercises is to produce state-level estimates, with small domain estimates as a secondary goal, and individual-level verisimilitude a tertiary goal. Thus, we will focus on aggregate results (part 1), and only touch on individual microdata results (part 2). Likewise, specific model specifications will not be presented here, as they have been presented elsewhere (Judson, 2011)[7].

Aggregate Results

Table 2 below, tabulates the four latent class models LC1-LC4 and SIPP imputation results, for the fifty states. For LC1, we have presented results for the simple rake factor (SRF) and complicated rake factor (CRF). Two SIPP models are presented, the first based on a right hand side specification described in Judson (2006, update in 2011); the second is based on submitting the same collection of right hand side variables to an automated "boosting" procedure—boosting being a data mining tool the selects variables and interaction effects automatically. (More information can be found in Schonlau [2005], or on http://www.schonlau.net/. A more general discussion is found in Ridgeway, 1999.)

[7] Details on model specifications are available from the authors.

Table 2: 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)	(LC1-CRF)	(LC2-CRF)	(LC3-CRF)	(LC4-CRF)	(SIPP)	(Boosted SIPP)
	Total	Total	Total	Total	Total	Total	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
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

As can be seen, despite the wide diversity of specific implementation, the point estimates at the state level are highly consistent with one another. Because they are so consistent with one another, we will cease presenting large sets of estimates, and focus on only three: Basic SIPP model, Latent class model 1, and Boosted SIPP model, each using the complicated rake factor.

The fundamental question for these estimates is: Are they consistent with other published estimates, particular those of OIS? Table 4 presents a comparison with published tables for the largest states in the United States. Note that OIS does not present results for states other than those presented in table 3, so further comparisons with OIS are not available at this time.

Table 3: Predicted State of Residence of the Unauthorized Immigrant Population

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.

As can be seen, both methods generate very similar results for the largest states. (Note, again, that the rake factors did not take state into account, only total population [for the simple rake factor] and broad age/sex groups [for the complicated rake factor].)

However, state totals are not the end of the matter; table 4 exhibits comparisons with published OIS data on period of entry, and table 5 exhibits comparisons with published OIS data on country of birth.

Table 4: Predicted Period of Entry of the Unauthorized Immigrant Population

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%**
	0	**0%**	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.

Here we begin to see some differences of note—basic SIPP, latent class, and boosted SIPP all seem to be pointing to more unauthorized persons with recent period of entry than that generated by the OIS residual method. Furthermore, the OIS system automatically makes anyone whose period of entry is 1979 or earlier authorized—they assume that the legal reforms of the 1980's would result in regularized legal status. Further, year of entry may not be conceptualized exactly the same across surveys and administrative records, both of which are used in the OIS system. However, the three models suggest that there might be hundreds of thousands of persons still unauthorized among those whose period of entry is 1979 or earlier.

Finally, we compare estimates by country of birth.

Table 5: Predicted Country of Birth of the Unauthorized Immigrant Population

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 Countries	**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.

Again we see some differences worthy of note—we see that the three models generate fewer Mexican unauthorized persons than the OIS residual method, and substantially more persons from other countries not on the list. The first finding is consistent with news reports averring that the recent Mexican census found more Mexicans, men in particular, than expected, and a suspicion is that this finding is a result of return migration and lower Mexican emigration in response to economic conditions (Cave, 2011). The second finding does not have the same "obvious" interpretation, but may reflect a trend in unauthorized immigration from other countries that, if true, could be of important policy interest.

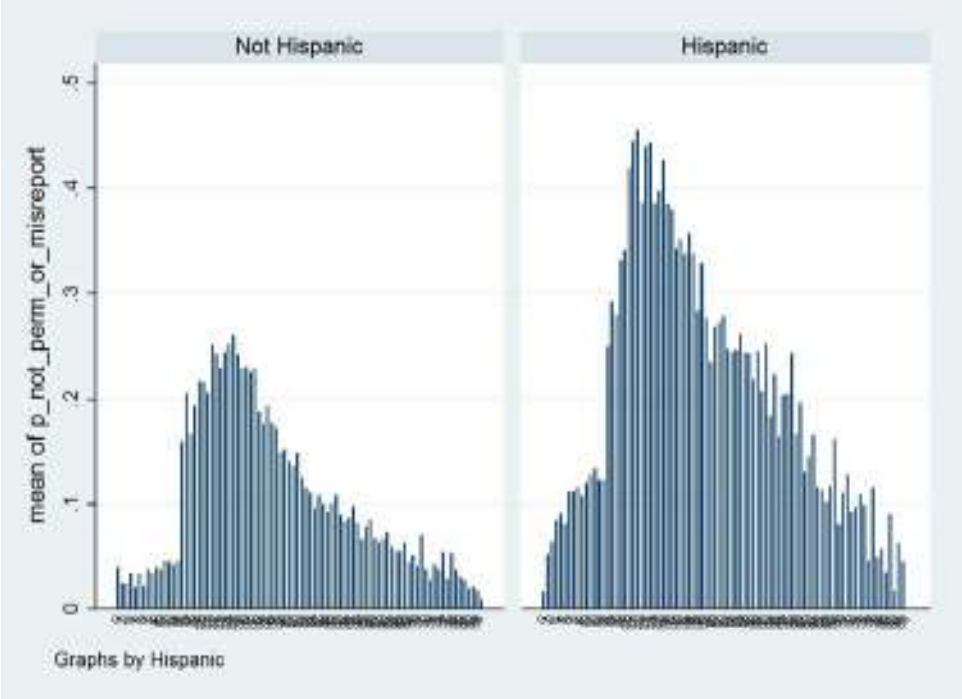
Graphical Analyses

Finally, we ask two questions:

1) Do the imputation probabilities, when added up to make population estimates, make demographic sense? That is, are the people that we expect to make up the bulk of the unauthorized, from other sources, in fact show up in our data?
2) Do the various imputation schemes, in particular SIPP with simple rake factor (SIPP-SRF), SIPP with complicated rake factor (SIPP-CRF), Latent Class with simple rake factor (LCA1-SRF), and latent class with complicated rake factor (LCA1-CRF), all hold together?

Figure 1 summarizes the age distribution of the SIPP model-based imputations by Hispanic/not Hispanic ethnicity.

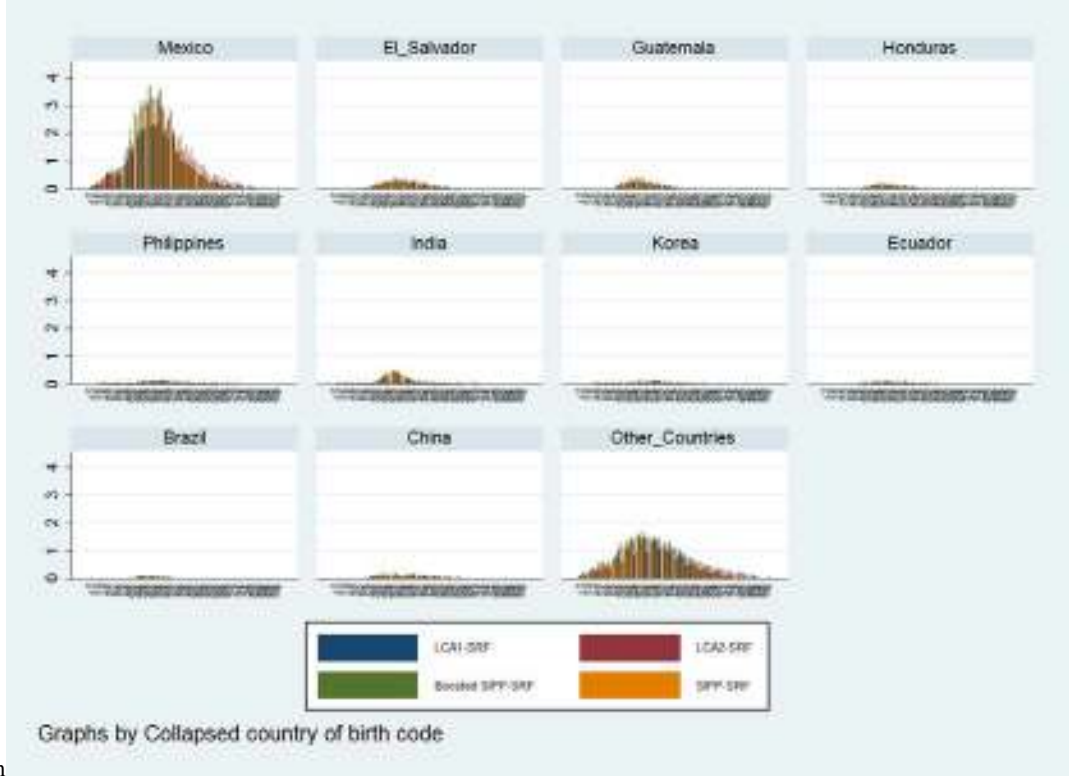
Figure 1: Predicted Age Distribution of the Unauthorized Immigrant Population, by Hispanic Ethnicity:



As can be seen in Figure 1, probabilities peak at approximately the prime migration ages, decline both before and after, and are more concentrated amongst Hispanics than non-Hispanics. Note that the 'spikiness' in the data represents sampling variability.

Figure 2 graphs the four models' results by age distribution, by country of birth code. (Note that the Y-scale is in 100,000's for readability.)

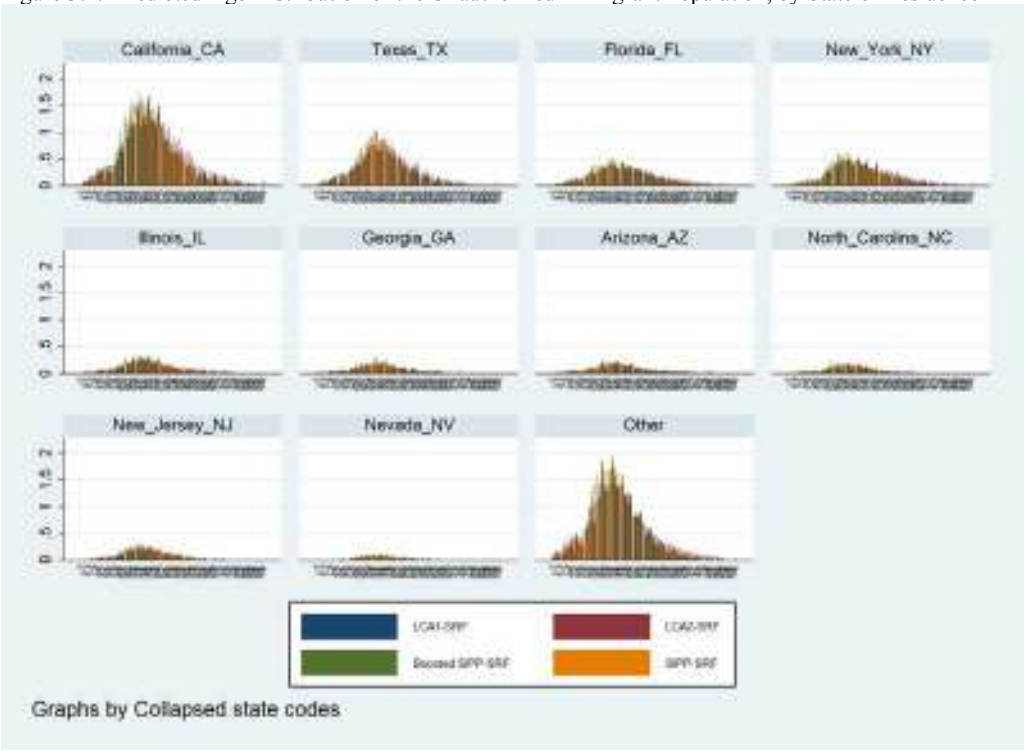
Figure 2: : Predicted Age Distribution of the Unauthorized Immigrant Population, by Country of Birth



Again, with slight exceptions, the four models are in general agreement, although there appear to be some areas of difference in Mexico at peak migration ages that is worth examining more closely.

Figure 3 performs the same exercise by state of residence (again, Y-scale in 100,000's).

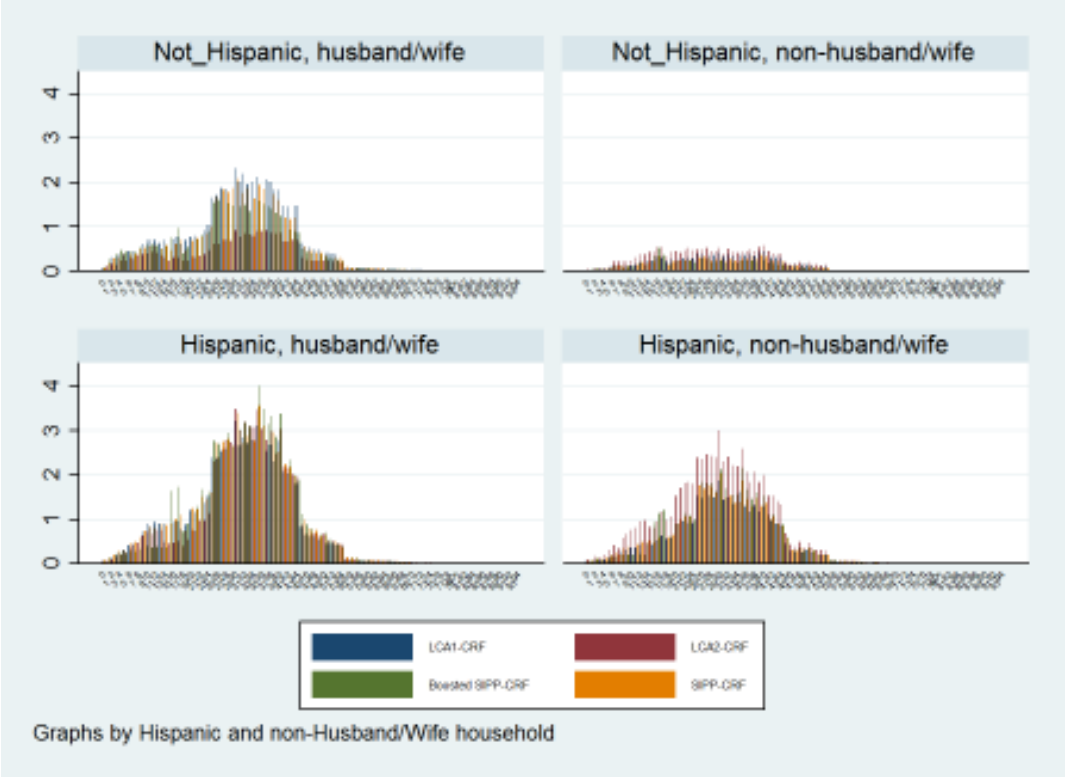
Figure 3: : Predicted Age Distribution of the Unauthorized Immigrant Population, by State of Residence



Again, broad consistency is obtained.

Our final two figures illustrate the "small domain estimation" capabilities of these models. Out of any number of ACS domains we could have chosen to illustrate, we have selected two: Numbers of persons living in non-husband/wife households, and occupation code by broad occupational groups. We present the estimated age distribution for each group.

Figure 4 : Predicted Age Distribution of the Unauthorized Immigrant Population, by Ethnicity and Household Marital Status



In Figure 4, "non-husband/wife" indicates persons living in non-husband/wife households, "husband/wife" the converse, and we can see some differences in the upper left panel and lower right panel. Substantively, the right upper and lower panels indicate that Hispanic persons, particularly those in prime migration ages, are far more likely to live in non-husband/wife households (lower right panel) than non-Hispanic persons are (upper right panel).

Finally, to illustrate the power of this system, we graph something very specific--distributions of unauthorized persons by occupation group and by sex (again, Y-scale in 100,000's) in Figure 5.

Figure 5: Predicted Age Distribution of the Unauthorized Immigrant Population, by Occupation Group and Gender

In this figure, occupations are:
MGR/BUS: Managerial business, and finance;
CMM/ENG/SCI: Computers, engineering and science;
CMS/EDU/LGL/ENT: Counselors, education, legal, and entertainment;

MED/HLS/PRT: Medical, health Services, public protection;
EAT/CLN/PRS/SAL: Eating, Cleaning, personal services, sales;
OFF: Office;
FFF/CON/EXT: Forestry, fishing, agriculture, construction, extraction;
RPR/PRD: Repair and production;
PRD: Production; and
TRN/MIL: Training and military.

As can be seen, our system estimates unauthorized persons almost exactly where anecdotal evidence says they should be (males in FFF/CON/EXT and in EAT/CLN/PRS/SAL; females in EAT/CLN/PRS/SAL but not FFF/CON/TEXT) and very few unauthorized persons in occupations unlikely to be attractive due to legal constraints.

Microdata Results

As noted above, record-by-record (i.e., individual observation in the survey) accuracy is only a tertiary goal of this project; however, to the extent that record-by-record results remain consistent with one another, then the overall validity of the approach is enhanced. To this end, we merely wish to examine simple correlation coefficients of the raked-probability imputations, to determine whether the systems "hang together" at a record level. Table 6, below, presents these simple (unweighted) correlations.

Table 6; Simple Correlations Across Models Predicting Unauthorized Immigration Status

	Boosted SIPP - SRF	Boosted SIPP - CRF	Base SIPP - SRF	Base SIPP - CRF	LCA Model 1 - SRF	LCA Model 1 - CRF
Boosted SIPP - SRF	1					
Boosted SIPP - CRF	0.8736	1				
Base SIPP - SRF	0.7683	0.7161	1			
Base SIPP - CRF	0.7387	0.7967	0.9501	1		
LCA Model 1 - CRF	0.4176	0.4073	0.5664	0.5315	1	
LCA Model 1 - SRF	0.5587	0.678	0.5553	0.7005	0.6077	1

As can be seen, for most of the imputations, the correlations are positive (as expected) and generally are strong. Where they are lower, graphical analyses suggest that a failure of linearity is at issue, as the pearson correlation coefficient is a measure of linear relationship, and these imputations appear to depart from linearity.

Conclusions

As can be seen, where external comparisons are available (published tables from the Office of Immigration Research), these (point) estimates generally fare well—both the residual-based OIS estimates and these estimates are very comparable, for the states for which OIS publishes data. The "simple rake factor" models, both SIPP-based and latent-class based, generate what appear to be more older unauthorized persons than would be expected based on residual methods. We are left with the question: Is this really true or is this a technical artifact of the different methods, or, further is this the result of a discrepancy between year of entry in the survey versus administrative data?

In order to demographically harmonize the OIS and these estimates, we introduced the (so-called) "complicated rake factor", which rakes by broad age and sex groups. These results are obviously demographically compatible with OIS results (by design), but they retain state-level values that are comparable to those obtained using the simple rake factor—no harm appears to be being done to the state-level estimates of the distribution of the undocumented population.

Next steps in this project include variance estimation, small domain estimation results (e.g. industry, occupation) to continue to assess the validity of the results relative to other data sources, and further development of the statistical models (both SIPP, boosted SIPP and latent class) that form the base of this estimation system.

These results, as they are, do generate novel outcomes by country of birth and period of entry—raising the question as to which set of estimates, OIS or model-based, are more "accurate." It is far too early to attempt to answer this question. But, we now have two completely different methods to estimate the size of the unauthorized population—two methods that are broadly consistent with each other, but differ in important ways—and so we now have new tools to use to try to understand the characteristics of that population. By taking advantage of the large sample size and relatively detailed characteristics available in the American Community Survey, we can say more about the unauthorized population than was possible before, providing much needed information for states as they prepare for the forthcoming changes under the Affordable Care Act.

References

Carroll, S. 2007. *Dark Matter, Dark Energy: The Dark Side of the Universe*. Chantilly, VA : Teaching Co.

Cassidy, R. 2004a. *Involuntary and Voluntary Migrant Algorithm*. https://www.sabresystems.com/whitepapers/AMS Deliverable 3 020305.pdf., last accessed, April, 2008.

Cassidy, R. 2004b. *Involuntary and Voluntary Migrant Estimates*. https://www.sabresystems.com/whitepapers/AMS Deliverable 5 020305.pdf., last accessed, April, 2008.

Cave, D. (2011). *Better Lives for Mexicans Cut Allure of Going North*. http://www.nytimes.com/interactive/2011/07/06/world/americas/immigration.html, last accessed, March, 2012.

Costanzo, J., Davis, C., Irazi, C., Goodkind, D., and Ramirez, R., 2001. Evaluating Components of International Migration: The Residual Foreign-Born. (Population Division Working Paper #61) (December 2001) U.S. Census Bureau. http://www.census.gov/population/www/documentation/twps0061.html, last accessed August, 2008)

Fortuny, K., R. Capps, and J. Passel. 2007. The Characteristics of Unauthorized Immigrants in California, Los Angeles County, and the United States. Washington DC: The Urban Institute (Available online, http://www.urban.org/UploadedPDF/411425 Characteristics Immigrants.pdf, last accessed April, 2008).

Gerstle, G. 2008. "The Immigrant as a Threat to American Security: A Historical Perspective." pp. 217- 245 in E. Barkan, H. Diner, and A. Kraut (EDs.). From Arrival to Incorporation: Migrants to the U.S. in a Global Era. New York City, NY" New York University Press.

Hoefer, M., Rytina, N., and Campbell, C. 2006. Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2005. http://www.uscis.gov/graphics/shared/statistics/publications/ILL_PE_2005.pdf , last accessed, July, 2008).

Hoefer, M., Rytina, N., and Baker, B. C. 2009. Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2008. http://www.dhs.gov/xlibrary/assets/statistics/publications/ois_ill_pe_2008.pdf, last accessed, July, 2009).

Judson, D. 2006. "Memorandum for E. Lamas (21 September): Subject: Preliminary Results and Evaluation of Experimental Estimates of the Foreign-Born Population by Legal Status. Document available upon request.

Judson, D. H. 2009. Coverage of the foreign born in censuses and surveys: What do we think, what do we know and what can we prove? Unpublished working paper in progress, available from the author.

Judson, D.H. 2011. Imputing the Legal Status of the Foreign Born Population Using Model-Based Methods. Paper presented to the 2011 Western Economic Association International, June 20th, 2011.

Judson, D.H., and Cornwell, Derekh (2010). Estimating Uncertainty Bounds for Complex Demographic Models: A Case Study of Estimates of the Unauthorized Population in the U.S. Paper presented to the 2010 Joint Statistical Meetings, August, 2010, Vancouver, B.C.

Judson, D., and Swanson, D. 2010. Estimating Characteristics Of The Foreign-Born By Legal Status: An Evaluation Of Data And Methods. New York, NY: Springer Briefs.

Korn, E.I., and Graubard, B.I. 1999. Analysis of Health Surveys. New York, NY: Wiley.

Larsen, L. 2004. "The Foreign-Born Population in the United States: 2003." Current Population Reports P20-555. Washington, DC: U. S. Census Bureau.

Lohr, S. 1999. Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press.

Malone, N., Baluja, K., Costanzo, J. and Davis, C. 2003. The Foreign-Born Population: 2000. Census 2000 Brief C2KBR-34. Washington, DC: U.S. Census Bureau.

National Research Council. 2006b. Multiple Origins, Uncertain Destinies: Hispanics and the American Future. Washington, DC: National Academies Press.

Orrenius, P. and M. Zavodny. 2006. "Did 9/11 Worsen the Job Prospects of Hispanic Immigrants?" Research Department Working Paper 0508. Dallas, TX: Federal Reserve Bank of Dallas (Available online, http://209.85.173.104/search?q=cache:QS902hgenocJ:www.dallasfed.org/research/papers/2005/wp0508.pdf+review+passel+2005+%22Estimates+of+the+Size+and+Characteristics+of+the+Unauthorized%22&hl=en&ct=clnk&cd=9&gl=us, last accessed May, 2008).

Passel, J. 2005. Estimates of the Size and Characteristics of the Unauthorized Population. Washington, DC: Pew Hispanic Center, Washington.

Passel, J., J. Van Hook, and F. Bean. 2004. "Estimates of the Legal and Unauthorized Foreign-Born Population for the United States and Selected States, Based on Census 2000." Arlington, VA: Immigration White Papers, Sabre Systems (available online, http://www.sabresystems.com/sd_whitepapers_immigration.asp, last accessed December 2007).

Passel, J. and Cohn, V. 2011. Unauthorized Immigrant Population: National and State Trends, 2010. (Available online, http://pewhispanic.org/files/reports/133.pdf, last accessed November 2011).

Schmidley, A.D. 2001. "Profile of the Foreign-Born Population in the United States: 2000." Current Populations Reports, Special Studies, P23-206. Washington DD: U.S. Census Bureau. (http://www.census.gov/prod/2002pubs/p23-206.pdf , last accessed April 2008).

State of Minnesota. 2005. The Impact of Illegal Immigration on Minnesota: Costs and Population Trends. St. Paul, MN: Office of Strategic Planning and Results Management, Minnesota Department of Administration (http://www.state.mn.us/mn/externalDocs/Administration/Report_The_Impact_of_Illegal_Immigration_on_Minnesota_120805035315_Illegal%20Immigration%20Brief%2026.pdf, last accessed May, 2008).

Ridgeway, G. 1999. The State of Boosting. Computing Science and Statistics, 31: 172-181.

Scheuren, F. 2005. Multiple Imputation: How it Began and Continues. The American Statistician. 59(4): 315-319.

Schonlau, M 2005. Boosted Regression (boosting): A Tutorial and a Stata plugin. The Stata Journal, 5(3):330-354.

Walashek, P. and D. Swanson. (2006). The Roots of Conflicts over US Census Counts in the late 20th Century and Prospects for the 21st Century." Journal of Economic and Social Measurement 31 (3, 4): 185-206.

Walsh, J. 2007. "Illegal Aliens: Counting the Uncountable." The Social Contract (4): 216-223. (http://www.thesocialcontract.com/artman2/publish/tsc_17_4/index.shtml, last accessed April 2008)