Estimating the Eligible-to-Naturalize Population

By Manuel Pastor and Justin Scoggins
March 8, 2016

This memo explains the method we at the University of Southern California (USC) Center for the Study of Immigrant Integration (CSII) use to estimate the eligible-to-naturalize population in the United States. This necessarily involves a rather lengthy discussion of estimating the undocumented population; that is the first and most crucial step to estimating the eligible-to-naturalize since once that group is determined, the remainder of the non-citizen foreign-born residents are mostly Lawful Permanent Residents (LPRs) and the criteria that can then be applied to that group to determine LPRs eligible to naturalize is fairly straightforward.

Accordingly, the bulk of this memo describes the CSII method of estimating the undocumented, comparing that at times to the approach taken by experts at the Migration Policy Institute (MPI) and the Center for Migration Studies (CMS). The memo also includes some discussion of estimates from the Pew Research Center but both MPI and CMS have been a bit more explicit about their methods and that facilitates comparison; comparison with CMS is particularly important since CMS has also provided data on the eligible-to-naturalize. This memo also compares our approach and theirs as a way of informing users as to relative strengths and contributions. Finally, the last section of this memo describes the approach taken to generate estimates of the eligible-to-naturalize adult population at the various levels of geography which are featured in a new CSII interactive map available at: http://dornsife.usc.edu/csii/eligible-to-naturalize-map/.

Estimating the Undocumented Population

The first step in determining who is eligible to naturalize is determining who in the non-citizen foreign-born population is likely to be an LPR—and that requires netting out those who are unauthorized or undocumented. Of course, estimating the undocumented population is a challenging exercise since it involves a series of assumptions and estimation strategies that must be combined to derive defensible numbers. Fortunately, the state of the art has evolved and researchers have become increasingly clear about their methods, allowing for other researchers to replicate and modify approaches.

In this exercise, we adopted an increasingly common strategy that involves first determining who among the non-citizen population is least likely to be undocumented due to a series of conditions (a process called "logical edits") and then sorting the remainder into documented and undocumented based on a series of probability estimates (applied in a way to reflect the underlying distribution of probabilities). We applied this technique to a pooled 2010-2014 version of the American Community Survey (ACS) microdata; the actual data we use came from annual ACS surveys provided by IPUMS-USA, which CSII "self-pooled" into a single sample (Ruggles et al., 2010). We sought to use a pooled sample to increase the sample size so that we could generate more reliable estimates at the Public Use Microdata Area (PUMA) level, which serves as the basis for estimating the eligible-to-naturalize at other levels of geography.

Logical Edits

We start the estimation by assuming that the aggregate total of undocumented adults in the U.S. is similar to those reported in the most recent estimates from the Office of Immigration Statistics (Rytina, 2013), the Migration Policy Institute (Capps, Bachmeier, Fix, & Van Hook, 2013), and the Center for Migration Studies (Warren, 2014). The resulting aggregate number, particularly from the most recent estimates by MPI and CMS (and also consistent with projected trends in the Pew estimates), is around 11 million.

We then take every non-citizen, non-Cuban foreign-born respondent in our pooled ACS microdata sample and assign to each of these respondents an initial documentation status based on certain characteristics for which information is available in the ACS microdata. In the literature, this process is called "logical edits." For example, we assumed that any non-citizen, non-Cuban immigrant with military experience is an LPR. Other characteristics that led a respondent into LPR status include whether or not the respondent: worked for the public sector; had an occupation that required documents (such as police officer); received social security or disability payments; or was a household head or spouse in a household receiving food stamps but did not have a child in the house (who could have been the legal source of the assistance).

We also assume those who immigrated as adults and were currently enrolled in higher education to be LPRs, on the grounds that they were likely student visa holders and not among the undocumented population. Additionally, we also assume, as do others (for example Warren, 2014) that any immigrant who arrived before 1982 was able to make their way to legal status through the Immigration Reform and Control Act of 1986. Since we are doing this as a pooled sample, we technically utilized the difference between the last year of the pool and 1982 and apply that to all sample years so we have a smattering of individuals who arrived prior to 1982 who are not eliminated in the logical edit procedure. Finally, we place respondents in the LPR category if they received Medicare, Veterans Affair Care, or Indian Health Services.

Compared to some other researchers, particularly at CMS, we apply a few less conditions to designate status to respondents. For example, we did not assume that reporting Medicaid was sufficient to designate one as documented (as does CMS (see Warren, 2014) for men older than 19 and women older than 19 who did not give birth to a child in the last year). The reason is that while the variable labels are reported as "No insurance through Medicaid" and "Has insurance through Medicaid," the actual question asked is ambiguous and could be interpreted as asking about any kind of public health assistance such as emergency care at the hospital. It is also the case that in several states, including California, which has a high share of undocumented residents, some undocumented immigrants in certain categories (some of which Warren considers, such as post-pregnancy) are eligible for the state version of Medicaid and would thus likely answer yes. Children are also eligible and parents could answer based on their children. In any case, previous research does suggest that there are users of such services who are undocumented, so this seems like one logical edit too many. The key point to remember here is that CMS has more individuals assigned by logical edits than we do, a point we return to below.

Probability Edits

The initial method of status assignment leaves us with an undocumented population that is significantly larger than it should be according to estimates by the Office of Immigration Statistics (OIS), MPI, CMS,

and others. That is, the logical edits we apply are not enough to capture all the LPRs and so we need to sort the remainder of the population into LPR and undocumented categories. To assign the rest, we first determine the probability of being undocumented using a technique similar to that in Capps, et al. (2013). Following the very clear directions kindly provided by those authors, we started with Wave 2 of the 2008 Survey of Income and Program Participation (SIPP) in which respondents offered answers with regard to whether they had LPR status upon arrival and whether they had ever achieved it later; those who answered no to both were considered to be undocumented.

In our estimation of the probabilities, we reduce that sample of immigrants in two ways. To understand why, it is important to realize the purpose: to take the estimates of the impact of certain variables on the probability of being undocumented and apply those to the ACS microdata. But recall that the sample to which we apply the estimates is a sample created *after* logical edits that exclude all pre-1982 immigrants and all those who are likely on student visas. Thus, we first drop from the SIPP sample the same potential individuals, narrowing the sample down to those who arrived after 1980 (because that is the break in the SIPP coding) and dropping from that all foreign-born residents who arrived in the last five years who are currently enrolled in undergraduate university or graduate school.

Capps, et al. (2013) use a similar approach to determining who in the SIPP is undocumented. They then essentially add these individuals to the American Community Survey and use a multiple imputation strategy to populate "missing" answers in the ACS microdata (which are basically all the answers). We instead utilize a logistic regression strategy in which the probability of being undocumented is determined by an equation in which the right-hand side variables include gender, age, years since arrival, education levels, marital status (whether never married and if married, whether married to a U.S.-born or naturalized citizen), whether or not the respondent has his or her own children in the house, English ability, and several dummy variables for broad region of origin. This specification is based in large part on the discussion in Van Hook, Bachmeier, Coffman, and Harel (2015).

We then applied the coefficients from that regression (utilizing sample weights on the observations to better approximate estimated populations) to the observations in the pooled ACS microdata. With all the observations tagged by conditions and probabilities, we were ready to utilize what some have termed "country controls" (Warren, 2014)—which essentially help to adjust the number of undocumented by country of origin to fit the country totals most observers believe to be the case.

The Role of Country Controls

In what might be the sort of "gold standard" for country-of-origin estimates, the approach used by Warren (2014) of CMS involves developing an independent estimate of these country totals; we instead make use of those and other estimates. For example, we take advantage of the fact that the OIS offers a breakdown of the top 10 nations of origin of the undocumented for 2012 (Rytina, 2013) and downweight by the decline in the aggregate number since that year. We then line up estimates from MPI and CMS and take the average from all three with some nuance: we drop the MPI estimate for Guatemalans because it is so much higher than the others and we also take a lower average for Dominicans based on work that suggests that the share of undocumented is surprisingly low in the Dominican community (Marcelli et al., 2009).

For the remaining countries, we take a variety of approaches. For example, to estimate the Brazilian undocumented, we utilize two-year average from 2009 and 2010 (their official numbers had fallen and so the count was not in the most recent OIS reports on the top ten); other studies have shown that

unauthorized Brazilians are a very large share of the non-citizen Brazilian immigrant population and we did not want to miss this nuance (Marcelli et al., 2009). We also base our estimate for Canadians on an estimate of undocumented Canadians that was generated by MPI in 2008. Aside from these special cases, we line up available country-of-origin estimates from MPI and CMS. For the smaller countries, the degree of divergence is small and so we take a simple average or just use one of the targets if that is all that is available.

For the rest of the unauthorized population, the easiest approach would be to assume that all nations of origin have exactly the same share of undocumented residents by comparing the remaining OIS numbers to the non-citizen, non-Cuban immigrant numbers in the ACS microdata. However, that is clearly not the case and for these, we use available information on similar countries in their same hemispheres (either from the overall data or from the information in the SIPP data) to target a percent undocumented and hence number undocumented. At the end of the targeting and assignment process, we have a total number of adult undocumented residents that is close to the MPI and CMS totals.

More precisely, it is close to that number assuming a degree of undercount. There is a widely-shared assumption that the undocumented are undercounted by around 10 percent in the decennial Census (see Marcelli & Ong, 2002) and more in other samples. To account for this, we had initially set the targets below the target adult numbers (nation-by-nation) so that when we reweighted all of those observations up, we would arrive at the anticipated final number. Warren and Warren (2013) contend, reasonably enough, that the undercount might be as high as 20 percent in recent years because the ACS is perceived as a more voluntary survey by respondents than is the Census. To implement this procedure, however, we stick closer to the earlier research and set the undercount assumption for adults at 12.5 percent.

Assigning Legal Status to the Rest of the Pool

Using the logical edits, we then assign the remaining adults to either documented or undocumented status until we reach the country controls discussed earlier. To do this, one logical approach might be to take all non-citizen, non-Cuban adult residents who had not been assigned to documentation by the conditions and sort them in order of the probability of being undocumented, using a random number assigned to all respondents to break ties where a large group of respondents share the same probability. One issue with this strategy, however, is that sorting and assigning based on probability estimates tend to generate an undocumented population that is younger and more male than other samples (because these groups tend to have the highest probabilities).

To account for this, we adopt a more complicated approach that takes into account the probability of being undocumented but in a way that is similar to the multiple imputation strategy used by other researchers (Bachmeier, Van Hook, & Bean, 2014; Batalova, Hooker, & Capps, 2014). To understand the strategy, note that each individual who has not yet been assigned to LPR status through logical edits has a particular probability of being undocumented. We round these to the second decimal and wind up with just over 60 possible strata (that is, individuals who share the same probability of being undocumented). We then select a sample for each country as follows (with several complications introduced in a minute). We take all those falling in the .60 stratum and randomly select 60% of those cases; we then go to the next stratum, say, .50, and select 50% of those cases randomly; and so on.

Essentially, what we are trying to do is mimic the underlying probability distribution of the undocumented. In the case where all those probabilities are exact, such a procedure yields a profile for

each country of the undocumented with various probabilities (some high, some low, but all based on the actual probabilities in the sample and with an average that is the country's average probability). Since we are not likely to have such perfect estimates, we instead stratify the sample, taking increasing slices of the country population distribution. To understand this, suppose we take a first slice at half the probability (so we chose 30 percent of those in the .60 stratum, 25 percent of those in the .50 stratum, etc.). This gives us a probability distribution that parallels the country in question but is likely below the country total. We can then move to the next slice, pulling another 20 percent (so an additional 12 percent of those in the .60 stratum, an additional 10 percent of those in the .50 stratum, and so on) until we bump up against the country control.

In fact, we start the first slice at 20 percent (so we choose, for example, 12 percent of those in .60 stratum, etc.). We repeat for twenty slices, within each slice ordering individuals from high probability to low probability and then selecting individuals from each stratum sequentially until we meet the country total. However, one runs out of individuals in the highest probability strata more quickly than in the lower probability strata—so after the first few slices, each slice generally has a top observation that has a lower probability. For each country, we also take the five percent of the observations with the lowest probability (and thus, most random chance of being undocumented) and assign them to the last slice. Thus, the minimum in each slice (until the last slice) may not be the lowest in the sample for that country. The breaks are set such that we never pull anyone from that last (least likely slice).

The most important point is that the process described above corrects for the bias of sorting by high probability and more or less simulates a multiple imputation procedure. It is no surprise, then, that our numbers are relatively close to those of MPI. However, the numbers are different for states and for other compositional elements offered by CMS because of the approaches CMS takes with regard to assigning documentation status—a topic we discuss later when we consider the eligible-to-naturalize population.

As a final step in our procedure, with individual adults tagged as undocumented, we turn to youth, assigning minor children as undocumented if one of their two parents is undocumented and neither parent is a U.S. citizen. After adding that number to the adult count, we make some very minor adjustments to the weights to better fit benchmarks on state totals. We ultimately come up with a total of 11,030,000 undocumented immigrants, a bit more than the 11,020,000 estimated by the MPI and the 11,010,000 estimated by CMS, and less than a 2012 estimate of 11,200,000 from Pew.

Estimating the Eligible-to-Naturalize

With all this in place, we then move to calculate the eligible-to-naturalize. This is actually far more straightforward: Those foreign-born non-citizen adults who are not considered to be undocumented are deemed eligible to naturalize if they meet certain conditions. The basic one is being in the U.S. for more than five years (or three years if married to a U.S. citizen), but we also exclude those who seem to have student visas (using an approach similar to the exclusion for the undocumented described above) or who are otherwise eligible but lived abroad or just got married to a U.S. citizen last year (the three-year condition requires three years of marriage).

Like CMS, we account for the fact that the ACS is an ongoing sample (i.e., the survey is conducted every month) and so the last half year of observations needs to be censored in the calculations (Warren & Kerwin, 2015). The reason is that someone who answers in January and reports that they arrived five years before could have arrived in December of that year and so would only have been in-country for a

bit over four years; since we don't know when they answered or arrived in that year, we simply randomize and choose half from those on the "edge" year.

We also add "derivative minors"—children who will automatically become citizens if their parents make the shift. These are foreign-born non-citizens under 18 living with a parent who is eligible to naturalize providing that the child is not undocumented (figured earlier) and that there is not another parent who is already a U.S. citizen (in which case—and this is a small number of cases—the child is already eligible and the non-citizen status might be a misreport). The number we derive for total eligible to naturalize is 8,803,000, very close to the 8,790,000 reported by the Office of Immigration Statistics (Rytina & Baker, 2014).

Comparison with CMS Estimates

Another major effort to offer estimates of those eligible to naturalize at the local level is that undertaken by the Center for Migration Studies, or CMS (Warren & Kerwin, 2015). While there are many similarities in the approaches, there are three key differences for the purposes of this memo:

- CMS developed independent country controls while CSII is essentially utilizing country estimates taken from work by CMS, OIS, and others; this is a major contribution of the CMS effort and builds on earlier pioneering work by Warren and Warren (2013).

- In determining the undocumented, CMS applies more logical edits and so winds up with fewer non-citizen foreign-born individuals to allocate to documented or undocumented status. Partly because of this, CMS assigns the remainder randomly. CSII applies fewer conditions and assigns the remainder non-randomly, taking advantage of an estimate of the probability of being undocumented.

- CMS calculations are presented for individual single years of the ACS while CSII calculations are done with and presented for a five-year pool.

What are the consequences of these differences? First, CMS country controls are likely superior but we are generally close for the bigger sending countries—and there is nothing innovative in the CSII approach compared to the CMS approach on the country control side of the equation. Given this broad similarity, many of the minor differences in the answers yielded by the two approaches will occur as a result of the second two factors: the conditions and assignment methods and the pooling issue.

Conditions and Assignment

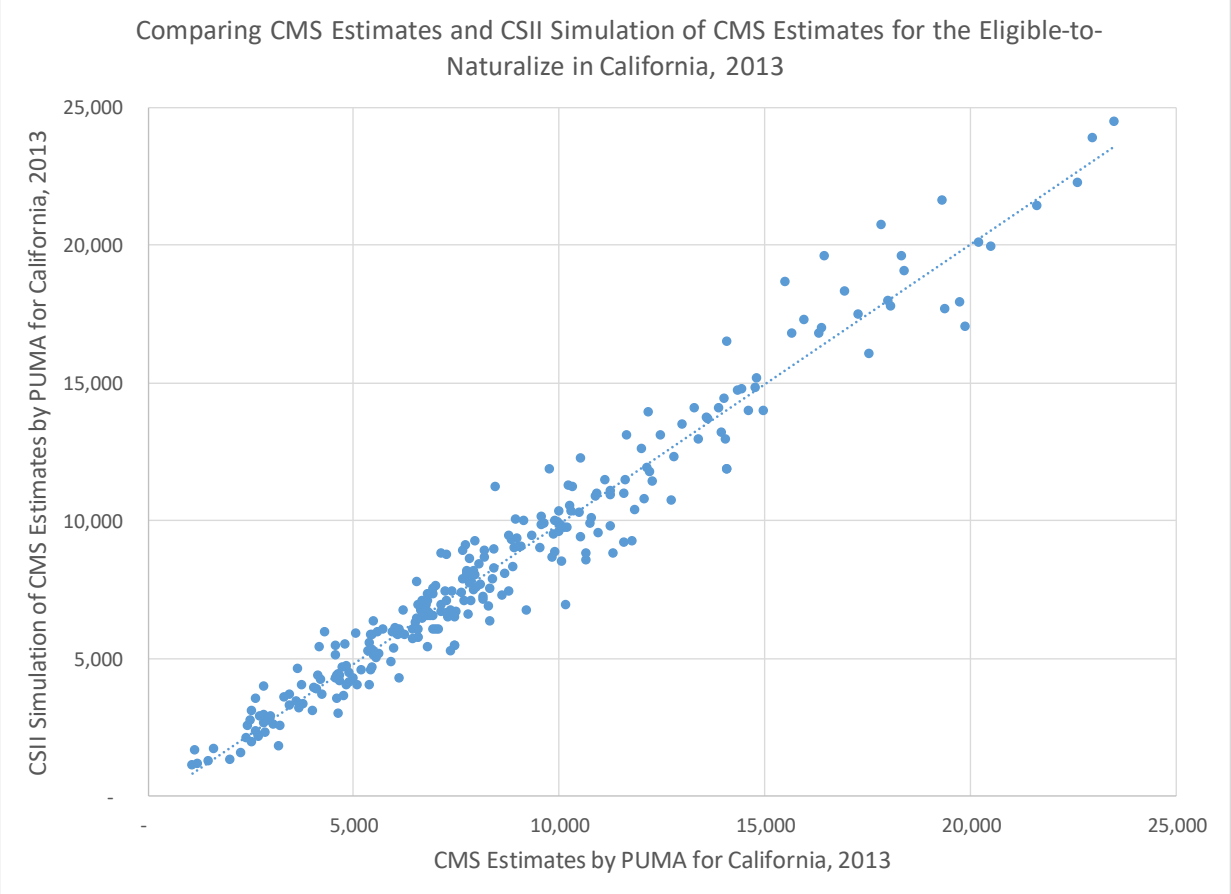
As noted, CMS applies more conditions to edit than CSII. The edit we were most concerned about was the choice to designate nearly all non-citizen adults who receive MediCaid as documented. In our view, that may be too strict, particularly in California where undocumented immigrants can access parts of the public health system and so may respond positively to a question about MediCaid when, in fact, they are undocumented. With fewer non-citizen adults left after logical edits, CMS assigns the remainder random numbers and proceeds to add up until a country control is hit.

To better understand the consequences of the two approaches, we sought to replicate the CMS data in the five-year pool by applying all the conditions CMS utilizes and then randomly assigning individuals subject to country controls. This results in totals very similar to CMS in terms of the number

undocumented and we also come up with, 8,609,000 individuals eligible to naturalize, very close to the 8,616,000 reported by CMS (but, of course, slightly lower than the number derived through the CSII method). We also replicate the one-year 2013 sample and come up with a total of 8,545,000, less than a one percent difference from the CMS number.

This simulation approach allows us to make certain comparisons that suggest that differences in answers we derive are due to the methods of assigning the undocumented (as well as the pooling, a topic we turn to below). To see this, we first compare CMS estimates and CSII simulations of the CMS estimates for 2013 for PUMAs in California. The pattern we show is also true nationwide but we confine our attention to the state with the most eligible to naturalize in order to report fewer but still quite a few observations in order to clarify the pattern. The first thing to note in the resulting scatterplot is that while there are differences, the fitted line is a nearly 45-degree line to the origin, suggesting that the aggregate fit between our simulation and the actual CMS numbers is marked by a good PUMA-level match as well (Figure 1).

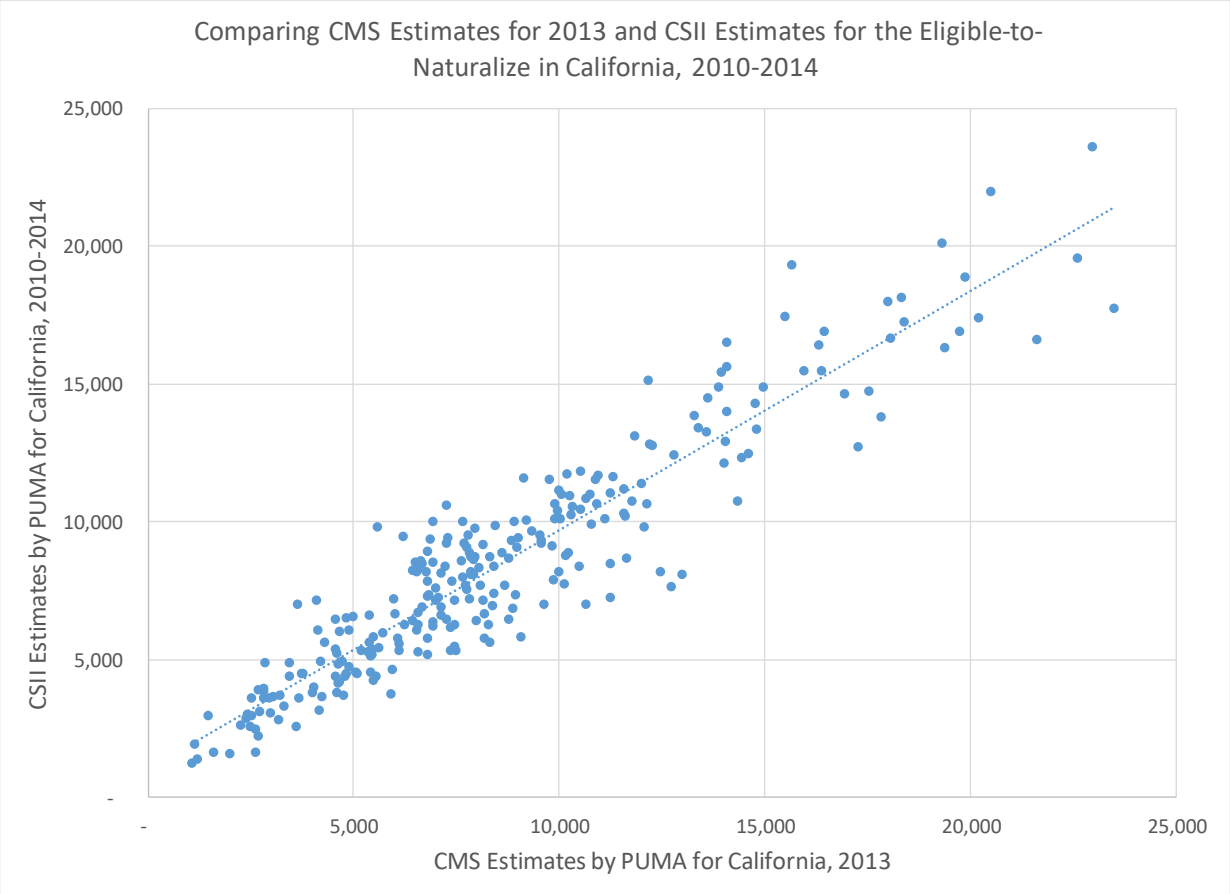
Figure 1



But what if we compare the CMS estimates with the actual CSII estimates (rather than our simulation of the CMS estimates)? As Figure 2 shows, this is a noisier pattern and one that has a fitted line that differs

from 45 degrees, suggesting that these are close but different estimates. Let's explore the differences and investigate the consequences.

Figure 2



Turning Back to the Undocumented Estimates

To understand the difference between the CSII results and the CMS results, it is useful to consider the precursor to determining who is eligible: how individuals are assigned as undocumented. For CMS, this is mostly through logical edits. For example, Warren reports that for Mexicans, about 20 percent of the non-citizen immigrants are tagged as documented; of the remainder, the country controls suggest that more than 95 percent are undocumented (Warren, 2014, p. 322). When we simulate the CMS approach, we come close to that: just under 19 percent are assigned by conditions and nearly 90 percent of the remainder tuck in under the country controls for undocumented. However, in the preferred CSII approach, with fewer conditions and the use of probability estimates to assign, only about 15 percent of Mexican non-citizen adults are assigned by logical edits and less than 80 percent of the remainder are assigned as undocumented by probabilities (as described above).

With such a high share of individuals being assigned undocumented status, we would expect modest differences; whether it's 95 percent or just under 80 percent being assigned status, we would not expect significant differences in aggregate outcomes for Mexicans. However, for other groups where the share being assigned undocumented status either randomly or based on probability after the logical edits is

smaller, we would expect bigger differences. For example, our simulation of the CMS approach suggests that about 50 percent of Asian-origin immigrants must be assigned status by CMS after the logical edits—and here randomization could yield differences.

Is there any advantage to our assignment procedure in which we use fewer conditions and make use of a probability-based versus randomized approach? One way to look at the issue is to examine the probability of being undocumented (according to our regression analysis) for three categories (documented by conditions [or logical edits], assigned as documented by the process, and assigned as undocumented by the process) under the two methods (the CSII method which uses probabilities and the CMS which uses random assignment). The best comparison would be to the actual CMS micro-data; since we don't have that, our comparison is to our simulation of the CMS process which, as noted above, seems to be close in terms of broad patterns (and in terms of gender and other breakdowns reported by CMS, explored in our comparison, and not reported to save space here).

The results are reported in Table 1 and there is an interesting pattern.

Table 1

	CSII Process				CSII Simulation of CMS Process		
	Documented by conditions	Documented by Process	Undocumented by Process		Documented by conditions	Documented by Process	Undocumented by Process
Mexico / Central America	.261	.302	.463		.301	.429	.430
Carribean	.097	.121	.172		.111	.141	.146
South America	.361	.373	.476		.365	.433	.437
Europe	.094	.125	.168		.096	.141	.140
Asia	.200	.188	.237		.192	.220	.212
Africa	.150	.134	.182		.147	.158	.159
Total	.198	.225	.406		.219	.275	.374

First, note that the probabilities are similar for the conditional (or logical) assignments for CSII and CMS; the probabilities are generally higher for CMS but that makes sense given that there are more conditions utilized and hence a higher likelihood of tagging some undocumented individuals as documented. Second, note that there is virtually no difference in the probability estimate for the CMS approach between those assigned as documented and those assigned as undocumented post-conditions. That makes sense since the assignment at that stage is random.

Third, and perhaps most important, note that there is a significant difference between those two groups in the CSII approach—and perhaps more strikingly, that the probabilities for those assigned as documented are very close to the probabilities for those assigned by conditions by either approach (recall that, in each case, the condition assignment was done with no consideration of probabilities). This suggests that the CSII approach may actually effectively separate out the documented from the undocumented—and by doing that, create a different pool of those from which we estimate the eligible-to-naturalize, and perhaps one more representative of the LPR population.

The differences this makes are slight but could be important for studies of sub-populations, particularly those not from Mexico or Central America. For example, the eligible-to-naturalize population skews older in the CSII data than in the CMS data likely because the non-randomly assigned undocumented population skews younger.

Pooling Versus Single Year

Another key difference between the CMS and CSII approaches has to do with the use of a single year of data versus a five-year pool. There can be advantages to using a single year, particularly because pooling across multiple years is not straightforward and can introduce complications. The reason is that the PUMA shapes changed as of 2012—so the first two years (2010 and 2011) are based on the 5 percent PUMAs specified in 2000 Census (2000 PUMAs) and the last three years are based on the PUMAs specified in the 2010 Census (2010 PUMAs). To reallocate individuals in the 2010 and 2011 ACS microdata samples from the earlier PUMAs to the newer PUMAs, we used a 2010 population-based crosswalk from Mable/GeoCorr12, randomly assigning individuals from their 2000 PUMAs to 2010 PUMAs in proportion to the 2010 population distribution (Missouri Census Data Center, 2012). While this introduces some unknown degree of geographic error in our 2010 PUMA-level estimates (given that the undocumented population—or, indeed, any smaller sub-population—is not likely to be randomly distributed within PUMAs), we feel that the gains in reliability from the increased sample size warrants pooling five years of data. We note that any geographic inaccuracies only apply to two of the five years included in our pooled sample (2010 and 2011)—so about 40 percent of the overall sample. And for those two years, there is only any geographic inaccuracy for about half the PUMAs (since nearly half of the 2000 PUMAs are completely contained by a single 2010 PUMA). This suggests that the geographic issues affect only about 20 percent of the overall sample, and for much of that, the issues are not severe since a large share of the remaining 2000 PUMAs are mostly contained by a single 2010 PUMA (i.e., 86 percent are at least 50 percent contained by a single 2010 PUMA based on 2010 population).

The decision to use a one-year sample or a five-year pool is unlikely to generate much divergence at an aggregate or national level; whether you use a one-percent (one-year) or five-percent (five year) sample, the numbers will be very close. However, the median population in a PUMA in the 2010-2014 sample is 126,000 while the median non-citizen population is 6,200—and the question of whether that number is being generated from a five-percent or one-percent sample is important.

To see what difference pooling makes, we investigate two potential cut-offs for reporting: do not report if there are less than 50 total non-citizen foreign-born in the raw count or if there are less than 25 eligible-to-naturalize in the raw count. For the CSII calculations, we know the raw count for both—we simply turn off the weights and summarize the observations. As it turns out, 95 percent of our PUMA estimates for the 2,351 PUMAs in the U.S. meet the threshold of at least 50 raw observations of non-citizens; 92 percent of our PUMA estimates meet the threshold of 25 we utilize for those eligible to naturalize.

We then apply the same standards for the CMS single-year approach. We can directly determine the cut-off for non-citizens utilizing just the 2013 sample: utilizing a cut-off of 50, 47 percent of the PUMA-level one-year numbers for non-citizens meet the threshold. As for the raw count of the eligible-to-naturalize, we do not have the exact numbers so we guess in two ways: (1) We take the CMS numbers for 2013 and divide by the sample weights (averaging the weights for naturalized and non-naturalized foreign-born), and; (2) we can utilize the raw numbers of the CSII simulation of CMS for 2013 as a second check. Utilizing a cut-off of 25 raw count of the eligible-to-naturalize, either 955 or 950 of the PUMAs (40.6% or 40.4% of the total) meet the threshold.

Generating Estimates of the Eligible-to-Naturalize at Various Levels of Geography

With our individual-level estimates of who is likely to be eligible to naturalize in the ACS microdata place, we then sought to generate consistent summary estimates of this data at various levels of geography, (including PUMAs, counties, metropolitan areas, and states) which are featured in our interactive map, available at: http://dornsife.usc.edu/csii/eligible-to-naturalize-map/.

For these estimates, we focused on the adult (age 18 or older) population for a variety of reasons: they are the only ones who are technically eligible to naturalize since minors (under 18) may only derive citizenship through their parents; they are the focus of most of the research on the socio-economic characteristics of those eligible-to-naturalize in comparison to the naturalized population (again, since they make the naturalization decision rather than their minor children); and they are the ones who would be eligible to vote, if naturalized. In addition to summarizing the total number of eligible-to-naturalize adults, we also summarized the numbers from seven broad regions of origin (to provide a rough sense of their language and cultural distribution), as well as the total number of adults and the total citizen voting-age population (CVAP). The latter two variables were necessary to calculate the eligible-to-naturalize as a share of all adults (to provide a sense of concentration) and the maximum percentage by which they could increase the voting-eligible population if they all naturalized (to provide a sense of potential impact on the electorate). In total, 10 variables were estimated for each geography.

In short, our procedure relies upon initial summary estimates at the PUMA level from the ACS microdata, distributes them across the Census tracts contained in each PUMA using tract-level information from the 2014 5-year ACS summary file, and then uses the resulting tract-level estimates as the "building block" to summarize the data to the aforementioned higher levels of geography, which are featured in our interactive map. While not reported in our interactive map, the tract-level estimates are a convenient bi-product of our procedure; while they are certainly not reliable enough to report for the entire U.S., they may be useful for examining the sub-PUMA distribution of eligible-to-naturalize adults in large cities and metropolitan areas with sizeable non-citizen foreign-born populations.

Now, for the longer version of that brief synopsis. Recall from above that the first part of our estimation procedure resulted in consistent 2010 PUMA tags applied to all the microdata, so the first step here of summarizing all 10 variables to the (2010) PUMA level was straightforward. We then prepared a crosswalk between 2010 Census tracts and 2010 PUMAs, again using underlying information from Mable/GeoCorr12 (Missouri Census Data Center, 2012). For the most part, Census tracts do not cross PUMA boundaries, but when they do we simply assigned the tract to the PUMA containing the largest share of its 2010 population. Next, we allocated each of the 10 PUMA variables to the tract level separately, using tract-level "proxy" variables from the 2014 5-year ACS (which covers the same period as our pooled ACS microdata). The proxy variables used to allocate each PUMA-level variable to the Census tracts contained in each PUMA are shown in Table 2, and were selected on the basis of being the best available variables in the ACS summary file to approximate each of the 10 PUMA-level measures.

For total adults and total CVAP, the tract-level proxy variables were identical to the PUMA-level measure. Due to lack of information in the ACS summary file on non-citizens from Africa, we had to estimate this proxy variable. To do so, we relied upon tract-level information in the ACS summary file on total foreign-born by country/region origin, which does break out Africa. We calculated the African share of all foreign-born after subtracting out those from Mexico, Central America, South American, the Caribbean, Asia, and Europe, and multiplied it by total non-citizens in each Census tract after first subtracting out the same regions to estimate the number of non-citizens from Africa.

Table 2

PUMA-level variable	Tract-level proxy variable
Total adults (age 18 or older)	Total adults (age 18 or older)
Total Citizen Voting Age Population (CVAP)	Total Citizen Voting Age Population (CVAP)
Total eligible-to-naturalize adults	Total non-citizen adults
Total eligible-to-naturalize adults by region of origin:	Total non-citizens by region of origin:
Mexico	Mexico
Central America	Central America
South American & Caribbean	South American & Caribbean
Asia	Asia
Africa	Africa (estimated)
Europe	Europe
Other	Other

In the initial allocation, we simply distributed the PUMA-level measures across the Census tracts contained in each PUMA in proportion to the tract-level proxy variable. We then took the tract-level estimates of total adults and total CVAP as our final estimates. We also took our initial tract-level estimate of total eligible-to-naturalize adults as final. An alternative could have been to set the final tract-level total to the sum of the initial counts by region of origin, but this would likely be far less accurate given that the counts by region of origin are based on much smaller samples (both at the PUMA and tract levels) and the proxy variables used to derive them are not as closely aligned as the proxy used to derive our initial estimate of the total.

Finally, in order to ensure that our tract-level estimates of the number of eligible-to-naturalize adults by region of origin summed (across regions of origin) to our final tract-level total and also summed (across tracts in each PUMA) to our PUMA-level totals, we utilized an Iterative Proportional Fitting (IPF) procedure. The resulting fitted estimates were then used as the "building block" to summarize the data to the higher levels of geography which are featured in our interactive map, including (2010) PUMAs, counties, metropolitan (metro) areas (using the U.S. Office of Management and Budget's February 2013 definitions), and states. As a check on our final estimates, we compared our results for 2010 PUMAs and states to what we get when we summarize to these levels directly from the ACS microdata to ensure they were in alignment.

We would stress that all estimates we provide are subject to at least two sources of error: sampling error since they are based on the ACS microdata, and error in our assignment of undocumented and documented status to non-citizen foreign-born respondents. In addition, the estimates we provide at the PUMA, county, and, to a far lesser degree, the metro area are also subject to a small amount of geographic error stemming from the fact that we estimate the 2010 PUMA for respondents in the 2010 and 2011 ACS microdata samples who reside in 2000 PUMAs that are not coterminous or completely contained in a single 2010 PUMA. To guard against reporting highly unreliable estimates, we do not report any PUMA- or state-level estimates if they are based upon fewer than 50 actual (unweighted) non-citizen adults in the ACS microdata, and we do not report county- or metro area-level estimates if they are based upon fewer than 1,000 (weighted) non-citizen adults—which is the equivalent of 50 unweighted observations assuming an average weight of 20 (which is the case in our pooled ACS microdata sample which covers approximately five percent of the total U.S. population). The reason for different cutoffs is due to the fact that we are able to calculate the unweighted counts for PUMAs and

states (and feel they are a better choice for censoring data), but are unable to for counties and metro areas given our estimation strategy. The censoring described above results in missing data for 149 out of 2,351 PUMAs, 2,146 out of 3,143 counties, and 21 out of 381 metro areas; all 51 states met the 1,000 non-citizen threshold.

Conclusion

This memo describes the steps utilized to generate our estimates of the eligible-to-naturalize and compares our method to that of CMS. The bulk of the work occurs in the estimation of the undocumented which then allows us to generate a pool of presumed LPRs, and from that, estimates of the eligible-to-naturalize population. Our process includes the use of country controls, conditional edits, and probability sorting for the first part of this procedure; both the use of a probability-based approach and a five-year pool should lead to reasonably accurate estimates and breakdowns at the PUMA level, which may make the data useful for targeted local efforts. Lastly, this memo documents our methodology for generating estimates of eligible-to-naturalize adults at various levels of geography (PUMAs, counties, metro areas, and states) that are featured in the new CSII interactive map, available at: http://dornsife.usc.edu/csii/eligible-to-naturalize-map/.

Acknowledgements

CSII would like to thank the Carnegie Corporation of New York, the James Irvine Foundation, The California Endowment, the California Community Foundation, and the California Wellness Foundation for providing the funding that has built the capacity to carry out this research. We would also like to thank CSII staff Rhonda Ortiz for keeping us on track to get these estimates out, Gladys Malibiran for her help in developing our new interactive map, and Madeline Wander for her assistance in preparing this memo.

References

Bachmeier, J. D., Van Hook, J., & Bean, F. D. (2014). Can We Measure Immigrants' Legal Status? Lessons from Two U.S. Surveys. *International Migration Review*, *48*(2), 538–566. http://doi.org/10.1111/imre.12059

Batalova, J., Hooker, S., & Capps, R. (2014). *DACA at the Two-Year Mark: A National and State Profile of Youth Eligible and Applying for Deferred Action*. Washington, D.C.: Migration Policy Institute. Retrieved from http://www.migrationpolicy.org/research/daca-two-year-mark-national-and-state-profile-youth-eligible-and-applying-deferred-action

Capps, R., Bachmeier, J. D., Fix, M., & Van Hook, J. (2013). *A Demographic, Socioeconomic, and Health Coverage Profile of Unauthorized Immigrants in the United States* (MPI Issue Brief No. 5). Washington, D.C.: Migration Policy Institute. Retrieved from http://www.migrationpolicy.org/research/demographic-socioeconomic-and-health-coverage-profile-unauthorized-immigrants-united-states

Marcelli, E. A., Holmes, L., Estella, D., da Roucha, F., Granberry, P., & Buxton, O. (2009). *(In)Visible (Im)Migrants: The Health and Socioeconomic Integration of Brazilians in Metropolitan Boston.* San Diego, CA: Center for Behavioral and Community Health Studies, San Diego State University. Retrieved from http://boston.com/bonzai-fba/Third_Party_PDF/2009/10/17/Marcelli_et_al_BACH_2009_Brazilian__1255753970_2565.pdf

Marcelli, E. A., & Ong, P. (2002). *Estimating the Sources of the 2000 Census Undercount among Foreign-born Mexicans in Los Angeles County*. Presented at the 2002 Population Association of America, Atlanta, GA.

Missouri Census Data Center. (2012). MABLE/Geocorr12: Geographic Correspondence Engine (Version 1.2). University of Missouri: Missouri Census Data Center.

Ruggles, S. J., Alexander, T., Genadek, K., Goeken, R., Schroeder, M. B., & Sobek, M. (2010). *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]*. Minneapolis: University of Minnesota.

Rytina, N. (2013). *Estimates of the Legal Permanent Resident Population in 2012* (Population Estimates) (p. 4). U.S. Department of Homeland Security; Office of Immigration Statistics.

Rytina, N., & Baker, B. (2014). *Estimates of the Lawful Permanent Resident Population in teh United Stes: January 2013* (Population Estimates) (p. 4). U.S. Department of Homeland Security; Office of Immigration Statistics.

Van Hook, J., Bachmeier, J. D., Coffman, D. L., & Harel, O. (2015). Can We Spin Straw Into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches. *Demography*, *52*(1), 329–354. http://doi.org/10.1007/s13524-014-0358-x

Warren, R. (2014). Democratizing Data about Unauthorized Residents in the United States: Estimates and Public-Use Data, 2010 to 2013. *Journal on Migration and Human Security*, *2*(4), 305–328. http://doi.org/10.14240/jmhs.v2i4.38

Warren, R., & Kerwin, D. (2015). The US Eligible-to-Naturalize Population: Detailed Social and Economic Characteristics. *Journal on Migration and Human Security*, *3*(4), 306–329. http://doi.org/10.14240/jmhs.v3i4.54

Warren, R., & Warren, J. R. (2013). Unauthorized Immigration to the United States: Annual Estimates and Components of Change, by State, 1990 to 2010. *International Migration Review*. http://doi.org/10.1111/imre.12022