USCDornsife
Equity Research Institute

Estimating the Undocumented Population and the Eligible-to-Naturalize

By Manuel Pastor, Thai Le and Justin Scoggins
August 18, 2021

We estimate the undocumented population by applying a series of conditions and probability estimates to individual observations in the American Community Survey (ACS). We start with a focus on noncitizen, non-Cuban foreign-born adults. We exclude Cubans because until 2017, any Cuban arriving in the U.S. regardless of whether they held a visa, was automatically allowed to pursue legal residency a year after arrival. We then apply a series of edits to our pool of noncitizens, using data on characteristics that are indicative of a legal or lawful resident (LR) rather than an undocumented resident. These characteristics include working in the public sector; having an occupation that requires documents (e.g., police officer); having received certain public benefits such as social security; or having received food stamps as a household head or spouse in a household without a child who could have been the lawful beneficiary of said public assistance (Pastor and Scoggins 2016).

Additionally, we consider observations to be an LR if they arrived to the U.S. before 1982 as these individuals are likely to have gained legal status through the Immigration Reform and Control Act of 1986. We also tag individual adults as documented if they report having received certain government-supported health services, including Medicare, Indian Health Services, or Veterans Affair Care (Warren 2014). We do not tag individual adults as necessarily documented if they indicate that they are receiving MediCaid. This is partly because it is possible to be undocumented and receive such assistance in states like California. However, the primary reason is that the actual survey question asks if one has received "Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability." Many undocumented immigrants are able to access local public clinics, particularly in California, and so answering "yes" to this question does not definitively mean they are documented or a lawful resident.

We also net out two groups that are likely to be in the U.S. legally though not as Lawful Permanent Residents (LPRs). The first group consists of individual who immigrated as adults and are currently enrolled in higher education; these are likely to be student visa holders and we tag them as such. The second group consists of those who are likely to be H1B visa holders. While this could be a hard task, the work is facilitated by the concentration of such visas: roughly 75 percent of H1B visa holders are migrants from India and another 12 percent are from China; about two-thirds of H1B are in high-tech occupations, with smaller but significant percentages in architecture, engineering, post-secondary education, and medical professions.

Given this information, We thus create a tag for those who likely have H1B visa status if they are working-age noncitizens from India or China with at least a bachelor's degree and in certain computer and technology occupations, or in one of the other specified occupations. They also must have been in the country for no more than ten years, which could be sufficient time to earn a degree and do two H1B visa stays. Through these conditional tags and screens, we are able to come up with estimates by occupations that approximate the Indian and Chinese H1B visa holders in those occupations in the U.S. For the remainder of the Indian and Chinese, we take a random sample of those who fit all the other conditions but are not in the specified occupation until we hit the expected country totals for India and China. For the smaller number of H1B visa from other countries, we select randomly from other (non-Indian and non-Chinese) non-citizens that meet the various educational, occupational, length of stay,

age, and other criteria until we fill out the remainder; the distribution by country of origin for the larger countries in this group lines up nearly perfectly with the actual figures. Identifying those who are most likely H1B visa holders improves the accuracy of our undocumented estimates as well as the accuracy of determining which legal residents are LPRs and so eligible to naturalize.

Through this logical editing process (that is, the application of screens), it can be assumed that observations exhibiting these characteristics are documented, but it cannot be assumed that observations not exhibiting these characteristics are undocumented. We thus rely on probability edits to impute who in the remainder of our dataset is likely to be undocumented. Similar to Capps et al.'s (2013) approach using the 2008 SIPP, we develop a statistical model that determines the probability of being undocumented based on data from the Survey of Income and Program Participation (SIPP). We apply the coefficients from the SIPP analysis to calculate the probability that a noncitizen in the ACS microdata is undocumented based on whether they exhibit similar characteristics.

In our previous work estimating the undocumented population (e.g., Pastor and Scoggins 2016), we followed Bachmeier et al. (2014), Batalova et al. (2014), and others and used the 2008 SIPP (wave 2) to determine which characteristics are associated with being undocumented. We are able to directly identify those who are likely undocumented in the SIPP because those who were noncitizens at the time of the survey and who arrived without permanent resident status were asked directly if they eventually achieved LPR status. Those indicating they did not we assumed to be undocumented. However, in the 2014 SIPP, respondents arriving as noncitizens answered a series of sociodemographic and socioeconomic questions, including their LPR status upon arrival, but they were not asked about subsequent adjustments. Since we could not assume that all those who arrived with non-permanent status were undocumented—as they may have adjusted their status since arriving to the U.S.—we needed to estimate whether adjustment had occurred through a logical editing process.

To do this, we first worked with the previous 2008 SIPP wave 2 data to examine characteristics associated with switching status and then applied a series of logical edits. These included whether one had arrived with non-permanent status but had subsequently served (or was serving) in the military; was in public employment or held a job that likely required documentation; was married to a U.S. citizen for ten years or more; had arrived before the IRCA cut-off and was thus likely to have adjusted status; or was not a current student, had at least eight years in the country, and held a Ph.D. Finally, we tagged those who received Medicare or were receiving Medicaid but had not recently given birth (since there is an exception for undocumented mothers) as immigrants who likely adjusted their status. We would like to note that we are comfortable using the MediCaid screen here because the SIPP question is specifically about health insurance coverage and is not worded as vaguely as in the ACS.

Though this method of ascertaining status adjustment through conditions involves a bit more guesswork, it allows us to add those likely to have adjusted status to the group we are certain arrived with permanent status, and we estimate the undocumented as the remainder. The advantage of the 2014 SIPP is that the data is more consistent with more recent ACS microdata. This is particularly true in terms of time in the U.S. For example, when we compare estimates applying the coefficients from the 2008 SIPP with the 2014 SIPP to the very same ACS microdata, we find that the length of time in the U.S. increased roughly at par with what one would expect in an undocumented population that peaked in 2007 and has been on a slow decline ever since—signaling an undocumented population that is "aging in place."

With the 2014 SIPP sample now consisting of those who arrived with permanent status, those who adjusted, and those who did not, we use a logistic model where the outcome is undocumented status (the last of the three groups) and the predictors are a series of sociodemographic and economic variables including gender, age, years since arrival, education levels, marital status, citizenship status of spouse, English speaking ability, and dummy variables for broad region of origin. This model predicts the impact of each variable on the probability of the respondent being undocumented. The resulting coefficients are then applied to the remaining observations in the ACS microdata (after the logical edits described above) to determine their probability of being undocumented, in a method akin to the approach of Van Hook et al. (2015).

With probabilities assigned, we then take the adult population that are not tagged as lawful residents by the conditions or screens discussed above, and use the probability model to estimate their status. We do this with benchmark goals in mind for undocumented adults by country of origin and state of residence. This "country control" procedure relies on different reputable estimates of the undocumented population, including those from the Center for Migration Studies (CMS) and the Migration Policy Institute (MPI), which we adjust to reflect only adults based on the age distribution of all noncitizens by country of origin in the ACS. We update our country and state controls each time we generate estimates based on a new release of ACS microdata depending on updates by others. The CMS country numbers are likely more reliable given the "residual" method used to calculate them so when there is much divergence, we lean in their direction.

At this point, one straightforward approach would be to tag as undocumented adults those with the highest probability of being undocumented until we reach a country total. However, that is likely to miss those who might, for example, be educated but were visa-over-stayers or who are otherwise somewhat atypical. We thus created probability strata (i.e., groups of individuals who share the same probability of being undocumented, rounded to the nearest whole percentage point). We pull from each strata in multiple iterations until each country control (i.e., threshold) is met. This can be repeated up to twenty times for each of the sixty strata to obtain a sample that mimics the probability distribution that we derive from the 2014 SIPP sample; generally, it does not take all twenty runs since we hit the thresholds earlier.

To understand this part of the process, assume that there are only three strata with probability ratios of .6, .3, and .1. In the first round, we sample from each stratum, taking .6 of those in the first stratum, .3 of those in the second stratum, and .1 of those in the third stratum. We then go through subsequent rounds until we hit each country's estimated threshold. Each subsequent iteration tags more of each strata's remaining observation as undocumented, repeating this until each country control is met. For each country, twenty percent of the observations with the lowest probability of being undocumented are assigned to the last iteration; we do this so that observations with the lowest probability of being undocumented are only tagged if the country controls are not met before we hit the last iteration.

We then adjust our estimates to fit state totals by creating estimated ranges for origin countries and states of residence, with the range being tighter for larger origin groups and state locations. We always shift the most likely to be undocumented in a country-state group to undocumented status, and shift those least likely to be undocumented in a country-state group to documented status. When we are done, we check and find, as one would expect, that those we tag as undocumented have a higher average probability that than those we tag as documented. More intriguing (and more assuring) is that those who are assigned as documented by conditions and those who are assigned as documented

through our model have nearly identical probability estimates. In short, the two groups (one whose status we are certain about through conditions and the others whose status we estimated through probabilities) are virtually the same in terms of this key characteristic.

We then match all the adult observations with any of their own children living in the same household. If a child has just one parent and that parent is undocumented, we tag the child as undocumented. We also tag the child as undocumented if both parents are undocumented. We then tag children as undocumented if their parents are not present in the household and more than half of the adult family members in the household are undocumented or two-thirds of the adult household members are. Since we also have a target number of undocumented children based on the aforementioned age structure of the noncitizen population, we fill in the rest with a share of the children who have one undocumented and one LPR parent present.

The resulting estimates are weighted up in a process similar to Warren (2014). However, to make this work, we implicitly apply the weights to the country targets as we apply them. In other words, we use the country controls and calculated number of adults as target totals. We apply an undercount estimate to reach these target totals using sample weight so that the properly weighted (or undercount-adjusted) number at the end of the process comes close to the country control. If the weight is uniform, this is straightforward; if the weights vary with time in country (and we assume they do), the target number will vary accordingly so that reapplication of the weights will meet the expected number.

We also estimate who in the undocumented population holds either DACA or TPS. For each of these, we apply the conditions needed for each status. In the case of DACA, we use current age, age at arrival, and educational attainment and enrollment status to get at the eligible population. We then use country and state numbers for actual recipients and randomly assign those we tagged as DACA eligible until we hit the target numbers published by the U.S. Office of Citizenship and Immigration Services (USCIS). We do the same to identify those who are likely TPS holders by using TPS-designated countries and time periods of arrival in which TPS could be granted. We randomly sort in those who are likely eligible until we hit targets based on official TPS registrant numbers published by the Congressional Research Service.

Once we identify those who are likely undocumented, we can determine which observations are eligible to naturalize with greater accuracy. Immigrants in the U.S. are eligible to naturalize when they 1) are at least 18 years old; 2) have resided in the U.S. as a LPR for at least five years or three if married to a U.S. citizen; 3) have been physically present in the U.S. for at least 30 months; 4) are deemed a person of good moral character by USCIS officers; 5.) have the ability to write, speak, and read in English; 6) are able to prove a basic understanding of U.S. civics and history; 7.) can demonstrate an attachment to the Constitution and its principles; and 8) are able to take the Oath of Allegiance (U.S. Citizenship and Immigration Services 2019).

Of these conditions, we can determine if an immigrant satisfies the age and LPR requirements using the ACS microdata. Though we know an immigrant's reported English speaking ability, we do not use this to determine eligibility because it is not a commonly used criterion by the Office of Immigration Statistics in estimating the eligible-to-naturalize population (Baker 2020) and because a significant share of immigrants who have naturalized still report limited English proficiency. Starting with our full ACS microdata, we first narrow our sample down to those who are foreign born and not yet citizens. Among noncitizens, we net out those who we estimated to be undocumented as they lack the required lawful permanent residency status to naturalize. Among our remaining observations, we initially identify an

immigrant to be eligible to naturalize if they are at least 18 years old and meet the five- or three-year residency requirement.

To refine these initial estimates, we then remove those we had previously tagged as student visa holders and H1B visa holders, those who lived abroad during the year prior to the survey, and deemed to be eligible to naturalize through marriage to a U.S. citizen but married during the year prior to the survey. Additionally, even if they do not meet the time thresholds, we include veterans and people on active military duty as eligible to naturalize, along with noncitizen children if they have at least one parent living in the same household who is eligible.

References

Bachmeier, James D., Jennifer Van Hook, and Frank D. Bean. 2014. "Can We Measure Immigrants' Legal Status? Lessons from Two U.S. Surveys." *International Migration Review* 48(2):538–66. doi: 10.1111/imre.12059.

Baker, Bryan. 2020. *Estimates of the Lawful Permanent Resident Population in the United States and the Subpopulation Eligible to Naturalize: 2015-2019*. Office of Immigration Statistics at the U.S. Department of Homeland Security.

Batalova, Jeanne, Sarah Hooker, and Randy Capps. 2014. *DACA at the Two-Year Mark: A National and State Profile of Youth Eligible and Applying for Deferred Action*. Migration Policy Institute.

Capps, Randy, James D. Bachmeier, Michael Fix, and Jennifer Van Hook. 2013. *A Demographic, Socioeconomic, and Health Coverage Profile of Unauthorized Immigrants in the United States*. 5. Washington D.C.: Migration Policy Institute.

Pastor, Manuel, and Justin Scoggins. 2016. *Estimating the Eligible-to-Naturalize Population*. Los Angeles: Center for the Study of Immigrant Integration, University of Southern California.

U.S. Citizenship and Immigration Services. 2019. *Naturalization Fact Sheet*. U.S. Citizenship and Immigration Services.

Van Hook, Jennifer, James D. Bachmeier, Donna L. Coffman, and Ofer Harel. 2015. "Can We Spin Straw Into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches." *Demography* 52(1):329–54. doi: 10.1007/s13524-014-0358-x.

Warren, Robert. 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):305–28. doi: 10.1177/233150241400200403.