New Approaches to Estimating Immigrant Documentation Status in Survey Data

Heeju Sohn and Anne R. Pebley

PWP-CCPR-2020-003

June 26, 2020

California Center for Population Research
On-Line Working Paper Series

New Approaches to Estimating Immigrant Documentation Status in Survey Data

Heeju Sohn
California Center for Population Research
University of California Los Angeles
hesohn@ucla.edu
Corresponding author

and

Anne R. Pebley
Department of Community Health Sciences
Department of Sociology
University of California, Los Angeles
pebley@ucla.edu

ABSTRACT

Approximately a quarter of the 43 million immigrants living in the United States are thought to be undocumented. Yet, the lack of accurate population-level information about undocumented immigrants provides fertile ground for public misconceptions, political and media hype, and false claims. The goal is to determine how well descriptions of the undocumented population are likely to mirror the reality of undocumented immigrants' lives in the US. We compare (1) the distribution of the population by documentation status and (2) distributions of the characteristics of undocumented and documented immigrants produced by two methods. The first method (the "decomposition method") is a commonly used strategy used in previous work and the second method is an alternative, independent method developed in this article. We used the Survey of Income and Program Participation (SIPP) and the Los Angeles Family and Neighborhood Survey (LAFANS). The existing decomposition method works reasonably well if the data contains information on whether respondents are naturalized citizens or and lawful permanent residents. However, when these variables are missing or problematic, the decomposition method produces biased results. The actual undocumented population in the US may be even more socioeconomically disadvantaged than studies based on existing decomposition methods indicate. This article evaluates methods to conduct reasonably accurate nationally representative, policy relevant research on the lives of undocumented immigrants without potentially jeopardizing members of this vulnerable population.

KEY WORDS Undocumented immigrants, Demographic methods, United States

In 2015, there were an estimated 43 million foreign-born residents in the United States (U.S. Citizenship and Immigration Services 2017), of whom approximately one quarter was thought to be undocumented (Passel and Cohn 2017; Rosenblum and Ruiz Soto 2015). Since the 1990s and especially in the last few years, undocumented immigrants have been the subject of intense, and often vitriolic, political debate. An anti-immigrant environment has led to increased US government deportation and detention efforts, militarization of the US-Mexico border, and even proposals to end birthright citizenship and restrict lawful permanent residence.

Although there is a rich literature based on qualitative research and small-scale surveys detailing the lives of undocumented immigrants (see for example, Menjívar and Abrego 2012; Arbona et al. 2010; Gonzales 2011; Gleeson 2010; Flippen 2012), the results of these studies, by their nature, cannot be generalized to the undocumented population nationwide. Considerably less is known about the undocumented population as a whole. This lack of information, coupled with the fact that many US natives have little personal contact with undocumented immigrants, provides fertile ground for public misconceptions, political and media hype, and false claims (Flores and Schachter 2018). For example, Flores and Schachter (2018) show that non-Hispanic whites are more likely to believe that immigrants are undocumented if they have criminal backgrounds, are poorly educated, work in the informal sector, and are from Mexico, Central America or Syria. They are also less likely to believe that Asian immigrants are undocumented. Yet undocumented immigrants have lower than average crime rates, an estimated 15 percent have a bachelor's or higher degree, many work in formal sector jobs, and 15-20 percent are Asian or white (Rosenblum and Ruiz Soto 2015; Light and Miller 2018;

Warren and Kerwin 2017; Gelatt and Zong 2018). Greater availability of credible information on the reality of undocumented immigrants' lives could contribute to a more evidence-based public discussion of immigration issues. On the other hand, collecting data on documentation status is itself a politically sensitive issue and, without adequate care and human subjects protections, may place immigrants at risk (Carter-Pokras and Zambrana 2006; Prentice, Pebley, and Sastry 2006; Mervis 2018). Thus, a key question is how we can conduct reasonably accurate, nationally-representative, policy relevant research on the lives of undocumented immigrants that will allow informed policy discussions without potentially jeopardizing members of this vulnerable population.

Researchers have developed several approaches to attribute documentation status to respondents in large national surveys in order to use the survey data to describe the socioeconomic and health characteristics of documented and undocumented immigrants. In this paper, our primary goal is to evaluate the quality of the estimates produced by a commonly used strategy – which we will call the "decomposition method" – for attributing documentation status in major national surveys. We do this by comparing estimates from a version of the decomposition method with an alternative, independent method developed in this paper. For each of the two methods we estimate and compare: (1) the distribution of the population by documentation status and (2) distributions of the characteristics of undocumented and documented immigrants. The ultimate goal is to determine how well descriptions of the undocumented population are likely to mirror the reality of undocumented immigrants' lives in the US. The alternative method that we use is a prediction equation model based on a sample survey that directly asked detailed questions about documentation status. Because this

method uses an entirely different strategy to estimate undocumented status, it can serve as a comparison for estimates produced by the decomposition method. We begin with a brief overview of methods for estimating the size and characteristics of the undocumented immigrant population in the US. More detailed reviews can be found in Van Hook et al. (2015) and Bachmeier et al. (2014).

Background

We use the term undocumented to denote that an individual has no legal authorization (i.e., citizenship, LPR, a valid visa, refugee/asylum status, temporary protected status, etc.) to be in the US, permanently or temporarily. Research on undocumented immigrant populations is difficult in all immigrant-receiving countries (Vogel, Kovacheva, and Prescott 2011): how can you study a population which is, by nature, hidden and wants to remain that way? In the US, a long line of research has focused on estimating the size and geographic origin and distribution of the undocumented population. Early examples include Passel and Woodrow (1987) and Warren and Passel (1987). These estimates generally relied on residual methods using the US Census or Current Population Survey (CPS) combined with administrative data on legal immigration and refugees from the Immigration and Naturalization Service (INS) or, more recently, the Department of Homeland Security (DHS). A recent example of this type of research is the detailed annual estimates of the components of change in the size of the undocumented population by state produced by Warren and Warren (2013).

A second and more recent line of research examines the socioeconomic characteristics of the undocumented population using large sample surveys and other data. This is the issue

we focus on in this paper. We illustrate the data limitations faced by studies in this literature with the schematic diagrams in Figure 1. The most commonly used method, i.e., the decomposition method, was developed by Passel and colleagues' and builds on the residual methods described above, using the Current Population Survey (CPS) or the American Community Survey (ACS) which are large surveys, representative of the US population, collecting data monthly or annually. CPS and ACS collect data only on citizenship and, thus, leave the documentation status of the majority of the foreign-born sample unknown, as represented in Figure 1a[1]. Passel et al. use these data and a series of assumptions about demographic, geographic, and socioeconomic characteristics to allocate the unknown category shown in Figure 1a – foreign-born non-citizens – to the categories "definitely lawful" and "potentially unauthorized." The most recent set of assumptions is outlined by Passel and Cohn (2018, pages 38 to 41) and includes attributes like working in occupations for which having a visa is highly likely (e.g., diplomat, high-tech worker). Then, cases identified as "potentially unauthorized" are probabilistically assigned to documented or undocumented status to make them consistent with the total numbers of people in each category estimated by the residual method described above. Passel et al.'s results are widely used and accepted as reasonably accurate.

Another approach to describing the undocumented population's characteristics is to use data from the Survey of Income and Program Participation (SIPP). Unlike other nationally representative surveys, SIPP asks foreign-born respondents about documentation status

[1] For verisimilitude, the size of each slice in Figure 1 is determined by Passel et al.'s estimate of the proportion of foreign-born US residents who are citizens, LPRs, visa holders, and undocumented immigrants. See Passel and Cohn. 2018. Key findings about U.S. immigrants. http://www.pewresearch.org/fact-tank/2018/09/14/key-findings-about-u-s-immigrants/ph_16-06-02_foreign-bornbreakdown-2/

directly. For example, the topical module in the second wave of SIPP 2004 asks foreign-born respondents about: naturalization, immigrant status at the time of arrival (lawful permanent residence, refugee or asylum status, non-immigrant status, i.e., held a visa, and other), and, for non-permanent residents at arrival, whether or not they had acquired permanent residence since. Unfortunately, for these purposes, SIPP only releases data on: (1) naturalization, (2) lawful permanent residence (LPR) at entry, and (3) whether R acquired LPR status since. Data on other legal statuses at entry including visas, Temporary Protected Status (TPS), and refugee status are not available to SIPP data users[2]. As shown in Figure 1b, availability of data on LPR status should considerably reduce the proportion of the sample with unknown documentation status. Even so, the size of the remaining "unknown" category is still large. The majority of respondents in this category are likely to be undocumented, but it also includes other groups such as H1B visa holders in high tech industries, university students, and refugees and asylees. Thus, the actual undocumented population in the US may be even more disadvantaged than estimates based on allocation of this unknown group suggest. A number of studies have used SIPP data to compare respondents who are assigned documented and undocumented status (Hall and Greenman 2015; Bachmeier, Van Hook, and Bean 2014; Van Hook et al. 2015). Although the availability of SIPP data on LPR status shrinks the size of the "unknown" category, these studies still need to allocate the sizeable non-citizen, non-LPR group by documentation status. To do so, they use a version of the decomposition method, described above, to categorize the non-citizen, non-LPR proportion of the sample into documented and

[2] Note that even if the more detailed SIPP data were available, they would provide information only on *status at the time of entry to the US* and not current status at interview. Since the frequency of visa overstaying is significant, a substantial number of people entering with a visa would be likely to no longer have a valid visa at the time of interview.

undocumented categories. Appendix A presents an illustration of one version of this decomposition method used by Hall et al. (2010).

Unfortunately, a sizeable proportion of SIPP respondents asked about LPR status at entry and subsequent adjustment to LPR status did not respond. In Figure 1c[3], we illustrate the problem that this non-response creates. Thus, there are two potentially quite different groups with unknown documentation status: the first due to non-response to LPR questions and the second due to responding "no" to both questions (i.e., no LPR at entry and no adjustment to LPR subsequently). Thus, Figure 1c is a more appropriate representation of actual data availability for documentation status in the publicly accessible SIPP data. The Census Bureau assigns values to missing data using statistical procedures (e.g., "hot decking"), but Bachmeier et al. (2014) persuasively argue these methods may not be appropriate for documentation status. Studies using SIPP to attribute documentation status must determine how to treat the non-response. Some use the Census Bureau allocations; others assign values for LPR status to these cases using either multiple imputation or Passel et al.–type decomposition methods. Also of concern, given the sensitivity for respondents of admitting to being undocumented on a government-sponsored survey, is whether all respondents replying positively to the LPR questions are being truthful – an issue discussed in detail by Bachmeier et al. (2014).

Another development is the use of SIPP data to impute LPR status in other, larger surveys. Although SIPP's sample is large, the American Community Survey (ACS) sample is far larger and, therefore useful for generating state or regional-level estimates and conducting multivariate

[3] Like the size of the other slices of Figure 1, the slices shown for non-response and unknown in Figure 1c are schematic and do not represent precise estimates of the actual percent in each group.

and/or multilevel analyses of complex associations. Van Hook and colleagues (Van Hook et al. 2015; Capps et al. 2013) use a cross-survey multiple imputation (CSMI) method to impute the probability that foreign-born non-citizen ACS respondents have LPR status. In CSMI, a donor data set (e.g., SIPP) is used to impute variables in a target data set (e.g., ACS). The donor and target survey samples must be drawn from the same universe and pairs of variables of interest must be jointly observed in one data set or the other (for details see, Van Hook et al. 2015; Rendall et al. 2013). Using the SIPP data and CSMI, it is possible to impute LPR status for ACS respondents, although some assumption must be made about the non-respondents on LPR status in SIPP[4]. Since SIPP does not provide any information on the "unknown" category in Figure 1c, Van Hook et al. (2015) use standard decomposition methods to assign documentation status (Van Hook 2019).

In summary, versions of the decomposition method are commonly used to estimate the characteristics of the undocumented immigrant population using large nationally representative sample surveys. The availability of data on LPR status in SIPP – despite the high rate of missing data – has the potential to improve these estimates, both in the SIPP and in other nationally representative surveys. Nonetheless, for the foreign-born sample who are not citizens and non-LPR, the decomposition method is, of necessity, used by almost all contemporary analyses to distinguish undocumented and documented immigrants.

Since researchers rely on the decomposition method for information on the undocumented population, it is important to assess the quality of the estimates it produces. Comparisons of the estimated distribution of documented immigrants from surveys with

[4] The details of the Van Hook et al. (2015) procedures for non-response in SIPP are not clearly spelled out in their article.

aggregate data from the Department of Homeland Security shows that the totals in each category are roughly comparable (Passel and Cohn 2018), but these comparisons are limited to data on legal immigration. Van Hook et al. (2015) assess the reliability of decomposition methods using simulated data in which they assign and vary the actual distribution by documentation status of respondents as well as their health insurance coverage status. Then they use the decomposition method to classify respondents by documentation status and health insurance coverage and compare it to the known distribution in the simulated data. Their results show that the decomposition method (i.e., their "demographic accounting method") results are biased both in cases when associations between health insurance and documentation status are in the expected direction and even more so when they are not.

In this paper, we use a different approach to assessing the strengths and limitations of the decomposition method: a prediction equation model based on a regional survey – the Los Angeles Family and Neighborhood Survey (L.A. FANS) – that collected detailed documentation status from a large stratified probability sample of respondents. Heer and Passel (1987) appear to have been the first to use a prediction equation model to estimate documentation status, followed by a number of others (Marcelli and Heer 1997; Capps et al. 2013). The method involves using a data set in which current documentation status is known for all respondents – including foreign born non-citizen, non-LPR respondents – to develop a prediction equation by regressing documentation status on a set of respondents' socioeconomic, demographic, and other characteristics. The estimated coefficients are then used to impute documentation status for sample members in a second data set based on their characteristics. If the two data samples are drawn from the same universe (i.e., the same population) then cross-sample

multiple imputation is a more appropriate approach because it is likely to yield less biased results (Little and Rubin n.d.; Van Hook et al. 2015). However, given CSMI's requirement that a donor and a recipient sample are drawn from the same population (Rendall et al. 2013), the dearth of national level data which meet that requirement, and the fact that even SIPP-based estimates (with or without CSMI to another data set) require use of the decomposition method for non-citizen, non-LPR, foreign born respondents, a prediction equation can be a useful alternative approach. Van Hook et al. (2015) show that both multiple imputation and prediction equation approaches (i.e., "single imputation") produce virtually unbiased estimates of the "true" associations in Monte Carlo-simulated data – when used in the same sample or, by extension, samples drawn from the same universe. In this paper we consider the use of a prediction equation approach using two samples drawn from different universes as described below.

To assess decomposition methods, we compare the prediction equation estimates with a decomposition approach used by Hall, Greenman, and Farkas (HGF) (Hall et al. 2010) using SIPP data. We chose HGF's version of the decomposition method because it is well described and straightforward. In HGF's method, foreign born respondents who are neither citizens nor lawful permanent residents (LPRs) are assigned to documentation status categories based on factors such as whether they are currently post-secondary level students, have received public social welfare benefits, are employed in particular occupations, are in the US military or veterans, or are likely to have refugee status based on origin country and year of arrival[5]. This method is similar to Passel et al.'s decomposition method, although it does not attempt to

[5] The HGF article does not make clear what they do about missing values on the SIPP LPR status questions. We assume they use the "hot deck" values assigned by the U.S. Census Bureau.

assign documentation status to respondents with unknown status using probabilistic methods to match the total distribution by status in the population. In the first part of our analysis, we use HGF's method to classify L.A.FANS respondents by documentation status and compare these estimates with L.A.FANS' respondents' reports of their own "actual" status, to determine how well the HGF method works with L.A.FANS. Second, we use L.A.FANS data to develop an equation that predicts documentation status among L.A.FANS respondents based on socioeconomic and demographic characteristics. We test this equation method in several ways. Then, we apply this prediction equation to SIPP data and compare our estimates of the proportion undocumented with HGF's estimates. Finally, we use SIPP to compare the estimated characteristics of undocumented immigrants from the prediction equation and the HGF method.

Our analysis makes an important contribution to research on the undocumented immigrant population because we evaluate methods that are the basis for most estimates of undocumented immigrants' characteristics and we compare them to an alternative method based on prediction equations. SIPP and other datasets combined with decomposition approaches to estimate the status of foreign-born, non-citizen (and in the case of SIPP, non-LPR) respondents are likely to provide the primary source of individual-level data on this population for the foreseeable future, particularly since human subjects concerns are likely to increase with growing threats of apprehension, detention, and deportation. Thus, understanding the strengths and weaknesses of the decomposition approach will improve the quality of research – and hopefully policy-making and public discussion –about for this vulnerable population.

Data

We use data from two sources: the Los Angeles Family and Neighborhood Survey (L.A. FANS) and the 2004 Survey of Income and Program Participation (SIPP). L.A.FANS collected detailed data on current documentation status for all respondents in a probability sample that is representative of the population of Los Angeles County in 2000-2001. We also searched for contemporary population-representative data sets from other areas of the US to include in the analysis as a contrast to Los Angeles but were unsuccessful.

A key assumption in using prediction equation models is that the associations among documentation status, the predictor variables, and socioeconomic and demographic variables of interest are approximately the same size and direction in the universes from which the "donor" (i.e., L.A.FANS) and the "recipient" data (i.e., SIPP) were drawn. Although no single subnational location can represent the experience of the undocumented population in the US, the undocumented population of Los Angeles County may be a better approximation than those of many other US locations, particularly in the early 2000s, before new immigrant destinations outside of the southwest US were less common than they are today. In the early 2000s, the undocumented population was heavily concentrated in urban areas. The Los Angeles metro area was estimated to have had almost twice the undocumented population as any other metro area – approximately 10% of all undocumented immigrants were estimated to live in Los Angeles (Fortuny, Capps, and Passel 2007; Migration Policy Institute 2018). Furthermore, as a major gateway for undocumented immigration to the US, many undocumented immigrants living elsewhere in the US had passed through or lived in Los Angeles sometime after arrival. Although in recent years, Los Angeles has been seen as friendlier to undocumented immigrants,

in the early 2000s, Los Angeles City and County law enforcement officials cooperated with the US Department of Homeland Security (DHS/ICE) in several programs targeting the undocumented population, including Secure Communities and the 287(g) program (Menjívar et al. 2018). Nonetheless, because the Los Angeles immigrant population is likely to be different from that of the US as a whole, we conduct several tests (described below) to determine whether the prediction equation we develop is likely to yield reasonable estimates.

In the first part of the analysis, we use a decomposition approach to classify L.A.FANS respondents by probable documentation status. Then, we compare this estimated documentation status to direct answers from respondents. Although we use the same two data sources as Bachmeier et al. (2014), our goals are quite different. Bachmeier et al. (2014) assessed the feasibility of collecting documentation status in other surveys and then used SIPP data to compare results from different imputation procedures on basic estimated sample characteristics. By contrast, our goal is to determine whether the decomposition method used with SIPP, specifically the variant used by Hall et al. (2010), actually do a good job of predicting documentation status in a sample where this status is known (at least to the extent that self-reports are reliable). In the second part of the analysis, we consider an alternative method to estimate undocumented status using SIPP with data from L.A.FANS or other surveys that collect data on documentation status. Then we compare the estimates of socioeconomic characteristics of the undocumented population produced by the Hall et al. (2010) method and our prediction equation method.

Los Angeles Family and Neighborhood Study (L.A.FANS)

We use data from wave 1 of the Los Angeles Family and Neighborhood Survey (L.A.FANS), a stratified, multistage, clustered random-sample survey of 3,100 households conducted between April 2000 and December 2001 in 65 census tracts in Los Angeles, California (Sastry et al. 2006). The county includes 88 separate cities and many unincorporated areas, spread over 4,083 square miles. The L.A.FANS sampling strategy was designed to select a representative sample of the population of Los Angeles County. At the time, Los Angeles County's population of about 9.5 million was 45% Latino, 31% white, 13% Asian-Pacific Islander, and 10% African American, but because of the oversample of poor neighborhoods, 57% of the L.A.FANS-1 sample (unweighted) was Latino (Peterson et al. 2004). Los Angeles was and is a major destination for immigrants. In 2000, about 30% of the Los Angeles County population was foreign-born, but the unweighted percentage is about 57% for the L.A.FANS-1 sample.
In sampled households, L.A.FANS-1 interviewed a total of 3,500 adults age 18 and older. Households in poorer neighborhoods and those with children were oversampled. Interviews were conducted in person in English and Spanish. Sample weights are used to correct for oversamples. L.A.FANS used the "peel the onion" method of determining citizenship and documentation status. All adult respondents were asked in the following order whether they were (a) native-born citizens, (b) naturalized citizens, (c) lawful permanent residents, (d) asylees, refugees, or had temporary protected immigrant status (TPS), or (e) had a tourist visa, student visa, work visa or permit, or another time-limited document. Those in the category (e) were asked whether their visa or other document was still valid or had expired. Respondents in who said they held one of these forms of authorized residence were skipped out of the rest of the sequence. The residual of respondents who did not report citizenship, LPR status, or a valid

(non-expired) visa or documented status were assumed to be undocumented. There is a reason to believe that the responses to these questions were fairly reliable. Spanish-speaking respondents in L.A.FANS were interviewed by Latino native Spanish-speaking interviewers (Peterson et al. 2004) and interviewers reported little respondent unease in replying to these questions. To protect respondents, L.A.FANS obtained a Certificate of Confidentiality from the National Institutes of Health and included that information in the interview consent materials. Furthermore, the response rates for these questions are high (Bachmeier, Van Hook, and Bean 2014). We use responses to this series of questions as the benchmark against which to assess documentation status estimated by the decomposition method.

Survey of Income and Program Participation (SIPP)

Although the L.A.FANS publicly-available data contains far more detailed information documentation status, SIPP is better suited to study immigrants in the US as a whole because of its nationally representative sample, large sample size, and the regular availability of data over a long period of time. SIPP is a longitudinal survey of a representative sample of the US population, conducted by the U.S. Bureau of the Census. Panels of respondents are followed for roughly four years and regularly interviewed (in "waves") during that time. All household members at sampled addresses become panel members and those 15 and older are interviewed in each wave. Interviews are primarily by telephone. We use the 2004 SIPP panel which collected detailed monthly information on employment, income, occupation, and other topics for over 45,000 households from October 2003 until December 2007. The SIPP survey is divided into core questionnaires and topical modules. In the 2004 topical model in wave 2, SIPP asks about citizenship status, naturalization, LPR and visa status at arrival, and whether the

respondent has adjusted his/her status to LPR since arrival. As noted above, only data on citizenship, naturalization, LPR status at arrival and LPR status at interview are publicly released. Data on program participation, income, employment, and other topics are collected in each wave. SIPP immigrant status questions include items on place of birth, citizenship, and LPR status.

We limit both our L.A.FANS and SIPP analysis sample to foreign-born respondents aged between 18 and 60. The LA.FANS sample comprises 1,871 foreign-born individuals and is weighted to represent the residents of Los Angeles County. The SIPP sample comprises 7,507 foreign-born individuals and is weighted to compensate for SIPP's initial selection probabilities and differential attrition between waves 1 and 2 across subpopulations.

Analytical Strategy

In the first part of the analysis, we evaluate the decomposition method employed by Hall, Greenman, and Farkas (Hall et al. 2010) ("HGF method") by comparing the estimate of documentation status with more complete self-reported status in the L.A.FANS. In the second part, we examine an alternative to the HGF estimates in SIPP, using a prediction equation from L.A.FANS. The third part of the paper compares estimates of undocumented immigrants' socioeconomic characteristics based on the HGF method with those from our prediction equation method.

Evaluation of Hall, Greenman, and Farkas (HGF)'s allocation method using the L.A.FANS

The methodology developed by Hall, Greenman, and Farkas (2010) and its subsequent variations (Bachmeier, Van Hook, and Bean 2014; Borjas 2017; Van Hook et al. 2015) identifies

undocumented persons in the SIPP by classifying as documented individuals from the pool of foreign-born respondents who are naturalized citizens, lawful permanent residents (LPRs), recipients of federal assistance, post-secondary school students, "high-ranking public officials", or married to a post-secondary student or "high ranking public officials" (Bachmeier, Van Hook, and Bean 2014). We apply this HGF method to L.A.FANS assuming that all we know is nativity, citizenship, and LPR status, as is the case in SIPP. We then compare the results of the HGF procedure with respondent reports of their own documentation status.

As described above, relying on the LPR status in the SIPP is problematic because this variable has a high non-response rate (25 percent). Non-response is also likely to be greater among respondents who are undocumented (Bachmeier, Van Hook, and Bean 2014). It is also true that HGF-type methods have been used with other survey data that do not collect LPR status data (e.g., Borjas 2017) and are likely to be used for this purpose in the future. For these reasons, it is useful to know how accurate the HGF method is if LPR status is unknown or inaccurate. Therefore, we examine a second scenario in which we apply the HGF method to L.A.FANS data assuming that we know only nativity and citizenship, and not LPR status. Then we compare the results of this application of the HGF method with respondents' reports of their documentation status.

HGF's method was specifically developed to study a relatively homogenous population: low-skilled Mexican immigrant workers. We selected HGF's approach as an exemplar of decomposition methods due to its clarity and reproducibility but recognize that some of the differences between the HGF results and ours are likely to be due to their focus exclusively on low-skilled Mexican immigrants. However, similar decomposition methods have been applied

to study the documentation statuses of broader immigrant populations living in the US (Borjas 2017; Bachmeier, Van Hook, and Bean 2014; Van Hook et al. 2015). Appendix A shows how HGF's allocation procedure works.

An Alternative Approach: L.A.FANS Prediction Equation. In the second part of the paper, we develop and test a prediction equation as an independent means of estimating the SES characteristics of the undocumented population in the SIPP. By comparing SES characteristics generated by the L.A.FANS prediction equation with those from the HGF method in the same sample, we can compare the strengths and weaknesses of the approaches and get a sense of the likely range of SES characteristics.

We estimate the odds ratios of being undocumented associated with the set of covariates included in both L.A.FANS and SIPP: year of immigration, educational attainment, marital status, number of children in the household, continuous health insurance coverage for past two years, place of birth, age, and sex (Appendix B). We selected the combination of predictors that would yield the lowest mean squared-difference between the self-reported documentation status (0 = documented and 1 = undocumented) and the predicted probability of being undocumented (0 to 1). This combination of covariates yielded the greatest accuracy in predicting the characteristics of the undocumented population in the L.A. FANS, as measured by mean-squared difference. In addition to the variables in this final model, we also considered models that included: receipt of federal assistance, spouse's occupation/student status, and an alternative coding of marital status, education, number of children, race/ethnicity, and place of birth. However, adding these variables reduced the fit, and the variables were eliminated from the model. The size and direction of the coefficients are consistent with prior descriptions of

the undocumented population in the US (Passel and Cohn 2009). We used the equation which contains all the variables listed to predict documentation status.

If SIPP respondents all or nearly all responded to the questions on LPR status (as L.A.FANS respondents did), a logical method of proceeding might be to construct a prediction model only for the portion of the population that reported being non-citizens and non-LPRs (and thus assuming accurate reporting of citizenship and LPR status). However, the high level of non-response for these questions in SIPP makes this difficult because there is no way of knowing which L.A.FANS respondents would or would not have answered the LPR questions in SIPP had they been asked. Thus, for this comparison, we decided to use only information on whether or not the respondent is a citizen. None of the models include any explicit immigration status variables (i.e., naturalized, LPR, or temporary visas) other than citizenship. To test the prediction equation estimates, we estimate the model on randomly-selected ninety percent of the L.A.FANS sample (training set) and then use the coefficients to predict the probability that a person is undocumented in the remaining ten percent (validation set). These predicted results are then compared to self-reports of documentation status. We repeated this procedure ten times so that all observations rotate through the ten-percent validation set. Appendix C describes this process, known as the n-fold method, in detail.

As a test of this procedure, we also predicted naturalization status—an immigration-related variable that has an almost 100% response rate in the SIPP – using a prediction equation estimated from the L.A.FANS. We compared actual self-reported naturalization status in SIPP with naturalization status estimated by the L.A.FANS prediction equation and found that 50 percent of naturalized immigrants had predicted naturalization probabilities of less than 0.04.

50 percent of naturalized immigrants had predicted naturalization probabilities greater than 0.65. The area under the ROC curve was 0.82. Appendix C describes the procedure and results. Unlike the HGF and other decomposition methods, the prediction equation yields a probability of being undocumented rather than a simple yes-or-no classification. These probabilities better represent the information available about individuals' documentation status because they reflect a level of uncertainty in classification. For some respondents, it is much clearer whether they are undocumented or not than for others.

After applying the prediction model, we present the estimated socioeconomic characteristics of the undocumented and documented population using SIPP, based on the prediction model, and compare them with the HGF results. For this comparison, we could dichotomize the predicted documentation status probabilities into documented or undocumented categories based on some arbitrarily chosen cut point. Instead, a better approach is to apply Bayes' Theorem (Bayes and Price 1763) to derive categorical characteristics such as occupation and weighted averages to calculate continuous characteristics such as income. By using predicted probabilities, this approach does not identify individual respondents in the SIPP as documented or undocumented but rather, derives the population-level summaries of the undocumented as a group. We present the profile of undocumented respondents derived from the L.A.FANS prediction method with the profile yielded by the HGF approach.

Results

Accuracy of HGF method in L.A.FANS Data

The first set of results compares the application of the HGF decomposition approach to L.A.FANS data and with the actual documentation status that respondents report in L.A.FANS. The goal is to determine how well we would do in predicting documentation status using the HGF method if respondents did not report it. For this exercise, we assume that only nativity, citizenship, and LPR status are known, as is true in SIPP. The results are presented in column 1 of Table 1. The first row shows that 27 percent of foreign-born L.A.FANS respondents in the sample report themselves to be undocumented. The HGF approach, in the second row, is largely consistent with respondents' self-reported documentation status (33 percent), in this scenario. The last two rows show the degree of misclassification. Two percent of foreign-born respondents are categorized by the HGF approach as documented but report themselves to be undocumented. About eight percent of foreign-born respondents who report themselves as documented are categorized as undocumented by the HGF approach. On further examination (not shown), we found that HGF's method overestimates undocumented status among people from Asia and the Pacific with postsecondary education, the group most likely to be on H1B[6] visas.

Table 1

The second column in Table 1 examines that case in which we assume only nativity and citizenship are known in L.A.FANS. When LPR status is unknown (as it is for about a quarter of eligible SIPP respondents and for all respondents in other surveys, like the American Community Survey), HGF's method performs poorly in identifying documented foreign-born

[6] It is worth recalling that HGF did not attempt to reclassify potential H1B visa holders in any way (as Passel and colleagues do in more complex versions of the decomposition method) because their interest was in Mexican and Central American immigrants with low educational attainment.

residents in L.A.FANS. This scenario produces an undocumented population that includes almost 55 percent of foreign-born residents in Los Angeles, in contrast to the 27 percent of foreign-born respondents who are undocumented by self-report.

In summary, the HGF approach does a good job for L.A.FANS respondents of predicting documentation status, but only if LPR status is known. Most of the errors produced by HGF are classifying documented residents as undocumented and much of this error may be due to misclassification of H1B and similar visa holders. However, if LPR is missing or unknown, the HGF method produces a high error rate which, not surprisingly, classifies a large portion of documented residents as undocumented[7].

Describing characteristics of the US undocumented population using SIPP

In the next part of the analysis, we predict documentation status in SIPP using the L.A.FANS-based prediction model and SIPP respondent attributes and estimate the socioeconomic characteristics of documented and undocumented SIPP respondents. The goal is to present an alternative picture of the SES distribution of the undocumented population.

Table 2

First, we present the relative odds of being undocumented associated with socio-demographic characteristics estimated from L.A.FANS in Table 2. As a point of reference, we also include the characteristics associated with being a naturalized citizen. Arriving in the US after 1980, not having continuous health insurance coverage, and having more children in the household were

[7] As a sensitivity test (not shown), we also tested the HGF method on the population that the method was originally developed for: immigrants from Mexico and Central America with low levels of schooling. Similar to our main findings, the HGF method performs well when LPR status is known and performs poorly when LPR status is unknown.

significantly associated with being an undocumented immigrant in the L.A.FANS. Compared to naturalized citizens, undocumented immigrants were more likely to men and divorce/separated.

Table 3 presents the sociodemographic and employment-characteristics of the undocumented population using the L.A.FANS-based prediction and the HGF method. The prediction model estimates that 23 percent (calculated by averaging the predicted probabilities) of foreign-born respondents in the SIPP are undocumented. In comparison, the HGF method assigns 18 percent of the foreign-born to undocumented status. Since we do not know the actual characteristics of the undocumented population in the US, there is no concrete basis on which to judge whether the HGF or L.A.FANS-based results are more accurate. Nonetheless, we can draw some conclusions.

Table 3

First, the characteristics of probable undocumented respondents are similar for both methods. For most of the variables, the differences between the samples produced by the two methods range between 0 and 30% of the L.A.FANS-based estimate. Reflecting the results evaluating the HGF method in Table 1, the largest differences are for the region of birth, race/ethnicity, and more highly skilled occupations. Second, many of the differences are due to higher proportions born in Asia who are more highly educated and in more highly skilled occupations. The HGF method is more likely to classify documented highly educated and higher income immigrants, who are often from Asia, as undocumented. It also produces significantly larger proportions (4 to 6.6 times larger than the L.A.FANS based predictions) of undocumented people working in computer, mathematical, architecture, engineering, and science occupations, and much larger shares other professional occupations in business and financial operations, healthcare, legal,

education, media, and sports (detailed occupation tabulations not shown). We speculated earlier that the reason is, at least in part, that in the HGF methodology, H1B visa holders would be classified as undocumented. Third, although the results of the analysis show that an L.A.FANS prediction equation does well in predicting naturalization status among SIPP respondents, the prediction equation for the undocumented population may, nonetheless, be biased. If the associations between documentation status and socioeconomic characteristics differ substantially between Los Angeles County and the US as a whole, we would expect to see a bias toward assigning higher probabilities of undocumented status to respondents who have the characteristics of the undocumented in Los Angeles. However, the extent to which these associations differ between Los Angeles and the US as a whole is unknown because of the lack of nationally representative data on documentation status. One case in which L.A.FANS may misclassify SIPP respondents is based on national origin. Undocumented status in Los Angeles is more likely to be associated with origins in Mexico and Central America than in the national population because of Los Angeles' predominantly Mexican and Central American-origin population, the proximity of the Mexican-US border, and the long history of circular migration between Mexico and the Southwest US. As Table 3 shows, the L.A.FANS prediction model yields far more individuals from Mexico and Central America than the HGF model. Some of this difference appears to be due to misclassification by HGF of well-educated Asian immigrants, but some may also be due to a bias toward allocating Mexicans and Central Americans to the undocumented status in the L.A.FANS-based model.

Discussion

Research on the undocumented population in the United States faces significant challenges, not the least of which is to try to understand the experience of this important population while protecting its members' privacy and confidentiality. In this paper, we evaluate a decomposition method for assigning documentation status to respondents in SIPP developed by Hall, Greenman, and Farkas – one version of a group of similar strategies employed by researchers to study the undocumented population. We also consider an alternative approach which uses a prediction model based on a sample in which respondents report their own documentation status.

Our results show that HGF's method does a fairly good job of assigning documentation status, as long as information on lawful permanent residence (LPR) is available for each respondent. However, even if LPR status data is available, the HGF method appears to misclassify well-educated, higher income respondents in high skilled occupations, many of whom are from Asia. As we have noted, Hall, Greenman, and Farkas' research focused on low skilled immigrants from Mexico and Central America and their research results, thus, were not affected by this apparent misclassification. Part of the misclassification of highly educated individuals may be driven by the probable bias of SIPP's hot-deck imputation of LPR status and duration of US residence for non-response, since we believe that Hall and his colleagues accepted the Census Bureau's allocation of respondents who were missing data on these questions. The most obvious source of potential misclassification is that in HGF's procedure, H1B visa holders are likely to be classified as undocumented. Researchers using the HGF method can correct for this bias by assigning some or all of the highly educated in high skilled jobs to the documented category, as some decomposition methods do (Passel and Cohn 2018).

Substantively, our results suggest that the actual undocumented population in the US may be even more disadvantaged than studies based on HGF-type procedures indicate.

In the absence of information on LPR status, however, the HGF method does a poor job in assigning documentation status: 30 % of documented respondents were classified as undocumented in L.A. FANS when we assume that no LPR information is available – which makes sense since the HGF method is designed to allocate individuals who are non-citizens and non-LPR. Thus, when LPR data is missing completely, as is true in the American Community Survey (ACS), the Current Population Survey (CPS), and most other major surveys, application of the L.A.FANS prediction equation method or a prediction equation from a similar survey may prudent to consider in addition to the decomposition method. If LPR status is available, but missing for a sizable proportion of the sample, as in SIPP 2004, multiple imputation prior to applying the HGF method would significantly improve HGF assignment. Since data on documentation status in a national sample is not available, we cannot determine which of these two methods would produce better results: (a) multiple imputation followed by the HGF method (with a correction for incorrect classification of highly skilled workers) or (b) use of the L.A.FANS-based procedure.

To develop a prediction equation from L.A.FANS which would work well in a national population, we followed several steps. First, we reduced the likelihood of producing a prediction model that is too specific to the L.A. FANS by estimating the prediction model from ten overlapping groups from the L.A.FANS and validated at each iteration onto a test dataset that did not contribute to the estimation. This step prevents a few individuals unique to the L.A.FANS from unduly influencing the prediction model. During this procedure, we determined

that adding certain variables (i.e., receipt of federal assistance and the spouse's student status) that have been widely used in the literature reduced the model's performance. Second, we repeated the prediction-validation procedure on the L.A.FANS 50 times to produce a distribution of possible prediction outcomes. Each outcome slightly varies due to the random process of dividing the L.A.FANS into the ten equal subsets. We use the expected values of these outcomes to predict immigration statuses of respondents in the SIPP.

We also tested the results of a similarly estimated L.A.FANS-based prediction equation for naturalization status on SIPP data in which naturalization status was reported. Results of this test showed that that the L.A.FANS-based method works well in SIPP's national sample. This validation test plus the process we undertook to develop the prediction models suggest that the prediction equation is also likely to work well for documentation status for SIPP and other surveys. However, the L.A.FANS-based procedure may be biased in unknown ways. We speculate that it may be somewhat more likely to assign Mexican and Central American immigrants to undocumented status than is actually correct.

References

Arbona, C, N Olvera, N Rodriguez, J Hagan, A Linares, and M Wiesner. 2010. "Acculturative Stress among Documented and Undocumented Latino Immigrants in the United States." *Hispanic Journal of Behavioral Sciences* 32 (3): 362–84.

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. https://doi.org/10.1111/imre.12059.

Bayes, Mr., and Mr. Price. 1763. "An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S." *Philosophical Transactions of the Royal Society of London* 53 (0): 370–418. https://doi.org/10.1098/rstl.1763.0053.

Borjas, George J. 2017. "The Labor Supply of Undocumented Immigrants." *Labour Economics* 46: 1–13. https://doi.org/10.1016/j.labeco.2017.02.004.

Capps, Randy, Michael Fix, Jennifer Van Hook, and James D. Bachmeier. 2013. "A Demographic, Socioeconomic, and Health Coverage Profile of Unauthorized Immigrants in the United States." https://www.migrationpolicy.org/research/demographic-socioeconomic-and-health-coverage-profile-unauthorized-immigrants-united-states.

Carter-Pokras, Olivia, and Ruth Enid Zambrana. 2006. "Collection of Legal Status Information: Caution!" *American Journal of Public Health* 96 (3): 399; author reply 399-400. https://doi.org/10.2105/AJPH.2005.078253.

Flippen, Chenoa A. 2012. "Laboring Underground: The Employment Patterns of Hispanic Immigrant Men in Durham, NC." *Social Problems* 59 (1): 21–42. https://doi.org/10.1525/sp.2012.59.1.21.

Flores, René D., and Ariela Schachter. 2018. "Who Are the 'Illegals'? The Social Construction of Illegality in the United States." *American Sociological Review* 83 (5): 839–68. https://doi.org/10.1177/0003122418794635.

Fortuny, Karina, Randolph Capps, and Jeffrey S. Passel. 2007. "The Characteristics of Unauthorized Immigrants in California, Los Angeles County, and the United States." Washington DC. https://www.urban.org/research/publication/characteristics-unauthorized-immigrants-california-los-angeles-county-and-united-states.

Gelatt, Julia, and Jie Zong. 2018. "Settling In: A Profile of the Unauthorized Immigrant Population in the United States | Migrationpolicy.Org." https://www.migrationpolicy.org/research/profile-unauthorized-immigrant-population-united-states.

Gleeson, Shannon. 2010. "Labor Rights for All? The Role of Undocumented Immigrant Status for Worker Claims Making." *Law & Social Inquiry* 35 (3): 561–602. https://doi.org/10.1111/j.1747-4469.2010.01196.x.

Gonzales, R G. 2011. "Learning to Be Illegal: Undocumented Youth and Shifting Legal Contexts in the Transition to Adulthood." *American Sociological Review* 76 (4): 602–19.

Hall, Matthew, E. Greenman, G. Farkas, and Amada Armenta. 2010. "Legal Status and Wage Disparities for Mexican Immigrants." *Social Forces* 89 (2): 491–513. https://doi.org/10.1353/sof.2010.0082.

Hall, Matthew, and Emily Greenman. 2015. "The Occupational Cost of Being Illegal in the United

States: Legal Status, Job Hazards, and Compensating Differentials." *Social Forces; a Scientific Medium of Social Study and Interpretation* 49 (2): 406–42. https://doi.org/10.1111/imre.12090.

Heer, D, and J Passel. 1987. "Comparison of Two Methods for Estimating the Number of Undocumented Mexican Adults in Los Angeles County." *The International Migration Review* 21 (4): 1446–73. http://www.ncbi.nlm.nih.gov/pubmed/12280920.

Hook, Jennifer Van. 2019. "Personal Correspondence."

Hook, Jennifer Van, 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. https://doi.org/10.1007/s13524-014-0358-x.

Light, Michael T., and T.Y. Miller. 2018. "Does Undocumented Immigration Increase Violetn Crime?*." *Criminology* 56 (2): 370–401. https://doi.org/10.1111/1745-9125.12175.

Little, Roderick J. A., and Donald B. Rubin. n.d. *Statistical Analysis with Missing Data*. Accessed October 4, 2019. https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data%2C+2nd+Edition-p-9780471183860.

Marcelli, Enrico A., and David M. Heer. 1997. "Unauthorized Mexican Workers in the 1990 Los Angeles County Labour Force." *International Migration* 35 (1): 59–83. https://doi.org/10.1111/1468-2435.00004.

Menjívar, Cecilia, and Leisy J. Abrego. 2012. "Legal Violence: Immigration Law and the Lives of Central American Immigrants." *American Journal of Sociology* 117 (5): 1380–1421. https://doi.org/10.1086/663575.

Menjívar, Cecilia, William Paul Simmons, Daniel Alvord, and Elizabeth Salerno Valdez. 2018. "IMMIGRATION ENFORCEMENT, THE RACIALIZATION OF LEGAL STATUS, AND PERCEPTIONS OF THE POLICE." *Du Bois Review: Social Science Research on Race* 15 (1): 107–28. https://doi.org/10.1017/S1742058X18000115.

Mervis, Jeffrey. 2018. "Plan for 2020 U.S. Census Is Fatally Flawed, Critics Say." *Science (New York, N.Y.)* 360 (6386): 250–51. https://doi.org/10.1126/science.360.6386.250.

Migration Policy Institute. 2018. "Unauthorized Immigrant Population Profiles." https://www.migrationpolicy.org/programs/us-immigration-policy-program-data-hub/unauthorized-immigrant-population-profiles.

Passel, J., and D. Cohn. 2018. "U.S. Unauthorized Immigration Total Lowest in a Decade." Washington, DC. http://www.pewhispanic.org/2018/11/27/u-s-unauthorized-immigrant-total-dips-to-lowest-level-in-a-decade/.

Passel, Jeffrey, and D'Vera Cohn. 2009. "A Portrait of Unauthorized Immigrants in the United States." Washington D.C. http://www.pewhispanic.org/files/reports/107.pdf.

Passel, Jeffrey S., and D'Vera Cohn. 2017. "20 Metro Areas Are Home to Six-in-Ten Unauthorized Immigrants in U.S."

Passel, Jeffrey S., and Karen A. Woodrow. 1987. "Change in the Undocumented Alien Population in the United States, 1979-1983." *International Migration Review* 21 (4): 1304. https://doi.org/10.2307/2546516.

Peterson, Christine E., Narayan Sastry, Anne R. Pebley, Bonnie Ghosh-Dastidar, Stephanie Williamson, and Sandraluz Lara-Cinisomo. 2004. "The Los Angeles Family and Neighborhood Survey." https://www.rand.org/pubs/drafts/DRU2400z2-1.html.

Prentice, Julia C., Anne R. Pebley, and Narayan Sastry. 2006. "PRENTICE ET AL. RESPOND."

American Journal of Public Health 96 (3): 399-a-400. https://doi.org/10.2105/AJPH.2005.079871.

Rendall, Michael S, Bonnie Ghosh-Dastidar, Margaret M Weden, Elizabeth H Baker, and Zafar Nazarov. 2013. "Multiple Imputation For Combined-Survey Estimation With Incomplete Regressors In One But Not Both Surveys." *Sociological Methods & Research* 42 (4). https://doi.org/10.1177/0049124113502947.

Rosenblum, Marc R., and Ariel G. Ruiz Soto. 2015. "An Analysis of Unauthorized Immigrants in the United States by Country and Region of Birth."

Sastry, Narayan, Bonnie Ghosh-Dastidar, John Adams, and Anne R Pebley. 2006. "The Design of a Multilevel Survey of Children, Families, and Communities: The Los Angeles Family and Neighborhood Survey." *Social Science Research* 35 (4): 1000–1024. https://doi.org/10.1016/J.SSRESEARCH.2005.08.002.

Snee, Ronald D. 1977. "Validation of Regression Models: Methods and Examples." *Technometrics* 19 (4): 415. https://doi.org/10.2307/1267881.

U.S. Citizenship and Immigration Services. 2017. "Immigration and Citizenship Data." 2017. https://www.uscis.gov/tools/reports-studies/immigration-forms-data.

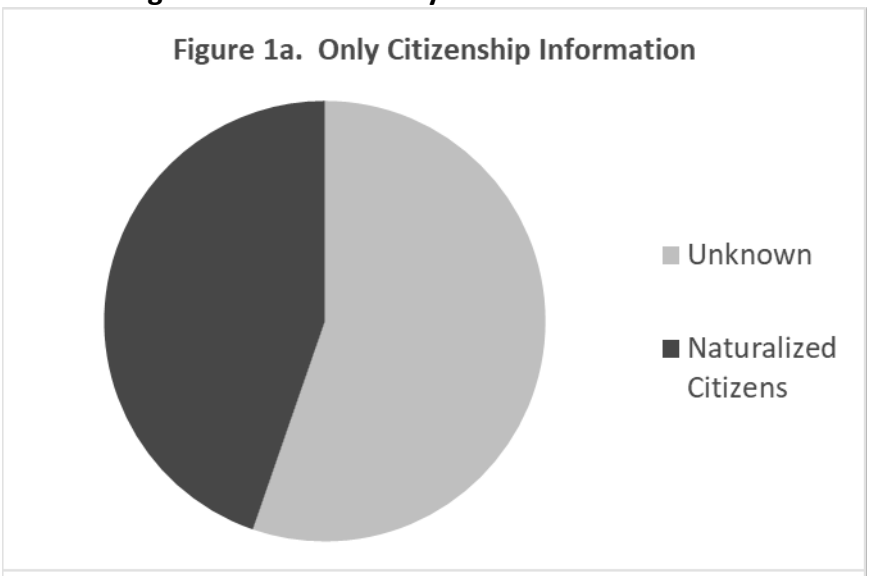
Vogel, Dita, Vesela Kovacheva, and Hannah Prescott. 2011. "The Size of the Irregular Migrant Population in the European Union – Counting the Uncountable?" *International Migration (Geneva, Switzerland)* 49 (5): 78–96. http://www.ncbi.nlm.nih.gov/pubmed/22167866.

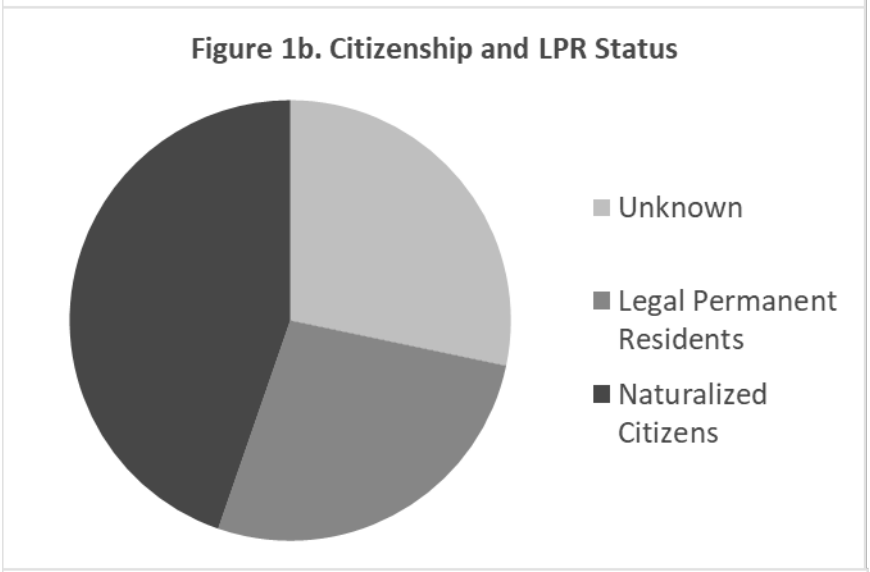
Warren, Robert, and Donald Kerwin. 2017. "Mass Deportations Would Impoverish US Families and Create Immense Social Costs." *Journal on Migration and Human Security* 5 (1): 1–8. https://doi.org/10.14240/jmhs.v5i1.71.

Warren, Robert, and Jeffrey S. Passel. 1987. "A Count of the Uncountable: Estimates of Undocumented Aliens Counted in the 1980 United States Census." *Demography* 24 (3): 375. https://doi.org/10.2307/2061304.

Warren, Robert, and John Robert Warren. 2013. "Unauthorized Immigration to the United States: Annual Estimates and Components of Change, by State, 1990 to 2010." *International Migration Review* 47 (2): 296–329. https://doi.org/10.1111/imre.12022.

Figure 1. Types of Information Available on the Documentation Status of Foreign-Born Respondents in Several Large US National Surveys





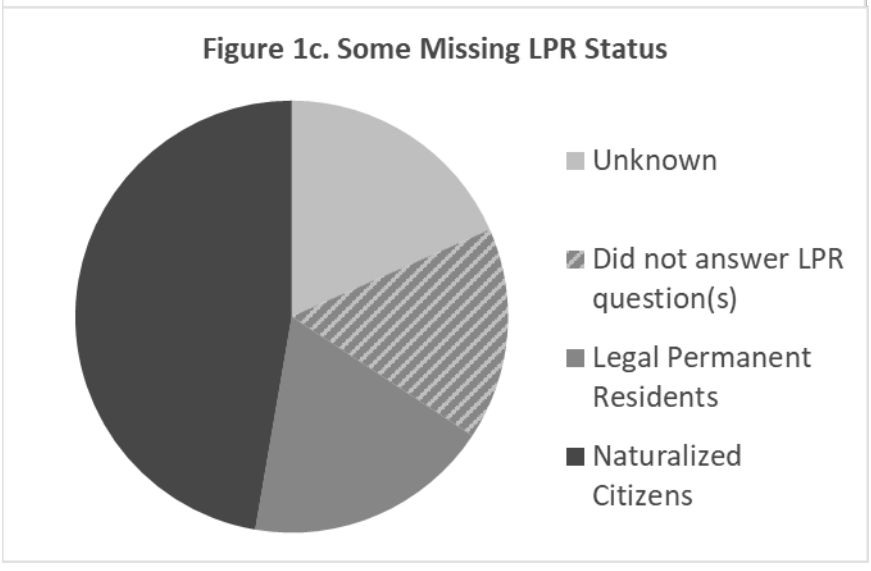


Table 1. Evaluation of Hall, Greenman, and Farkas (HGF)'s methodology using the L.A.FANS

	% of Total Foreign-born in L.A.FANS	
	Assuming LPR status *known* (1)	Assuming LPR status *unknown* (2)
Total undocumented in the L.A.FANS, self-report	27.2	27.2
Total undocumented in the L.A.FANS using HGF method	33.3	55.0
L.A.FANS undocumented, assigned documented by HGF method	2.0	2.0
L.A.FANS documented resident assigned as undocumented by HGF method	8.2	30.7

Data Source: L.A.FANS. Notes: Sample is limited to foreign-born respondents aged 18-60. (1) HGF method assigns documentation status assuming that LPR status is known and accurate. (2) When LPR status is completely unknown, the proportion of foreign-born respondents inaccurately categorized as undocumented increases to 30.7.

Table 2: Relative odds of being naturalized or undocumented associated with socio-demographic characteristics predicted from L.A.FANS

Variable	Naturalized	Undocumented
Gender		
Male	1.00	1.00
Female	1.65	0.79
Age	1.04	1.06
Age-squared	1.00	1.00
Marital Status		
Married	1.00	1.00
Separated	0.79	2.54
Widowed	0.46	0.43
Divorced	0.61	4.24
Cohabitating	0.64	1.56
Never Married	0.59	1.91
Number of children under 18 in household	0.90	1.21
Immigration Year		
Before 1980	1.00	1.00
1980-1989	0.30	7.99
1990-1994	0.08	22.20
1995 or later	0.00	46.68
Unknown	0.00	2.41
Region of birth		
North America	1.00	1.00
Mexico and Central America	2.22	12.25
Latin America and the Caribbean	10.70	1.75
Asia and Pacific	18.24	0.35
Africa, Europe, and Others	34.11	1.00
Educational Attainment		
High School Graduate or below	1.00	1.00
Some college or more	2.31	0.31
Health Insurance		
Had gaps in coverage during past 2 years	1.00	1.00
Continuously covered during past 2 years	1.56	0.32

Data: L.A.FANS. Notes: Analysis is limited to foreign-born immigrants aged 18 to 60 in the L.A.FANS.

Table 3. Sociodemographic characteristics of the undocumented population aged 18-60 by estimation method, Survey of Income and Program Participation (SIPP)

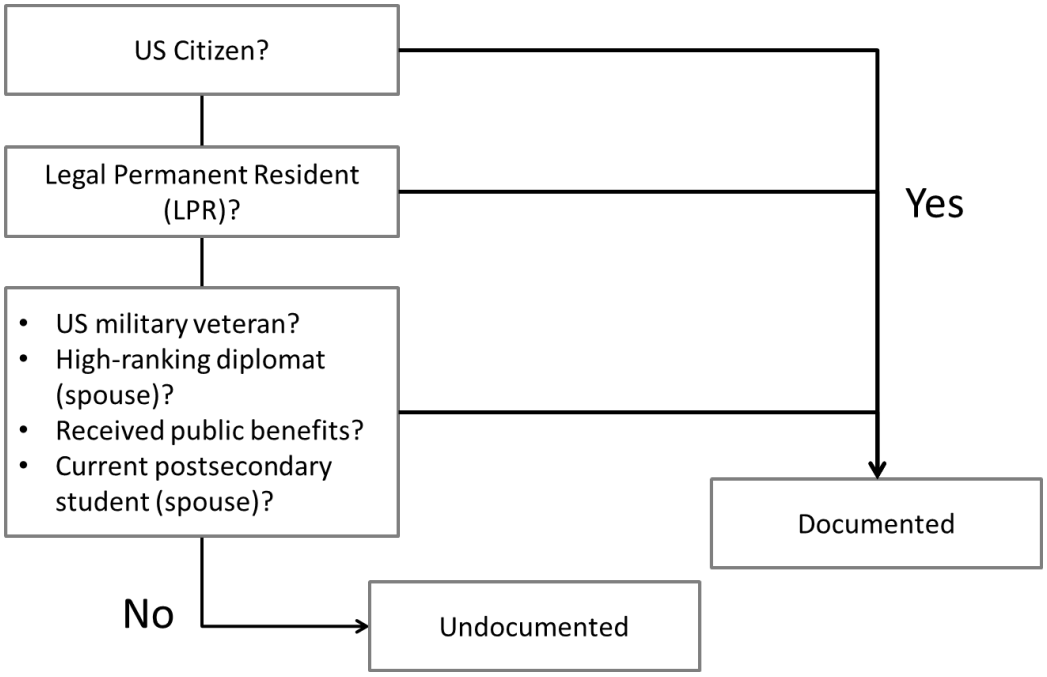
	LA-FANS based Prediction (% or means)	HGF Assignment Method (% or means)	Difference (col2-col1)	Difference as a Ratio of LA.FANS (=col3/col1)
Female*	40.8	42.9	2.1	0.1
Age Group*				
Less than 25	25.5	21.2	-4.3	-0.2
25-39	55.4	54.6	-0.8	0.0
40-54	17.8	20.5	2.7	0.2
55+	1.3	3.7	2.4	1.9
Marital Status*				
Currently (legally) married	54.1	58.0	3.8	0.1
Separated	3.9	2.8	-1.1	-0.3
Widowed	0.4	0.6	0.3	0.8
Divorced	3.3	2.1	-1.2	-0.4
Cohabiting	8.1	7.6	-0.6	-0.1
Never legally married	30.2	28.9	-1.2	0.0
Education*				
Less than High School	59.5	46.8	-12.7	-0.2
High School Graduate	27.1	21.3	-5.9	-0.2
Some College	9.1	13.9	4.8	0.5
College Graduate	3.3	10.7	7.4	2.3
Advanced Degree	1.1	7.4	6.3	5.7
Region of Birth*				
North America, excl. us	0.2	1.3	1.1	5.3
Mexico and Central America	89.1	66.1	-22.9	-0.3
Latin America and the Caribbean	6.2	10.7	4.5	0.7
Asia and Pacific	1.3	12.2	10.9	8.6
Africa, Europe, Other	3.3	9.8	6.4	1.9
Race/Ethnic Origin				
Non-Hispanic white	12.3	18.2	5.9	0.5
Latino/a	81.6	63.3	-18.3	-0.2
African American	3.9	5.6	1.8	0.5
Asian, Pacific Islander	1.9	12.5	10.7	5.7
Other	0.3	0.3	0.0	0.0
Number of Children in Household*				
None	33.7	40.8	7.1	0.2
One	20.1	22.4	2.3	0.1
Two	24.0	20.4	-3.6	-0.2
Three	14.2	11.0	-3.3	-0.2
Four or more	7.9	5.5	-2.4	-0.3
Poverty Level				
Below poverty	30.3	27.2	-3.1	-0.1
100-200%	37.7	34.5	-3.1	-0.1
200-400%	25.6	26.2	0.5	0.0
Greater than 400%	6.4	12.1	5.7	0.9
Health Insurance Coverage				
Covered in own name	15.1	24.4	9.3	0.6
Covered by someone else's plan	6.6	9.8	3.2	0.5
Covered in own name and someone else	0.7	0.8	0.1	0.2
Not covered	77.6	65.0	-12.7	-0.2

Table 3, continued

		LA-FANS based Prediction (% or means)	HGF Assignment Method (% or means)	Difference (col2-col1)	Difference as a Ratio of LA.FANS[1] (=col3/col1)
Income					
Total person earned income		14,449	16,642	2192.7	0.2
Total family earned income		36,075	37,581	1505.9	0.0
Total family income[1]		37,523	38,946	1422.4	0.0
Primary Occupation	*SOC code[2]*				
Did not have a job or business	na	24.6	25.7	1.1	0.0
Management, business/financial operations, computer/mathematical, architecture/engineering, and science	1X-XXXX	2.7	7.4	4.7	1.8
Community/social service, legal, education/training/library, art/design/media/sports, and healthcare	2X-XXXX	1.7	4.4	2.7	1.6
Healthcare support, protective services, food prep/serving, building grounds cleaning/maintenance, and personal care/service	3X-XXXX	23.4	21.7	-1.7	-0.1
Sales, office administrative support, farming/fishing/forestry, construction/extraction, and installation/maintenance/repair	4X-XXXX	29.4	24.8	-4.6	-0.2
Production, transportation/material moving, and military specific	5X-XXXX	18.2	16.0	-2.2	-0.1

Notes: * Variable is included in the predicting naturalization probabilities. 1 Total income includes earned income, property income, means-tested cash transfers, and other income. 2 First digit of Standard Occupational Classification (SOC) code

Appendix A: Hall, Greenman, and Farkas (HGF)'s method of assigning immigrants' documentation statuses



Appendix B: Implementation of the n-fold method on the L.A.FANS

We applied a validation procedure commonly referred as the n-fold method. This approach reduces the risk of over-fitting the model to the data and tests the model on portions of the data that were excluded from the prediction process. It also prevents a few individuals unique to the L.A.FANS from unduly influencing the prediction model. This is the preferred method for fitting and evaluating a prediction model from a moderately-sized dataset (Snee 1977).

First, we divided the L.A.FANS into 10 equal groups.

1	2	3	4	5	6	7	8	9	10

Second, we estimated the prediction equation from 9 out of the 10 groups (the training set) and using this equation predicted the documentation status of the respondents in the remaining 10th group (the validation set).

1	2	3	4	5	6	7	8	9	**10**

We repeated this procedure 10 Training set times rotating the group Validation set that serves as the validation set.

1	2	3	4	5	6	7	8	9	**10**
1	2	3	4	5	6	7	8	**9**	10
1	2	3	4	5	6	7	**8**	9	10
1	2	3	4	5	6	**7**	8	9	10
1	2	3	4	5	**6**	7	8	9	10
1	2	3	4	**5**	6	7	8	9	10
1	2	3	**4**	5	6	7	8	9	10
1	2	**3**	4	5	6	7	8	9	10
1	**2**	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

We derived our prediction model by averaging the coefficients across all 10 groups. Similarly, we calculate the model's performance by averaging the mean-squared error (predicted – actual) across the ten validation sets.

TableB1. Validation of Documentation Status Prediction Models using the L.A.FANS

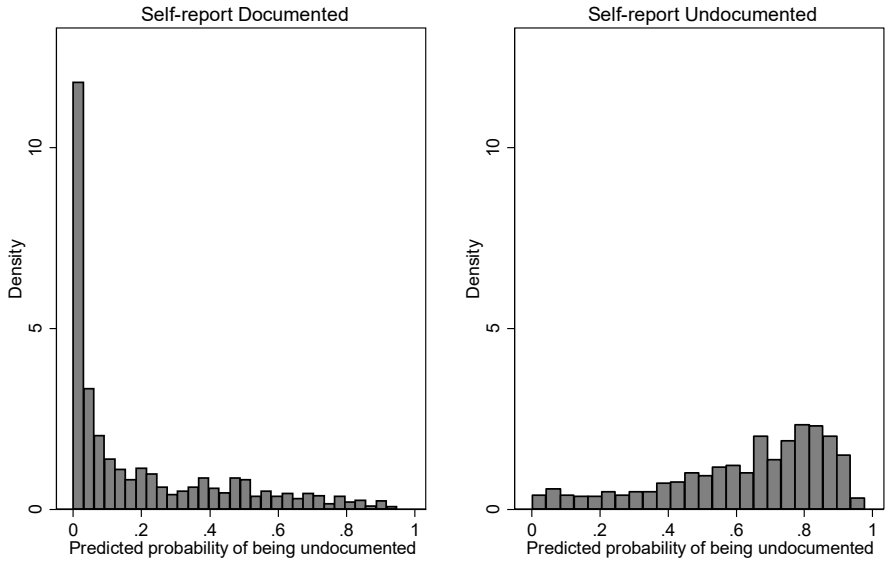
	Model 1 (baseline)	Model 2	Model 3	Model 4	Model 5
Documentation Status					
Mean squared-error[1]	0.199	0.173	0.114	0.111	0.109
SD of squared-error[2]	0.001	0.001	0.001	0.001	0.001
Naturalization Status					
Mean squared-error[1]	0.225	0.191	0.126	0.125	0.124
SD of squared-error[2]	0.001	0.001	0.002	0.002	0.002

Data: L.A.FANS. Notes: Analysis is limited to foreign-born immigrants aged 18 to 60 in the L.A.FANS. Model 1 predicts the probability of being undocumented or naturalized from a randomly generated binary variable. Model 1 serves as the baseline for evaluating models 2 to 5. Model 2 predicts documentation status from sex, age, and age-squared. Model 3 adds marital status, total number of children in the household, year of immigration, and place of birth to the predictors in model 2. Model 4 adds educational attainment to model 3's predictors. Model 5 adds continuous insurance coverage for the past 2 years. Performance is measured using the mean squared-error. Error is the difference between the predicted probability net the actual documentation status (0 = undocumented, 1 = documented). Standard deviation of the squared errors are derived from 50 repeated predictions. Each prediction yield slightly different results due to the random assignment of data into 10 prediction/test groups.

The first two rows of Appendix B Table 1 summarize the results of applying the prediction models to L.A.FANS to calculate the probability of being undocumented from observed characteristics and associated coefficients. We compared the performance of five regression models. The first predicts documentation status using random dummy variable[8]. This serves as the baseline for subsequent models as it predicts documentation status from a single unrelated variable. The second includes basic demographic variables. The third adds family and immigration background, the fourth model adds education, and the fifth adds continuous insurance coverage. The fifth model (the model that we will use to predict documentation status in the SIPP) performs substantially better than the baseline model (model 1). The baseline model, which uses a randomly generated binary variable to predict documentation status, yields a mean squared-error of 0.199. These values represent the average squared difference between the predicted probability (ranging from 0 to 1) and self-reported status (binary variable, 0 or 1). The mean squared-error decreases as demographic and socioeconomic characteristics are added as predictors. Model 5 performs almost twice as well as Model 1 yielding a mean squared-error of 0.109 for documentation status. The exact values of these performance measures depend on how the data is divided into ten random groups (using the n-fold procedure described above) during the prediction and validation process. In this analysis, we performed 50 iterations of this validation procedure randomizing the data each time. The performance outcomes did not vary much with standard deviations at 0.001 or below across all models.

[8] The dummy variable is a binary variable (0/1) is generated using a random number generator and has no correlation with documentation status.

Figure B1



Data: L.A.FANS. Notes: 50% of the people who report being documented have predicted probability of 0.026 and lower of being classified as undocumented. 50% of the people who report being undocumented have a predicted probability of 0.70 and higher of being undocumented.

Appendix B Figure 1 shows the distributions of probabilities of being undocumented which are predicted by the coefficients of the prediction equation for two groups of L.A.FANS respondents: those who report a documentation status and those who do not (and are assumed to be undocumented). If predicted and self-reported documentation status was identical, we would expect that all self-reported documented respondents would have a predicted probability of zero and self-reported undocumented would have a predicted probability of one. The first panel shows that the overwhelming majority of respondents who are documented have close to zero probability of being undocumented. Fifty percent of documented L.A.FANS respondents had less than 0.03 predicted probability of being undocumented. The second panel shows that the predicted probabilities of being undocumented are clearly higher for L.A.FANS respondents who are reported as undocumented. Our model predicted more than 50 percent of undocumented L.A.FANS respondents as having a greater than 0.70 probability of being undocumented.

Appendix C: Predicting naturalization status in the SIPP using L.A.FANS-based method

To determine whether it is feasible to use a prediction equation using data from Los Angeles County to produce national level estimates, we conducted a test in which we estimated a prediction model from L.A.FANS data for a variable which, unlike documentation status, *is* reported in SIPP and has a relatively high response rate: whether or not foreign-born respondents became US citizens through naturalization. We first estimate a model, based on the L.A.FANS foreign-born sample, in which we predict naturalization using the same variables described above for the documentation status prediction equation. Second, we apply the results of this equation to the SIPP data to predict naturalization. Third, we compare the predicted probabilities of naturalization for SIPP respondents and SIPP respondents' own reports of whether or not they naturalized. Finally, we compare sociodemographic characteristics of naturalized SIPP respondents using the predicted naturalization probabilities and self-reports to assess the accuracy of the prediction. As a sensitivity test, we repeated this test on two alternative subsets of the SIPP: respondents living in California and respondents living in Los Angeles, Riverside, and Orange Counties[9]. Our prediction equation performs very well and virtually identically on the national, California, and Los Angeles/Riverside/Orange County samples.

Table C1. Validation of Naturalization Status Prediction Models using the L.A.FANS

	Model 1 (baseline)	Model 2	Model 3	Model 4	Model 5
Documentation Status					
Mean squared-error[1]	0.199	0.173	0.114	0.111	0.109
SD of squared-error[2]	0.001	0.001	0.001	0.001	0.001
Naturalization Status					
Mean squared-error[1]	0.225	0.191	0.126	0.125	0.124
SD of squared-error[2]	0.001	0.001	0.002	0.002	0.002

Data: L.A.FANS. Notes: Analysis is limited to foreign-born immigrants aged 18 to 60 in the L.A.FANS. Model 1 predicts the probability of being undocumented or naturalized from a randomly generated binary variable. Model 1 serves as the baseline for evaluating models 2 to 5. Model 2 predicts documentation status from sex, age, and age-squared. Model 3 adds marital status, total number of children in the household, year of immigration, and place of birth to the predictors in model 2. Model 4 adds educational attainment to model 3's predictors. Model 5 adds continuous insurance coverage for the past 2 years. Performance is measured using the mean squared-error. Error is the difference between the predicted probability net the actual documentation status (0 = undocumented, 1 = documented). Standard deviation of the squared errors are derived from 50 repeated predictions. Each prediction yield slightly different results due to the random assignment of data into 10 prediction/test groups.

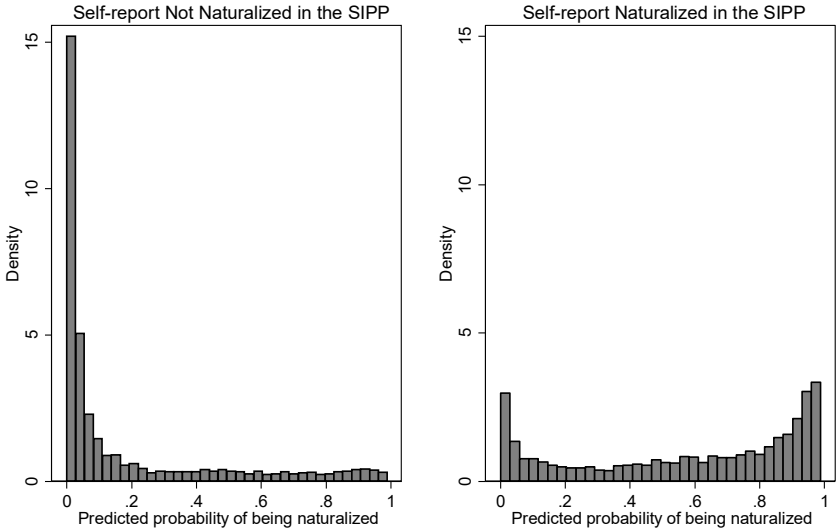
The results, shown in Table C1, are similar to those for documentation status: the baseline model (Model 1) produces a mean squared-error of 0.225 and Model 5 – which includes all the prediction variables – performs much better than Model 1 with a mean squared-error of 0.124. These results suggest that the prediction model does a good job in predicting naturalization status in L.A.FANS.

Next, we use the prediction equation to predict the probability of naturalization among foreign-born SIPP respondents based on their socio-demographic characteristics and the coefficients of the prediction model and then compare the results to their self-report of naturalization status. Figure C1 shows the distributions of predicted probabilities of naturalization for SIPP respondents using the L.A.FANS-based

[9] We used the 2001 panel of the SIPP to test residents of Los Angeles, Riverside, and Orange Counties as the variable identifying specific metro areas was eliminated in the 2004 SIPP.

equation for those reporting being naturalized and not naturalized in SIPP. The prediction model assigned more than 50 percent of *non-naturalized* SIPP immigrants to a probability of naturalization of less than 0.04. It also assigned more than 50 percent of *naturalized* SIPP immigrants to a predicted probability of greater than 0.65. Thus, the prediction model does a better job predicting respondents who report that they *are not* naturalized than predicting those who *are* naturalized. This result is discussed below. A crucial finding for our purposes is that the prediction equation performs as well in predicting naturalization status among SIPP respondents as it did predicting documentation status among respondents in the L.A.FANS. The second panel of Figure C1 shows a small group of SIPP respondents (less than 2 percent) who were predicted to have a low naturalization probability but reported that they were naturalized.

Figure C1



Notes: We apply multiple imputation to account for missing year of immigration to the US (non-response rates ~=25%). In this imputation, we use respondent's place of birth, age, age-squared, sex, education, fluency in English, marital status, number of own children in the household, type of health insurance coverage, and homeownership status to predict year of immigration. We then use the L.A.FANS-based naturalization prediction equation on the multiply imputed samples. 50 percent of unnaturalized immigrants had predicted naturalization probabilities less than 0.04. 50 percent of naturalized immigrants had predicted naturalization probabilities greater than 0.65. The area under the ROC curve is 0.82.

This high-error group (n=106) who reported that they were naturalized citizens in SIPP but are predicted not to be by our model are: relatively young with a median age of 28, have lived in the US on average 5.3 years, and are more likely to be men than women. Their socioeconomic status is, on average, low: median household income is less than $2,450 dollars per month, almost all of which is earned income. Almost all households reported zero income from property, unemployment, and social security. Less than a third completed high school and another third did not attend school beyond 6th grade. These characteristics – particularly the relatively short tenure in the US[10] and low education and income – are much more comparable to those of undocumented immigrants than to other naturalized citizens, suggesting at least two possible explanations. First, the group may, in fact, include undocumented immigrants who report themselves

[10] For example, lawful permanent residents must live in the US for at least 5 years in order to apply for citizenship and spouses of US citizens must be resident for at least 3 years.

as naturalized to avoid further questions about legal status. Second, this group may include individuals who immigrated to the US in special statuses as asylees and refugees.

Table C2 compares the characteristics of naturalized SIPP respondents using the L.A.FANS based prediction model with those of self-reported naturalized respondents. The L.A.FANS based model performed reasonably well in describing the naturalized population in the national sample interviewed in SIPP. Our model successfully predicted the distribution of naturalized citizens by variables —current health insurance coverage, income, and family poverty level – which were not included in the prediction equation. The largest differences between actual and predicted characteristics are in the age distribution and region of birth. Our prediction yielded an older age distribution and fewer people born in North and Central America and more people born in Africa and Europe.

The L.A.FANS-based model performed particularly well in predicting the primary occupation of the naturalized population in the United States. It predicted the distribution of naturalized citizens within 10 percent of the actual distribution for all 6 out of 6 occupational categories including smaller occupation groups that each accounts for less than 15 percent of the naturalized population.

Table C2. Sociodemographic characteristics of the ***naturalized population*** aged 18-60 using self-reported variables and predictions derived from the LA.FANS, Survey of Income and Program Participation (SIPP)

	Self-Reported in the SIPP (% or means)	**LA-FANS based Prediction** (% or means)	**Difference** (col2-col1)	**Difference as a Ratio of Self-Report**[1] (=col3/col1)
Female*	51.6	54.9	3.3	0.1
Age Group*				
Less than 25	7.5	3.3	-4.1	-0.6
25-39	32.3	25.9	-6.4	-0.2
40-54	45.9	52.1	6.2	0.1
55+	14.3	18.7	4.4	0.3
Marital Status*				
Currently (legally) married	69.0	71.5	2.5	0.0
Separated	2.2	3.0	0.9	0.4
Widowed	1.2	1.3	0.1	0.1
Divorced	8.1	8.5	0.5	0.1
Cohabiting	4.0	3.6	-0.4	-0.1
Never legally married	15.6	12.0	-3.5	-0.2
Education*				
Less than High School	19.6	17.6	-2.1	-0.1
High School Graduate	19.5	16.2	-3.3	-0.2
Some College	30.2	31.9	1.7	0.1
College Graduate	19.5	21.2	1.8	0.1
Advanced Degree	11.1	13.1	2.0	0.2
Region of Birth*				
North America, excl. us	1.8	1.3	-0.5	-0.3
Mexico and Central America	30.5	20.7	-9.8	-0.3
Latin America and the Caribbean	17.2	18.4	1.2	0.1
Asia and Pacific	28.7	30.1	1.4	0.0
Africa, Europe, Other	21.8	29.5	7.8	0.4
Race/Ethnic Origin				
Non-Hispanic white	28.2	32.8	4.5	0.2
Latino/a	32.8	24.5	-8.3	-0.3
African American	10.1	11.7	1.5	0.1
Asian, Pacific Islander	26.7	28.6	1.9	0.1
Other	2.2	2.5	0.3	0.1
Number of Children in Household*				
None	46.8	51.1	4.3	0.1
One	20.6	19.9	-0.7	0.0
Two	20.4	19.0	-1.5	-0.1
Three	8.2	7.2	-1.0	-0.1
Four or more	4.0	2.9	-1.0	-0.3
Poverty Level				
Below poverty	13.7	12.7	-1.0	-0.1
100-200%	19.3	16.3	-3.0	-0.2
200-400%	30.4	31.2	0.8	0.0
Greater than 400%	36.7	39.9	3.2	0.1
Health Insurance Coverage				
Covered in own name	49.1	48.9	-0.2	0.0
Covered by someone else's plan	21.8	24.9	3.1	0.1
Covered in own name and someone else	4.0	4.1	0.1	0.0
Not covered	25.2	22.2	-3.0	-0.1

Table C2, continued.

		Self-Reported in the SIPP (% or means)	LA-FANS based Prediction (% or means)	Difference (col2-col1)	Difference as a Ratio of Self-Report[1] (=col3/col1)
Income (Means)					
Total person earned income		31,862	32,853	991.3	0.0
Total family earned income		71,387	73,622	2235.2	0.0
Total family income[2]					
Primary Occupation	*SOC code[3]*				
Did not have a job or business	na	19.5	20.0	0.5	0.0
Management, business/financial operations, computer/mathematical, architecture/engineering, and science	1X-XXXX	15.9	17.2	1.4	0.1
Community/social service, legal, education/training/library, art/design/media/sports, and healthcare	2X-XXXX	11.7	12.6	1.0	0.1
Healthcare support, protective services, food prep/serving, building grounds cleaning/maintenance, and personal care/service	3X-XXXX	15.2	14.7	-0.5	0.0
Sales, office administrative support, farming/fishing/forestry, construction/extraction, and installation/maintenance/repair	4X-XXXX	25.2	23.6	-1.6	-0.1
Production, transportation/material moving, and military specific	5X-XXXX	12.6	11.8	-0.8	-0.1

Notes: * Variable is included in the predicting naturalization probabilities. 1 Calculated as (LA.FANS estimate -self-report)/self-report. 1 Calculated as (LA.FANS estimate -self-report)/self-report 2 Total income includes earned income, property income, means-tested cash transfers, and other income. 3 First digit of Standard Occupational Classification (SOC) code