New Approaches to Estimating Immigrant Documentation Status in Survey Data

Heeju Sohn and Anne R. Pebley
California Center for Population Research, UCLA

There were 43 million foreign-born residents in the United States in 2015 (U.S. Citizenship and Immigration Services 2017). About 11 million were estimated to be undocumented migrants (Passel and Cohn 2017; Rosenblum and Ruiz Soto 2015), a figure that appears to have remained relatively constant in the past decade. Roughly 70% of undocumented migrants were born in Mexico and Central America and 15% in Asia (Rosenblum and Ruiz Soto 2015). Undocumented immigrants are in an increasingly precarious position because of more rigorous enforcement of employer sanctions, an increasingly hostile political climate, and heightened deportation – and threats of deportation – under the Obama and Trump administrations.

The undocumented population is the subject of intensive interest and dispute in the political, policy, and research communities. On one hand, politicians and policy makers debate the consequences of undocumented immigration for labor markets, wages, public services (e.g., schools), and crime rates (Hanson 2007; Larsen et al. 2009; U.S. Immigration and Customs Enforcement 2017) and the effects of public policy on undocumented immigrant flows (Chishti and Bolter 2017; Felter and Renwick 2017). On the other hand, there are serious concerns about the effects of the hostile enforcement and political environment on the mental and physical health and human rights of undocumented migrants themselves (Amuedo-Dorantes and Pozo 2014; Becerra 2016; Williams and Medlock 2017). Researchers are also interested in a range of related questions including the consequences of legal status for: differentials in occupation, wages, and work conditions (Flippen 2012; Gentsch and Massey 2011; Hall et al. 2010a; Hall and Greenman 2015), immigrant incorporation (Abrego 2006; Gonzales 2011; Panel on the Integration of Immigrants into American Society 2015), and the effects of deportation, and fear of deportation, for family stability and child development (Enriquez 2015; Gulbas and Zayas 2017; Warren and Kerwin 2017).

Research on the undocumented population, and immigrants, more generally, has been hindered by the lack of data on immigrant legal status in the types of large-scale, nationally-representative survey data generally used to answer many of these questions. As Bachmeier, Van Hook, and Bean (2014) note, a common reason given for not collecting these data is the belief that questions about documentation are sensitive and produce high levels of misreporting and that they may have a chilling effect on respondents' willingness to participate in the survey. However, the results of their analysis suggest that these two issues are less problematic than the literature suggests. Another important reason for omitting questions on documentation status, however, particularly in government surveys, is the possibility of endangering undocumented respondents if they respond honestly and data including direct or indirect identifiers are somehow released (Carter-Pokras and Zambrana 2006; Prentice, Pebley, and Sastry 2006).

In the absence of direct and up-to-date data on documentation status for a large nationally-representative sample of individuals in the US, several researchers have recently begun to use extant survey data and indirect methods to estimate respondents' documentation status – something Van Hook et al. (2015) label "spinning straw into gold." Most of these studies are based on Survey of Income and Program Participation (SIPP) which collects data on whether respondents are citizens or legal permanent residents (Bachmeier et al. 2014; Hall et al. 2010a; Hall and Greenman 2015; Van Hook et al. 2015), although other datasets have also been used (Borjas 2017). Methods of estimating which respondents are undocumented vary but are all based on work by Passel and colleagues (Passel, Cohn, and Rohal 2014; Warren and Passel 1987). The approach using SIPP starts with the group of respondents who say that they are neither citizens nor legal permanent residents (LPRs) and assigns them to documented or undocumented status categories based on factors such as whether they are currently post-secondary level students, have received public social services not available to the undocumented population, 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.

The approach appears to work well in the aggregate in the sense that the total numbers of undocumented immigrants are comparable to other estimates (Bachmeier et al. 2014). Nonetheless, for analyses that depend on individual data, there are potential problems, described below, with using SIPP as a prime data source on undocumented immigrants. For example, as Bachmeier et al. (2014) note, there is substantial non-response for questions on legal permanent residence (LPR) status in SIPP. The way in which the non-response is handled could have an important effect on the results. Second, the approach potentially misallocated categories of immigrants who could be important in some substantive analyses including undocumented post-secondary students (both those who came to the US as young children and those overstaying their visas) who would be classified as "documented," and high tech H1-B visa holders (U.S.

Citizenship and Immigration Services 2017),agricultural and other workers on H2-A and H2-B visas all of whom would be classified as undocumented.

Our goal in this paper is to assess how well the indirect methods outlined above for use with SIPP work in a dataset that collected complete data on current documentation status – the Los Angeles Family and Neighborhood Survey (L.A.FANS). We also examine whether it is possible to improve the procedures for allocating non-response in SIPP and whether we can use data from L.A.FANS to improve the precision of assignment of SIPP respondents to documented and undocumented categories.

Our analysis uses the same two data sources as Bachmeier et al. (2014) – SIPP and L.A.FANS – but our goals are quite different. Bachmeier et al. (2014) used data from both surveys to assess 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 indirect methods used with SIPP by Bachmeier et al. (2014) and others actually do a good job of predicting documentation status in a sample where documentation status is known (at least to the extent that self-reports are reliable) and whether there are ways to improve the quality of this prediction.

Our analysis makes an important contribution to the understanding of the undocumented immigrant population, and the immigrant population more generally, in the US because it provides an evaluation of the quality of the indirect methods that are now being used by migration researchers to study documentation and its consequences. SIPP and similar data sets used with indirect methods are likely to provide the primary source of detailed individual level data on this population for the near future, particularly since human subject concerns are only likely to increase with greater threats of apprehension, detention, and deportation. Thus, understanding the strengths and weaknesses of this approach and methods of making potential improvements is particularly important.

Data

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

We use the first wave of the L.A.FANS data to evaluate Hall, Greenman, and Farkas (HGF)'s methodology (Hall et al. 2010a) and to estimate probabilities of being undocumented associated with observed socioeconomic characteristics. Unlike other large scale representative surveys, L.A.FANS asks a series of direct questions about respondents' documentation status. Foreign-born respondents are asked, sequentially, whether they are naturalized, permanent residents, or in the US on a temporary visa (and if the visa is still valid). Respondents who replied no to all these categories are assumed undocumented. An advantage of these questions is that L.A.FANS can identify residents with non-immigrant visas whereas the SIPP cannot. L.A.FANS also has a high response rates for immigration-related questions than SIPP. The non-response rates for these immigrant status questions are relatively low at 1.4 to 6.8 percent compared to the SIPP's 2.6 to 25.3 percent (Table 3, Bachmeier, Van Hook, and Bean 2014).

Survey of Income and Program Participation (SIPP)

Despite the L.A.FANS' distinct advantages over the SIPP in determining immigration statuses of its respondents, the SIPP is better suited to study immigrants in the US as a whole because of its nationally representative sample, large sample size, and regular availability of data over a long period of time. We use the 2004 panel of the SIPP to examine the characteristics of undocumented immigrants. The SIPP 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. The 2004 SIPP also asks citizenship status and immigration histories, albeit with limitations. While the SIPP collected data longitudinally across the 4-year study period, our analysis only uses data from Wave 2 when the survey asked questions on current citizenship status and immigration histories.

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 individuals and is weighted to represent the residents of Los Angeles County. The SIPP sample comprises 7,507 individuals and is weighted to compensate for SIPP's initial selection probabilities and differential attrition between waves 1 and 2 across subpopulation.

Analytical Strategy

We first evaluate the HGF method by comparing its allocated documentation status against self-reported status in the L.A.FANS. Next, we estimate a model that predicts documentation status from socioeconomic

characteristics of respondents in L.A.FANS. Then, we apply the coefficients of this model to the characteristics of respondents in the SIPP. Our approach improves upon the HGF method in two ways. First, we impute missing or non-responses of key variables in the SIPP using multiple imputation. Second, our models assign *probabilities* of being undocumented given a set of characteristics rather than deterministically categorizing each individual as either documented or undocumented. Appendix 1 presents a summary of our approach.

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 et al. 2014; Borjas 2017) identifies undocumented persons in the SIPP by eliminating individuals from the pool of foreign-born respondents based on their permanent resident status, receipt of federal assistance, attendance at a post-secondary school, and whether the respondent's spouse is high-ranking official or a post-secondary student. Using the LPR status at face value in the SIPP can be problematic because this variable has a high non-response rates (25 percent). Furthermore, non-response is likely greater among respondents who are undocumented (Bachmeier et al. 2014). We apply the same HGF imputation strategy to respondents in the L.A.FANS under two scenarios. In the first scenario, we assume that the LPR status is known. In the second, we assumed that LPR status is unknown or inaccurate. We then compared the imputed documentation status against the reported status among respondents in the L.A.FANS. During this step, we evaluate the effectiveness of each criterion in identifying documentation status among respondents. To do so, we compare respondents' reported documentation status and their status produced by the HGF procedure.

Constructing a prediction equation from the L.A.FANS to assign probabilities of being undocumented to SIPP respondents

We take advantage of the relatively complete immigration status variables in the L.A.FANS to estimate a model that predicts documentation status as a function of observed socioeconomic characteristics and their corresponding coefficients. Each co-variate is associated with an odds ratio that increases or decreases the probability of being undocumented. The variables that included in the prediction model are derived from findings in the literature on the characteristics of undocumented persons in the United States: year of immigration, receipt of federal assistance, educational attainment, spouse's occupation/student status, race/ethnicity, marital status, number of children in the household, past and current health insurance coverage, place of birth, age, and sex. We then apply the coefficients to the socioeconomic characteristics of respondents in the SIPP. Amount of time since moving to the United States—a significant predictors of documentation status—has a high non-response rate in the SIPP. The likelihood of a non-response is also skewed towards people who are likely undocumented. We impute missing values using multiple imputation that considers the 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. The final step applies the coefficients estimated from our prediction model from L.A.FANS to calculate the probability of being undocumented for each individual in the SIPP.

Estimating socioeconomic characteristics of the undocumented population of the United States

Averaging the probabilities of being undocumented across individuals in a particular group, say the unemployed, yields the probability of being undocumented conditional on being unemployed. We apply Bayes' Theorem (Bayes and Price 1763) to reverse these probabilities so that we can estimate the proportion of undocumented immigrates who unemployed. This approach does not attempt to identify undocumented individuals in the SIPP but rather, derives the population-level summaries of the undocumented as a group.

Our substantive focus is on two key policy issues related to immigrant documentation status. First, we examine the prevalence of mixed documentation status families and the characteristics of undocumented parents of U.S. citizen children. Mixed-status families have been the focus of tremendous attention since the new administration was inaugurated because of concerns about the effects of increased apprehension and deportation for these families and particularly the children in them. Second, we compare the occupational distributions of documented and undocumented adult workers by broad educational attainment categories. The occupational distribution of undocumented workers, particularly for the pre-Great Recession period, is the focus on public debate because of concern that less immigration enforcement during this period lead to displacement of U.S. citizen and legal immigrant workers by undocumented workers.

Preliminary Results

The HGF-approach in identifying undocumented immigrants in the L.A.FANS is largely consistent with their responses to immigration status questions in the L.A.FANS when their LPR statuses are known. About six percent of foreign-born respondents are categorized by HGF as documented despite reporting undocumented status. About 7.6 percent of foreign-born who reported that they have a temporary visa to live in the US are categorized as undocumented by HGF (Table 1). However, when LPR status is unknown (as it is for about a quarter eligible SIPP respondents), HGF's method performs poorly in identifying legal residents in the L.A.FANS. On one extreme, without any knowledge of respondents' LPR statuses, the HGF approach produces an undocumented population that encompasses over 60 percent of foreign-born residents in Los Angeles. Responses in the L.A.FANS show that about 32 percent of foreign-born are undocumented.

In the second part of the analysis, using a multivariate logistical regression with L.A.FANS data, we estimate the odds ratios of being undocumented associated with each covariate. The resulting coefficients are consistent with prior descriptions of the undocumented population in the US (Passel and Cohn 2009). Arriving in the US prior to1980, receiving federal assistance, being enrolled in or graduated from college, having continuous health insurance coverage, and living with fewer children in the household were significantly associated with being legal US resident. The covariates in the model accounted for almost half (R-squared value of 0.48) of the variation in documentation status among the foreign-born. We then used these coefficients to predict the probability that a person is undocumented in both the L.A.FANS and the SIPP. We emphasize that these prediction models do not include any explicit immigration status variables (i.e., naturalized, LPR, or temporary visas) other than whether the respondent was born a US Citizen or not.

Figure 1 is shows the results of applying the prediction models to L.A.FANS data, specifically, the distributions of predicted probabilities of being undocumented as determined by our model separately by respondents' actual reported status in the L.A.FANS. Approximately 10 percent of legal residents were assigned a probability of being undocumented of 0.5 and higher. About 75 percent of undocumented residents were assigned a greater than 0.5 probability of being undocumented. This means that if we were to apply a dichotomy for documentation status with a cutoff value of 0.5, 10 percent of legal residents would be incorrectly categorized as undocumented and 25 percent of undocumented immigrants would be categorized as documented. Shifting the cutoff values to a higher probability will reduce the likelihood of false positives but increase false negatives. Next, we applied the coefficients estimated from L.A.FANS to the observed characteristics of SIPP respondent. Figure 2 shows the distributions of predicted probabilities of being undocumented for the entire foreign-born population. The weighted average of these probabilities provides an estimate of the undocumented foreign-born as 28 percent of the foreign-born population; this is well within the range of estimates derived from prior estimates that used residual methods, multiple imputations, and logical assignments on the SIPP and the CPS (Bachmeier et al. 2014).

Our last set of results (Table 2) describes the sociodemographic profile of undocumented immigrants living in the US. To illustrate the approach we will use in this section of the paper, we compare the distributions of age, duration in the US, and region of birth of the undocumented population resulting from our methodology against a profile resulting from HGF's approach.

Conclusion

Our analysis suggests that HGF's method produces estimates of the undocumented population that are fairly consistent with our own and others. However, depending on the type of analysis to be undertaken, the HGF methods used with SIPP can misclassify some types of immigrants in terms of occupation status. In particular, HGF's approach yields an undocumented population that is younger, with fewer years in the US, and is more biased towards people born in Mexico and Central America. As described above, in the final paper, we will also examine the characteristics of parents in mixed-documentation status families and the occupational distribution of immigrants by documentation status and education using both the HGF method and our approach.

References

Abrego, Leisy Janet. 2006. "'I Can't Go to College Because I Don't Have Papers': Incorporation Patterns Of Latino Undocumented Youth." *Latino Studies* 4(3):212–31. Retrieved September 27, 2017 (http://link.springer.com/10.1057/palgrave.lst.8600200).

Amuedo-Dorantes, Catalina and Susan Pozo. 2014. "On the Intended and Unintended Consequences of Enhanced U.S. Border and Interior Immigration Enforcement: Evidence From Mexican Deportees." *Demography* 51(6):2255–79. Retrieved April 14, 2017 (http://link.springer.com/10.1007/s13524-014-0340-7).

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.

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. Retrieved September 21, 2017 (http://rstl.royalsocietypublishing.org/cgi/doi/10.1098/rstl.1763.0053).

Becerra, David. 2016. "Anti-Immigration Policies and Fear of Deportation: A Human Rights Issue." *Journal of Human Rights and Social Work* 1(3):109–19. Retrieved September 27, 2017 (http://link.springer.com/10.1007/s41134-016-0018-8).

Borjas, George J. 2017. "The Labor Supply of Undocumented Immigrants." *Labour Economics* 46:1–13. Retrieved May 30, 2017 (http://www.sciencedirect.com/science/article/pii/S0927537117301057).

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. Retrieved September 27, 2017 (http://ajph.aphapublications.org/doi/10.2105/AJPH.2005.078253).

Chishti, M. and J. Bolter. 2017. *The Trump Administration at Six Months: A Sea Change in Immigration Enforcement.*

Enriquez, Laura E. 2015. "Multigenerational Punishment: Shared Experiences of Undocumented Immigration Status Within Mixed-Status Families." *Journal of Marriage and Family* 77(4):939–53. Retrieved September 27, 2017 (http://doi.wiley.com/10.1111/jomf.12196).

Felter, C. and D. Renwick. 2017. *The U.S. Immigration Debate*.

Flippen, Chenoa A. 2012. "Laboring Underground: The Employment Patterns of Hispanic Immigrant Men in Durham, NC." *Social problems* 59(1):21–42. Retrieved May 30, 2017 (http://www.ncbi.nlm.nih.gov/pubmed/22844159).

Gentsch, Kerstin and Douglas S. Massey. 2011. "Labor Market Outcomes for Legal Mexican Immigrants Under the New Regime of Immigration Enforcement*." *Social Science Quarterly* 92(3):875–93. Retrieved May 30, 2017 (http://doi.wiley.com/10.1111/j.1540-6237.2011.00795.x).

Gonzales, Roberto G. 2011. "Learning to Be Illegal." *American Sociological Review* 76(4):602–19. Retrieved September 27, 2017 (http://journals.sagepub.com/doi/10.1177/0003122411411901).

Gulbas, L. E. and L. H. Zayas. 2017. *Exploring the Effects of US Immigration Enforcement on the Well-Being of Citizen Children in Mexican Immigrant Families*.

Hall, Matthew, E. Greenman, G. Farkas, and Amada Armenta. 2010a. "Legal Status and Wage Disparities for Mexican Immigrants." *Social Forces* 89(2):491–513.

Hall, Matthew, E. Greenman, G. Farkas, and Amada Armenta. 2010b. "Legal Status and Wage Disparities for Mexican Immigrants." *Social Forces* 89(2):491–513. Retrieved April 14, 2017 (https://academic.oup.com/sf/article-lookup/doi/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.

Hanson, G. H. 2007. *The Economic Logic of Illegal Immigration: Council on Foreign Relations*. New York, NY.

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. Retrieved September 27, 2017 (http://www.ncbi.nlm.nih.gov/pubmed/25511332).

Larsen, Knud, Krum Krumov, H. Le, Reidar Ommundsen, and Kees der Veer. 2009. "Threat Perception and Attitudes Toward Documented and Undocumented Immigrants in the United States: Framing the Debate and Conflict Resolution." *European Journal of Social Sciences* 7:115–34.

Panel on the Integration of Immigrants into American Society. 2015. "Legal Status and Immigrant Integration." Pp. 93–158 in *The Integration of Immigrants into American Society*, edited by M. Waters and M. Pineau. Washington D.C.: National Academies Press.

Passel, Jeffrey and D'Vera Cohn. 2009. *A Portrait of Unauthorized Immigrants in the United States*. Washington D.C. Retrieved (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., D'Vera Cohn, and Molly Rohal. 2014. *Unauthorized Immigration Totals Rise in 7 States, Fall in 14*. Washington D.C.
Prentice, Julia C., Anne R. Pebley, and Narayan Sastry. 2006. "PRENTICE ET AL. RESPOND." *American Journal of Public Health* 96(3):399-NaN-400. Retrieved September 27, 2017 (http://ajph.aphapublications.org/doi/10.2105/AJPH.2005.079871).
Rosenblum, Marc R. and Ariel G. Ruiz Soto. 2015. *An Analysis of Unauthorized Immigrants in the United States by Country and Region of Birth*.
U.S. Citizenship and Immigration Services. 2017. "Immigration and Citizenship Data." Retrieved (https://www.uscis.gov/tools/reports-studies/immigration-forms-data).
U.S. Immigration and Customs Enforcement. 2017. *DHS Announces Launch of New Office for Victims of Illegal Immigrant Crime*.
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. Retrieved September 27, 2017 (http://jmhs.cmsny.org/index.php/jmhs/article/view/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. Retrieved September 27, 2017 (http://link.springer.com/10.2307/2061304).
Williams, David R. and Morgan M. Medlock. 2017. "Health Effects of Dramatic Societal Events — Ramifications of the Recent Presidential Election" edited by D. Malina. *New England Journal of Medicine* 376(23):2295–99. Retrieved September 27, 2017 (http://www.nejm.org/doi/10.1056/NEJMms1702111).

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

	% of Total Foreign-born in L.A.FANS
Total undocumented in the L.A.FANS	31.9
Total undocumented in the L.A.FANS using HGF method	38.1
L.A.FANS undocumented, assigned legal by HGF method	6.0
L.A.FANS legal resident assigned as undocumented by HGF method	7.6
Total	100

Notes: Sample is limited to respondents aged 18-60. HGF method assigns documentation status assuming that LPR status is known and accurate. When LPR status is completely unknown, the proportion of foreign-born respondents inaccurately categorized as undocumented increases from 7.6 to 31.9.

Figure1. Distribution of predicted probabilities of being undocumented by immigration status from L.A.FANS

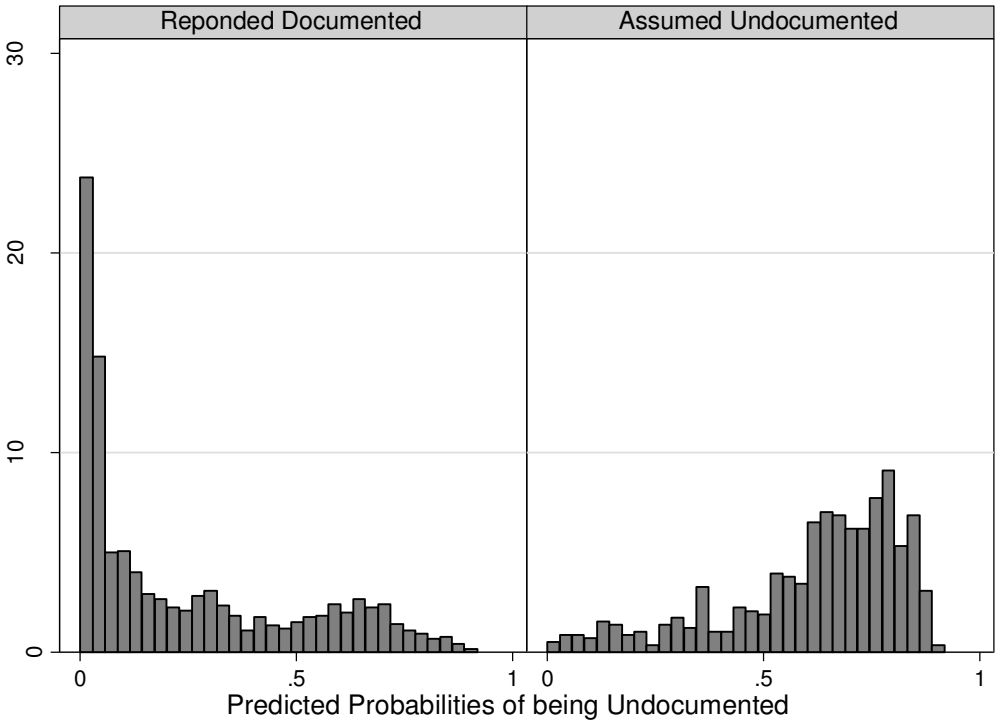
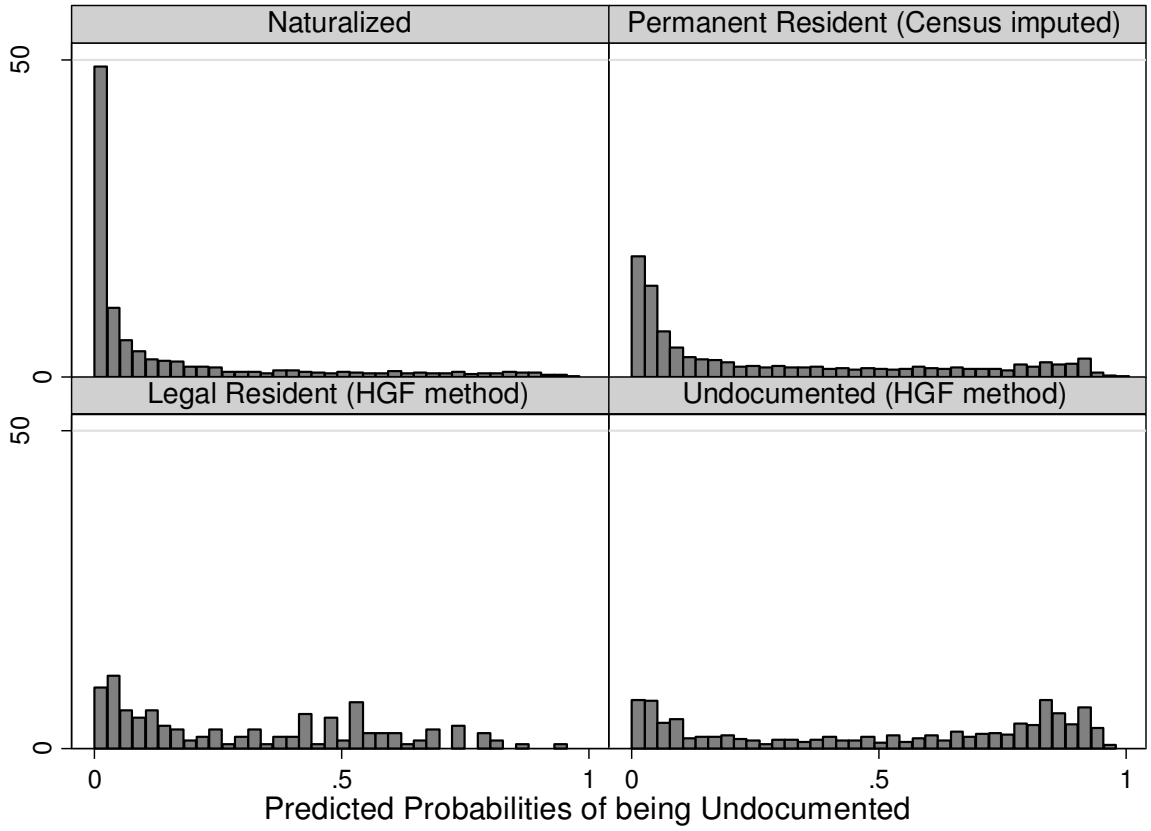


Figure 2. Distribution of predicted probabilities of being undocumented by immigration status determined by HGF's approach in the SIPP



Notes: Categorizations use Census-imputed LPR variables to replicate HGF's methodology

Table2. Socioeconomic characteristics of the undocumented population in the SIPP derived from predicted probabilities compared against results from HGF

	Predicted Probabilities	HGF Method
Age group		
18-24	17.92	20.67
25-34	37.58	41.94
45-44	21.06	19.85
54-55	12.46	11.48
55+	6.96	6.06
Years in the US		
0-5	34.32	41.58
10-16	17.76	20.68
11-15	6.96	7.82
15+	40.96	29.92
Region of birth		
Asia	12.45	12.3
Caribbean	5.54	3.77
Central America	8.84	9.57
Europe/Canada	7.69	5.68
Mexico	51.22	56.15
Middle East, Africa	6.04	5.13
South America	8.22	7.4

Notes: Represents the characteristics of respondents in the second wave of the 2004 SIPP. Probabilities are derived from probabilities of being undocumented assigned to each individual. Prediction equation is estimated from L.A.FANS.

Appendix Approaches to assigning documentation status

Respondents	Documentation Status Allocation Method	Analysis	Results
L.A.FANS	HGF method using LPR status from L.A.FANS	Compare reported and allocated documentation statuses	Table 1
L.A.FANS	HGF method without using LPR status	Compare reported and allocated documentation statuses	Footnote of Table 1
L.A.FANS	Prediction model[1] without immigration status variables	Compare reported and predicted documentation statuses	Figure 1
SIPP	Apply prediction model[1] without immigration status variables; impute missing variables used for prediction using multiple imputation	Compare documentation statuses from HGF method and prediction method	Figure 2

1 Prediction model is a logistical regression that predicts documentation status as a function of year of immigration, receipt of federal assistance, educational attainment, spouse's occupation/student status, race/ethnicity, marital status, number of children in the household, past and current health insurance coverage, place of birth, age, age-squared, and sex. Coefficients are estimated from responses in L.A.FANS.